INSITUTE OF CHEMICAL TECHNOLOGY, PRAGUE
Faculty of Food and Biochemical Technology
Department of Biochemistry and Microbiology

BACHELOR THESIS

Potential energy stabilizing a hydrophobic core of protein and its contribution to overall stability

Author: Boris Fačkovec

Supervisor: doc. Ing. Vojtěch Spiwok, Ph.D.
Consultant: RNDr. Jiří Vondrášek, CSc.
Study program: Chemistry
Year: 2010

Boris Fačkovec

Prague, 4th June 2010

Tato bakalárska práca bola vypracovaná v Centre biomolekúl a komplexných molekulových systémov na Ústave organickej chémie a biochémie AV ČR, v.v.i. pod vedením RNDr. Jiřího Vondráška CSc. v období nevember 2009 - jún 2010.

The thesis was worked out in the Centre of Biomolecules and Complex Molecular Structures at the Institute of Organic Chemistry and Biochemistry AS CR under supervision of RNDr. Jiří Vondrášek CSc. since November 2009 till June 2010.

Vyhlasujem, že som bakalársku prácu vypracoval samostatne s vyznačením všetkých použitých prameňov a spoluautorstva. Súhlasím so zverejnením bakalárskej práce podľa zákona č. 111/1998 Sb., o vysokých školách, v znení neskorších predpisov. Bol som zoznámený s tým, že se na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorský zákon, v znení neskorších predpisov.

I hereby declare that the thesis presented on following pages is presentation of my original work and that all the sources I used are properly cited.

Boris Fačkovec

V Praze dne 4. června 2010
Prague, 4th June 2010

SUMMARY

Interaction energy matrix concept is a promising approach to study protein folding, binding and function. The concept of hydrophobic core provides valuable opportunity to unify thermodynamic, kinetic, evolutionary and structural points of view. Method guaranteeing transferable and objective identification of key residues can make possible their further investigations and is the main purpose of this work.

The thesis introduces statistical mechanics, molecular modeling and structural biology backgrounds essential for theoretical modeling of the protein folding. Graph representation of proteins in energy space is utilized to characterize energy proportions of side-chains and importance of particular contact types. New definition of contact is proposed and finally, established energy quantities are used to determine residue importance.

SÚHRN

Koncept matice interakčných energií predstavuje perspektívny nástroj štúdia energetiky proteínov. Koncept hydrofóbneho jadra je mostom medzi termodynamickým, kinetickým, evolučným a štruktúrnym prístupom. Jednotná a objektívna metóda identifikácie kľúčových aminokyselín, ktorej vývoj je hlavným cieľom tejto práce, je nevyhnutná pre ich hlbšie štúdium.

Elaborát začína uvedením základných princípov štatistickej mechaniky, molekulárneho modelovania a štruktúrnej biológie nevyhnutných pre modelovanie stability proteínov. Reprezentácia proteínu kompletným grafom s hranami váženými interakčnými energiami bočných reťazcov je využitá na charakterizáciu energetických pomerov a posúdenie významu jednotlivých párových príspevkov pre celkovú stabilitu. Novo zadefinované energetické veličiny sú využité pri odhade významu jednotlivých reziduí pre celkovú stabilitu.

ACKNOWLEDGEMENT

I would like to thank my supervisor dr. Jiří Vondrášek for introducing me to the field and for his devoted supervision and support before and during working on the thesis. Many thanks are owed to dr. Karel Berka for his guidance in the time of my scientific beginnings and for introduction to experimental methods. I am also grateful to all my colleagues in the Center, namely Jiří Vymětal, for enriching discussions. Finally, I need to thank prof. Pavel Hobza for offering me the opportunity to work in the outstanding group of researchers and people and to dr. Vojtěch Spiwok for the trust and patience he bestowed upon me.

Contents

1 Introduction

Proteins are without doubts the most complex biomolecules. Their roles in organisms range from energy deposition through cellular scaffolding to molecular catalysis. Their structure is primarily coded in a sequence of DNA. An apparent question emerges - how this simple receipt can lead to such huge variety of shapes, function and behavior and last but not least to a complexity demonstrated by any living matter.

A protein chain is single-stranded heteropolymer composed of 20 different amino acids (AA) and the sequence is held by peptide bonds. The full information for 3D structure and function is written in linear sequence of composing amino acids. The genetic code is also linear and it is easy to switch from the alphabet of nuclear bases to the AA alphabet. The most interesting yet complicated part comes with the way how a protein is formed during process of structure formation called protein folding (PF). It is the folding itself that knots the information thread between inheritance and function.

According to Anfinsen's dogma (Anfinsen, 1973) protein structure is fully determined by the sequence and the structure represents a thermodynamic (free energy) minimum of all possible folds in the particular environment. In the light of structure-function paradigm a protein has to adopt proper conformation called native state (N) to perform its biological function. Deeper understanding of protein energy equilibrium is essential for further understanding and development of biological sciences. The Gibbs free energy minimum is acquired by interresidual interactions and interactions with an environment in three spatial dimensions. Major role in protein's structure and stability is played by non-covalent interactions between protein's building blocks. Their strength is in average less than 1/20 of covalent interactions energy but they are so numerous that they drive the molecule to final, stable and unique conformation.

The relation between the sequence and the fold space is non-linear. Similarity in sequence does not necessarily guarantee structure similarity and vice versa. Protein folding and binding together with nonlinear dynamics of enzymatic kinetics might stand behind the need of changing the paradigm from physical law system to biological system with "Elan vital". Protein folding problem (PFP) surely deserves the tag of being "the holy grail of biochemistry" .

There is already a theoretical solution of the protein folding problem. An exhaustive conformational search of structure/energy space. However, this solution is computationally expensive due to the exponential increase of computational time upon protein size. Moreover, evaluation of cooperative effects and entropy is extremely difficult and inaccurate, so we are

limited to find reasonable and accurate evaluation of the enthalpy term and find a way how to determine the entropy part in the Gibbs free energy equation for protein equilibrium.

To get closer to the solution of protein folding phenomenon new theories and models are systematically explored. They are required to provide reasonably precise information and efficiency regarding the whole configuration space sampling. The literature concerning energy in proteins and their processes is comprehensive. It somehow reflects the impact of results to the almost endless variety of their utilization (Dill, 1990; Baldwin, 2007).

The chapter "The current state of the art" introduces statistical mechanics, molecular modeling and structural biology backgrounds essential for theoretical modeling of the PF process. Overview starts with theory of potential energy calculation of simple microstates and continues to a discussion of pitfalls and perspectives of particular protein modeling approaches. It also provides the basic principles and their main characteristics. Interactions accessible to both experimental and theoretical treatment are discussed later. Cooperativity of interactions results in protein thermodynamics and macroscopically observed properties are also discussed.

In order to gain most of the information from available data, we introduce interaction energy matrix concept. The matrix is constructed in such way that it contains information about energy proportions in a protein of known structure. Interaction energy matrix is vital and still developing concept of current protein science field.

The most common and intuitive model of globular proteins is a sphere with charged and polar amino acid (AA) residues on the surface and hydrophobic residues in the interior. Generally, the interior of a protein is called hydrophobic core (HC). It has been recently described that hydrophobic core is at some extent proportional to a stability of thermostable proteins (Vondrášek and Hobza, 2007) and it plays important role in the kinetics of protein folding. Investigation and description of hydrophobic core properties is the main purpose of this work.

We start our work investigating energy proportions in proteins. Distributions of side chain interaction energies on representative set of protein structures are made. Then we discuss descriptors of residue importance and introduce contact order and relative residue interaction energy based on interaction energy matrix. Finally we propose four transferable and objective methods of identification key residues in protein structure.

2 The current state of the art

2.1 Protein modeling

Despite their fundamental role in the nature, the full understanding of proteins' behavior and their folding is still not complete. The field evolves for decades and numerous models of proteins have been proposed. To our knowledge none of the model satisfies criteria of simplicity, accuracy, productiveness and robustness.

What are the requirements for a good model? The model should elucidate some features not easily accessible or explainable. It should also provide a good deal of mathematical abstraction of the problem and use numerous data to support proposed mechanisms. While dealing with proteins we face a problem of complexity of the issue. There is a need for a coherent description that would have intuitive physical background and ability to explain and predict.

2.1.1 General representation of a protein

In this work we concern only globular, fully structured proteins. A protein molecule often consists of hundreds to tens of thousands of atoms. The only exact way how to model a protein is all-atom quantum dynamics simulation in natural protein environment. Obvious inconveniences coming with this approach imply a reduction and approximation. Even a use of quantum mechanics (QM) calculations to evaluate single point energy is too expensive because non-covalent interactions are difficult to calculate accurately and we must use expensive methods (at least CCSD(T) with CBS limit correction (Müller-Dethlefs and Hobza, 2000)). Berka et al. (Berka and Vondrášek, 2009) calculated side chain interaction energies (IE) of representative clusters of AA residues in proteins. They showed that most of the non-covalent interaction energy contributions are well described already at the force fields (FF) level as for example AMBER and OPLS. Although they severely fail to evaluate the electrostatic energy terms, they still represent very powerful method to get an idea of protein and its energy balance.

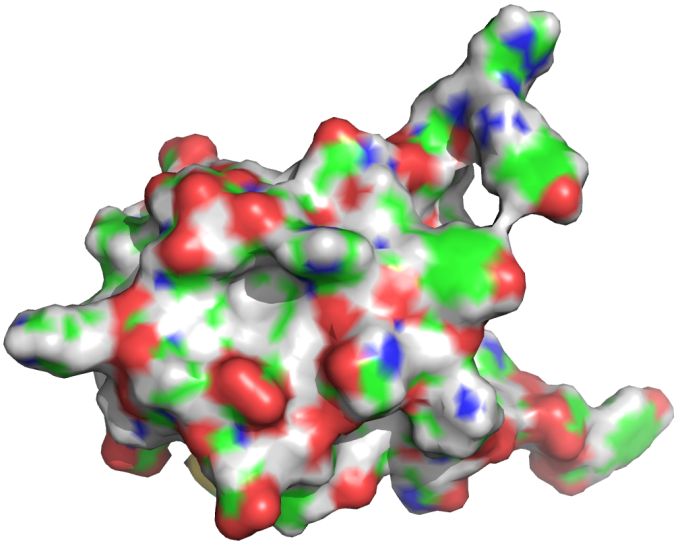


Figure 1: Surface of C-terminal domain of p73 (PDB ID: 1COK) from PDB database imaged by PyMol. Red spots represent positively charged sites, blue negatively charged ones.

Generally, folded (F) globular protein has a diameter ranging from 1 to 10 nm and increasing with cube root of protein sequence length. The full protein chain can be folded into one compact structure or into a structure composed of pseudo-independent domains. The domain - usually comprising of secondary structure motifs is connected by loops of various length. Currently there are about 60 000 structures in Protein Data Bank (PDB)[1] and it is generally accepted that determined X-ray structures keep important structural features identical to those in solution.

The structure reflects also the hierarchical principle of protein arrangement - from primary to quarternary structure. This representation of a protein - the full atom representation, is the most comprehensive among other representations. On the other hand, it is a bit misguiding. It provides a picture of a protein as atoms (represented by position and mass) connected by harmonic springs. This model of a protein lacks some important features and involves too much redundant information. The need for better models is driven by idea of simplicity at least to eliminate exhaustiveness and redundancy to reduce computational cost.

2.1.2 Features of protein models

Three main features should characterize sufficient protein model:

[1] http://www.pdb.org/

1. Space description
2. Particle representation
3. Energy function form

Let us assume that the classical molecular mechanics (MM) model usually used for proteins is the most complex one. All atom force fields (FFs) treat every atom as a mass particle with charge and van der Waals diameter. The atoms are allowed to move to infinite number of directions in a continual 3-dimensional space.

The energy functional form as well as the corresponding force constants are result of years of optimization. The force field developers are continuously solving conflict between the accuracy and the computational efficiency. The energy function is constructed as a sum of independent terms having at some extent a physical meaning. Various FFs usually differ from each other by a set of parameters but still utilize almost the same functional form[2]. Traditional approach to obtain reasonable force field parameters is a combination of high level ab initio calculations with experimentally determined vibrational constants. Parameters are further optimized to reproduce or fit some system properties (heat capacity, viscosity, relaxation times, etc). One of the most problematic FF features is their transferability. The parameters usually lack universality because the way how they were obtained is system specific. Moreover, some important interaction terms like induction are missing. It origins in ability of electron cloud to adapt to outer electric field applied. Polarizable force fields are expected to handle this inconvenience and are currently in development.

2.1.3 Simplified models

A reduction of any of the three above defined features leads to a reduction of the configurational space. Coarse grained models bind some degrees of freedom approximating some groups of atoms (mainly functional groups; specific for particular FF) to one particle. Most common way applied on proteins is to represent one AA residue by one or two spheres along the backbone. Functional forms depend crucially on the simplification expressed in a particular coarse grain model and they are the real pitfall of the method. The example of such reduction can be illustrated on the example of two phenylalanine residues. The problem of shape approximation by 2 (or 4) spheres is obvious. Coarse grained models are useful for modeling extremely large systems or for long simulation trajectories of the system or for some special uses (Vendruscolo

[2]Sometimes, special potential for improper dihedrals is added; in case of CHARMM FF, Urey-Bradley potential is added; there are also special FFs for small molecules (Merck Molecular FF) with various forms (Halgren, 1996)

and Paci, 2003). Representation of more residues by one solid particle is a ground for diffusion-collision model, that is used for PF kinetics studies (Karplus and Weaver, 1994).

Discretisation of the space leads to lattice models. Their simplicity and low computational cost allows intensive study of selected protein properties. Both, the different lattices (Yue and Dill, 1995; Dill and Chan, 1995; Ding and Dokholyan, 2005) and various particle representations can be found in the literature. An extreme of a simplification represents so called HP heteropolymer model, where only 2 types of residues are considered - hydrophobic (H) and polar (P) ones. Even this lattice model conserves the most important feature of a protein - its polymer character. The model was applied on studies of cooperativity, folding kinetics, micelle dissolution and even crystal collapse. Density of states is also discrete and provides comprehensive view on the thermodynamics model. Lattice protein models also make possible investigation of the complete fold space theoretically and even analytically (Rathore and de Pablo, 2003; Chan, 2000).

2.2 Energy terms

Empirical potential between two particles can be expressed in various functional forms. Classical chemical classification of applied forces take advantage of their physical meaning and different origin.

It must be mentioned at the very beginning that two important reductions should be made. First, many-body interactions is completely neglected so the whole model is build up using only pairwise contributions. The most accurate way how to describe interatomic potentials is a multipole expansion involving all non-covalent interactions. Based on relevant physical considerations and need of computational efficiency, the higher contributions are usually neglected and they do not appear in the final form of the particular FF.

Here is the most common functional form of the empirical force field utilized for example in AMBER type of force fields

$$U_{bonds} = \sum_{bonds} \frac{1}{2} k_b (l - l_0)^2 \tag{1}$$

$$U_{angles} = \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 \tag{2}$$

$$U_{torsions} = \sum_{torsions} \frac{1}{2} V_t [1 + \cos(n\omega - \gamma)] \tag{3}$$

$$U_{coulomb} = \sum_{i=1}^{i<N} \sum_{j=i+1}^{j<N+1} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{4}$$

$$U_{vdW} = \sum_{i=1}^{i<N} \sum_{j=i+1}^{j<N+1} 4 \cdot \varepsilon_{i,j} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{5}$$

Where k_b and k_a are force constants of bond stretching and angle bending respectively. l_0 and θ_0 are equilibrium values of bond lengths and bond angles. Torsion angle potentials are given as a sum of adequate cosines[3]. Non-bonding interactions are calculated for every pair of atoms in the system. r_{ij} is distance between atoms, σ_{ij} and ε_{ij} are calculated for every pair from particular atom parameters ε and σ. ε_0 is vacuum permittivity, q_i is partial charge of i^{th} atom.

[3] there is equivalent Ryckaert-Bellemans notation using power expansion of cosines

2.2.1 Covalent bonding interactions

Covalent interactions are described by energy terms spanning covalent bonds, bond angles and to torsion angles. Their number increases linearly with system size and represents significantly limited freedom of the system. It implies that states where harmonic potential differs significantly from the real potential are populated poorly. Parameters for bonds and angles are probably most accurate parameters in most of the force fields taking into account numerous experimental as well as theoretical studies contributing to their accuracy.

On the contrary, torsional angles potentials are usually soft and therefore energetics of protein is quite sensitive to these parameters. Moreover, they seem to be biased to prefer native structures and therefore causing bias in conformational space sampling (Best and Hummer, 2008).

2.2.2 Electrostatic interactions

Acids and bases as the first known denaturants were competently expected to change protein net charge. In the early times the electrostatic forces were supposed to account for native state stability. The first model of folded protein was a sphere of low dielectric constant in medium with higher dielectric constant (water) with charges placed on its surface. The size of electrostatic interactions was supposed to be strong enough to be the most important force mainly realized through interactions of pairs of close oppositely charged residues called salt bridges (SB). QM calculations show that contribution of this interaction calculated in vacuo can be stronger than covalent bond (!).

However, protein stability is defined by a difference between energy of native state and the denaturated state. Common view of unfolded states is such that charged residues are either fully solvated or interacting with oppositely charged residues. Energy of electrostatic interaction is according to Coulomb law proportional to reciprocal distance which seems not to change a lot by denaturation. Attractive and repulsive interactions can compensate each other at long distances and solvation of ions is usually the same in both states either. Although very strong (Dadarlat and Post, 2008) in some cases generally the electrostatics doesn't contribute to the protein stability at all.

2.2.3 Hydrogen bonds

Prevailing opinion about protein stability issue is that hydrogen bonds (HB) and hydrophobic effect are the major players in the field. It is thought that backbone hydrogen bonding accounts

significantly for stability of secondary structures. The concept of hydrogen bonding importance in proteins was first introduced by Pauling in 1936 (Mirsky and Pauling, 1936). Hydrogen bonds (HB) are usually described in all-atom force fields by electrostatic interaction of partial charges fitted especially for this purpose (Hermans, 2006). However, contribution of HB to the overall protein stability is also a subject of debate. HBs can be systematically overestimated due to their behavior in native as well as in denaturated states. On the other hand, short-range interactions are generally more selective and less resistant to higher temperatures. This implies less considerable compensation in the unfolded state.

2.2.4 van der Waals (vdW) interactions

vdW interactions are usually modeled by Lennard-Jones or Buckingham potential. It consists of 2 terms, each representing different physical nature. The repulsive term comes from Pauli exclusion principle and decreases exponentially with distance but can be modeled by potential depending r^{-n} (n>6) dependence because probability of incidence of other atom in repulsive space is in all cases negligible. This interaction is at some extent represented by atomic vdW radii. The attractive part is called London dispersion force and is proportional to r^{-6} . Dispersion interactions are usually weak compared to electrostatic interactions. However, interactions between aromatic rings are comparably strong as a hydrogen bond. The similar statement is true for interactions of aromatic residues with aliphatics side chains. Surprisingly, they are qualitatively well approximated in most of FFs. Moreover, they do not have counterparts in unfolded proteins and they are short-ranged. It implies that they are less frequent at higher temperatures and residues with high vdW interactions with other are quite evolutionary conserved.

2.3 Protein energetics

The stability of protein is defined as the difference of Gibbs free energy between the native and the denatured state of the protein. The energy corresponds to the equilibrium between functional state and odd states including unfolded state, misfolded states and molten globules lacking function. Optimum conditions for stability and function are characteristic for particular protein. Protein structure can be destroyed by various chemical denaturants or by a change of physical conditions. Stability determined by thermal denaturation is the most important characteristic for hydrophobic core (HC) concept so we will consider only thermal denaturation in our model.

2.3.1 In vitro experiments

The proteins stability can be experimentally characterized by differential scanning calorimetry (DSC). A typical calorimetric curve is shown on Figure 2 (green curve represents heat capacity change, which is subtracted from experimentally measured curve (black) to get red curve which can be integrated to get unfolding enthalpy - ΔH).

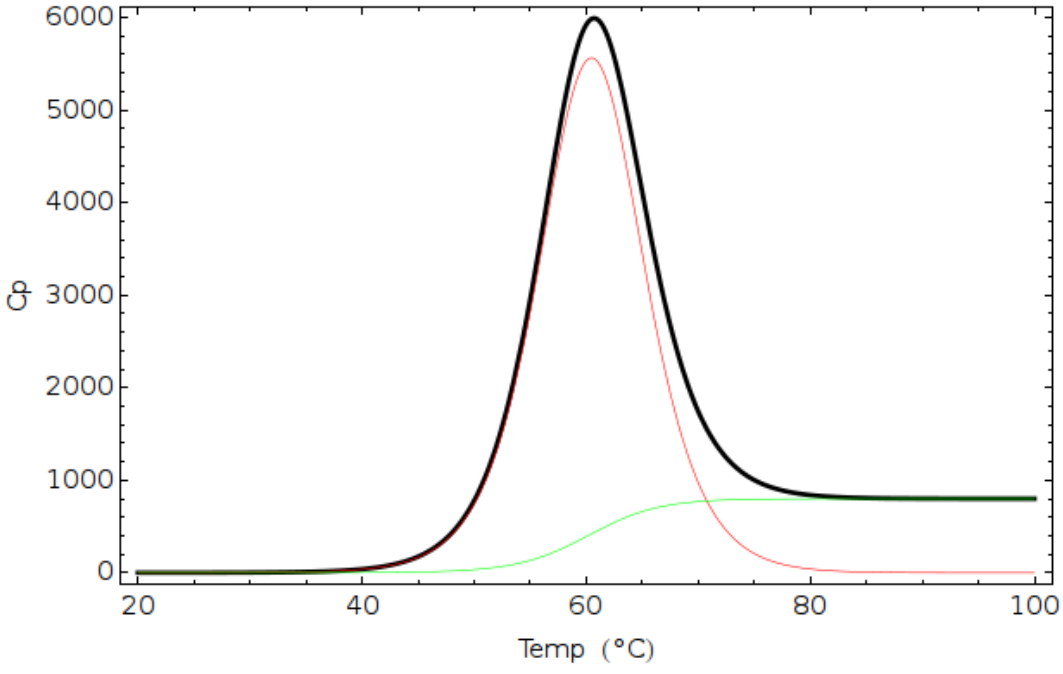


Figure 2: Calorimetric curve of protein unfolding

Three physical values are necessary and sufficient to characterize usual denaturation curve - ΔH (unfolding enthalpy), ΔC_p(unfolding heat capacity difference) and T_m (the

melting/unfolding temperature). T_m values are considered to be the most important and easily measurable and very often only they are measured as protein characteristics. There is discussion about correspondence between calorimetry determined enthalpy and van't Hoff enthalpy (Horn and Murphy, 2001) of protein unfolding. Similar curves are measured for phase transitions and it is a basis for description of protein folding and unfolding as a phase transition. Alternatively, quasi-chemical approximation of PF process is employed when studying folding kinetics.

2.3.2 Statistical mechanics of protein

Whole calorimetric curve can be obtained from partition function dependence on temperature or energy degeneration function called density of states (DOS). DOS is very important quantity and can be conceived as an one-dimensional projection of potential energy surface (PES). The quantities can be calculated as follows. PES is defined as a graph of function:

$$E = f(\vec{x}) \tag{6}$$

(energy as a function of geometry). Energy E can be replaced by potential energy U or enthalpy H depending on ensemble we apply our model. Transform of PES to DOS (in formula g(E)) can be written:

$$g(E) = |\{\vec{x}_i : f(\vec{x}_i) = E\}| \tag{7}$$

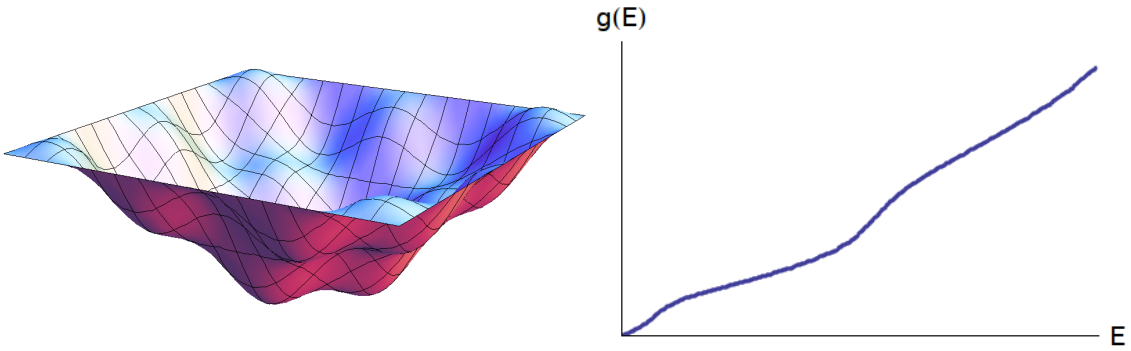


Figure 3: Density of states from simple PES

partition function is other way how to describe the systems

$$Q(T) = \int g(E) \cdot e^{-\frac{E}{kT}} dE_E \tag{8}$$

having partition function, every thermodynamic quantity can be easily calculated

$$F(T) = -kT \ln[Q(T)] \tag{9}$$

If isobaric-isothermal ensemble is sampled, free (Gibbs) enthalpy is obtained instead of free (Helmholtz) energy (equation 9). Negligible difference occurs between values of Gibbs and Helmholtz energy in condensed phase.

2.3.3 Sampling problem

The knowledge of PES is the most important characteristic of many protein models. Huge number of local minima implies high computational cost of the sampling procedure which increase exponentially with system size. Typical molecular dynamics simulation of solvated protein in explicit water lasts recently about a day for 1 ns trajectory. The most utilized sampling methods are molecular dynamics (MD), Monte Carlo (MC), Langevin dynamics and metadynamics. The most intuitive is MC method. Coordinations of a conformational space point are randomly generated and energy for the conformation is evaluated. However, steep potential causes oversampling of the highest energy states which can be overcame by utilization of Metropolis algorithm. MD evaluates forces from actual system geometry and uses them to accelerate particles in 3D space to propagate dynamics according to Newton's law. At some extent we can extrapolate the evolution of the system along the time axis[4]. According to ergodic hypothesis time averages of particular quantities for long time simulation is to ensemble averages. Central part of MD algorithm is an integrator that evaluates new positions and velocities. Langevin dynamics is similar to MD but uses artificial stochastic forces (Bussi and Parrinello, 2007). Metadynamics - the biased dynamics, returns directly free energy dependence on some chosen coordinates.

2.3.4 Solvent modeling

Natural environment of globular proteins is mainly cytosol and blood. These soups of enormous types of biomolecules, organic and inorganic ions and small molecules solvated in water. Environment is usually approximated by explicit involvement of water molecules in models with small amount of ions compensating the protein charge. Polarizable water models are only under development, so various non-polarizable water models (TIPnP, $n \in \{2, 3, 4, 5, 6\}$,

[4]However, prediction of events far in future is principally impossible according to chaos theory; the order of error of prediction depends also on integrator

SPC/E) are used as the more complex approximation. High increase in degrees of freedom which accompanies explicit solvation drives some researchers to use implicit involvement of solvent using models originating in Poisson-Boltzmann equation (Fogolari and Molinari, 2002) or generalized Born (Bashford and Case, 2000; Onufriev and Case, 2000) model with regard to cavity formation energy evaluated using surface area calculations (GB/SA).

2.3.5 Hydrophobic effect

Certainly the most mysterious phenomenon influencing the protein energetics are hydrophobic interactions (HI). Oil and water do not mix. An intuitive explanation would be simple - energy of interaction between water molecules are such strong that avoid oil molecules to interfere. Surprisingly, enthalpy measurements of hydrophobes in water show that solution[5] enthalpy is much lower in value than expected. It seems that in systems where hydrophobe is forced to interfere water (e.g. by translation entropy) water molecules prefer to form "ice berg" around the solute instead of losing HBs with each other. Enthalpy loss is then compensated by entropy increase[6]. When solute is so big that HB can not be satisfied because of geometrical reasons, hydrophobe-water interface behaves similarly as water-air interface.

HIs are subject of debates since their discovery (Hildebrand, 1979; Chalikian, 2001; Widom and Koga, 2003). Thousands of measurements have been made (Hummer *et al.*, 1998; Matulis and Bloomfield, 2001). They provide a great deal of experimental data about behavior of HIs in particular systems. HIs show strong dependence on temperature, both enthalpic and entropic terms. (Cooper, 2005). Advances in computations and MM models made possible calculations of HI at molecular level. However, simplified models can predict accurately only free energy of interaction (or solvation). Unfortunately, only the decomposition to enthalpic and entropic term is crucial for protein energetics.

HIs are not actually interactions in a classical way. We can not express the HI term similarly to the force field functional form. We can better approximate them as a result of cooperative events of all non-covalent interactions, mainly electrostatic. The only force that can be measured is potential of the mean force. Some researchers claim that the term "hydrophobic interaction" should be substituted by the term "hydrophobic effect". This highlights the problem of evaluation of the hydrophobic effect. Contemporary approach to get enthalpy term is to simulate protein in explicit solvent. Despite the know fact that implicit solvent models have

[5] I use term solvation for transfer of molecule from vacuum to solvent (water) and solution for transfer from its bulk to solvent

[6] This compensation is expected to stand behind the H-S compensation of protein unfolding (Qiu and Still, 1997)

apparent pitfalls they significantly decrease number of degrees of freedom and improve the sampling.

HIs have been thought to be the driving force for protein folding for a long time (Kauzmann, 1959). There are many reason to support this theory. First, in experimentally determined protein structures the hydrophilic residues are mostly at the protein surface and the hydrophobic residues are buried in the protein interior. Second, if free energy contribution to protein stability were additive[7] observed free energy difference between folded and unfolded state would be even lower than the calculated one. Third, denaturation of a protein structure might be described as micelle dissolution which provide some knowledge of protein unfolding.

However, thermodynamic models based on hydrophobic effect often presume that unfolded structure is almost fully extended. It was also shown that the reverse hydrophobic effect (Gromiha, 2009) based on the same physical principle also shows that simple mutation of AA residue in protein to more hydrophobic one can decrease the protein stability.

Water is therefore not "only" an environment, but also an active constituent (Ball, 2008) in PF. It is sometimes called the 21^{st} natural AA as it has perhaps more significant effect on protein structures than some AAs.

2.3.6 Enthalpic-entropic (H-S) compensation

To illustrate how challenging is the search for new models let us mention few paradoxes that accompany protein folding process. Problem of unexpectedly high rate of protein folding, which stands for funnel-like profile of PES, is called Levinthal's paradox (Zwanzig and Bagchi, 1992). New experimental data continuously upgrade our view of the protein folding problem towards a complex system behavior. One of the example is famous Privalov's puzzle published in 1979. The fact that specific entropies of protein unfolding converge to a common value when extrapolated linearly to the same temperature was explained in 1986 by liquid hydrocarbon model. It suggests that it comes from a fact that unitary entropy of transfer of hydrocarbons to water becomes zero at the same temperature (Baldwin, 1986). Second part of the puzzle was cleared (Baldwin and Muller, 1992) by splitting hydrophobic interaction into 2 terms. The most problematic issue in the protein energetics seems to be enthalpy - entropy (H-S) compensation (Sharp, 2001). The values of unfolding enthalpy and entropy are almost two orders of magnitude higher than the value of unfolding free energy and almost compensate each other. A theoretical solution to this problem was proposed by Liu, Yang and Guo (2001) but the problem for protein

[7] surely they are not fully additive (Dill, Additivity principles in biochemistry., 1997b)

energetics remains. It is very sensitive to enthalpy calculation accuracy and the idea of free energy additivity is meaningful.

2.3.7 Synthesis of energy terms

Unfortunately, it is not possible to determine generally which of the two terms is the most important in protein stabilization. But a need for such strategy is high and there are already some hypotheses conceived. However, the strategies are not mutually consistent and it is not clear which terms are stabilizing and which of them are destabilizing. One of the decomposition considering Gibbs energy additivity[8] can be illustrated by following table (Pace and Gajiwala, 1996):

Table 1: A rough estimate of the contribution of various forces to the conformational stability of RNase T1 (adapted from Pace and Gajiwala, 1996)

	Energy term	ΔG / [kcal/mol]
Destabilizing:		
	Conformational entropy:	-177
	Peptide groups buried:	-81
	Polar groups buried:	-28
	Total destabilizing:	**-286**
Stabilizing:		
	Histidine ionization:	4
	Disulfide bonds:	7
	Hydrophobic groups buried:	94
	Hydrogen bonding:	166
	Total stabilizing:	**+271**
Sum:		
	ΔG estimate:	**-15**
	ΔG experimental:	**+9**

Contribution of particular energy terms is often investigated by mutation of original AA to AA which is supposed to possess desirable properties (e.g. hydrophobicity or HB) (Guerois and Serrano, 2002). Importance of such AA or its influence on a property is based on increase or decrease of stability during defined denaturation process. Limitation of this approach is that mutations of AAs do not represent exclusively the desired property change so some other algorithm has to be applied to obtain specific contribution of particular AA.

[8] assuming independence of terms

Molecular modeling approach starts from calculation of enthalpic terms of single microstate and entropy is calculated by energy sampling as a statistical effect. The enthalpy is constructed as a sum of independent terms, so no decomposition is needed. Advances in computational modeling methods make theory end experiment closer and which helps to apply the method widely.

Native structure represents only a narrow range of well defined states that are important for function and whose population is definitely lower than the number of all possible states. The ratio is in orders of 10^{30} to 10^{300} depending on molecule size, so entropy of unfolded state is much higher than that of folded state. The corresponding enthalpy of unfolding is in order of hundreds to thousands kJ/mol. Simple theoretical comparison of unfolded and folded energy states is not possible mainly because we do not know how to represent structures of the denatured state (Shortle, 1996; Mittag and Forman-Kay, 2007). We aim to propose an appropriate model for unfolded state that would come from folded structure or molecular simulations.

Synthesis of the energy terms and their assignment is complicated because of the following problems:

1. Counter-effect of interactions receded into unfavorable positions in order to enable stronger interactions
2. Compensation of interactions that are conserved both in folded and unfolded form (Native contacts)
3. Enthalpic-entropic compensation

These effects together with inaccurate description of torsion and electrostatics contributions (discussed above) are most troublesome for successful match of the theory and experiments.

Unfolding enthalpy is often correlated with change of solvent accessible surface (SAS) of AA residues (Robertson and Murphy, 1997). Hydrophobic effect is supposed to contribute approximately by 8 kJ/mol per residue to free energy of unfolding (Makhatadze and Privalov, 1995; Honig and Yang, 1995). On the other hand, folding enthalpy does not necessarily origin in solvent - protein interactions. The higher the value of contact surface, the higher the dispersion interactions between residues. PF might be kind of solid-fluid energy transition of hydrophobic residues. Expecially aromatic residues are known to possess high SC-SC IE (Kannan and Vishveshwara, 2000; Vondrášek and Hobza 2005). Value of benzene fusion enthalpy (about 10 kJ/mol) supports this concept.

We need an appropriate computational representation of interaction energy in proteins. One adequate method is a characteristics of intramolecular energy of proteins represented by interaction energy matrix.

2.4 Interaction energy matrix (IEM)

2.4.1 Contact matrix

Contact matrices (CM) or contact maps are symmetric boolean N×N matrices matrices describing contact for every amino acid pair in protein of N residues. They represent a protein structure as a graph where number a_{ij} in matrix represents an edge between vertices i and j being is 1 if there is a contact based on some contact definition, otherwise 0. Many definitions of contacts have been proposed. They are mainly based on interatomic distance of special atoms, on all atoms of residues or on contact area.

It has been shown (Vendruscolo, Najmanovich and Domany, 1999; Vendruscolo and Lebowitz 1999) that CM representation contains less redundant information than 3D structure representation and the change from CM representation to 3D coordinates representation is a solved problem (Vendruscolo and Domany, 1997; Vassura and Casadio, 2008). CM transform PFP from image of one-dimensional discrete sequence space to 3-dimensional fold space into image of 1D fold space to space of square matrices of corresponding order. There are even attempts to represent any protein structure by one-dimensional information - eigenvector of contact matrix (Kabakçioğlu and Domany, 2002; Porto and Vendruscolo, 2004; Bastolla and Vendruscolo, 2005; Miyazawa and Kinjo, 2008).

The energy of particular conformation is commonly represented by pairwise contact potentials:

$$E(\mathbf{C},\mathbf{S}) = \sum_{i=1}^{N}\sum_{j>i}^{N} \mathbf{C}_{ij}\mathbf{U}(\mathbf{S}_i,\mathbf{S}_j) \tag{10}$$

Where $\mathbf{C}$ is contact matrix, $\mathbf{S}$ is protein sequence, $\mathbf{U}$ is definition of pair potential, E is energy of state characterized by $\mathbf{C}$ and $\mathbf{S}$. Let our protein model have M different types of amino acids and our protein consists of N amino acids. $\mathbf{U}$ is a symmetric matrix of size $M \times M$. $\mathbf{S}$ is a vector of size N and there is M^N possible sequences. This potential has been investigated and successfully used in lattice models but it has obvious limitations to real proteins.

Graphs with weighted edges can be represented by corresponding matrices with real numbers instead of boolean values to get distance matrices or interaction energy matrices. They contain much more information and can overcome problems rising from discrete potential.

2.4.2 Interaction energy matrix (IEM)

IEM is a symmetrical matrix (size of N) of real numbers representing IEs obtained with a particular method. It stands for real energetic circumstances in proteins. Every value depends on positions of all atoms of particular interacting residues pair. Despite the theory for IEM space properties is still developing IEMs have already found remarkable application in protein science. Use of physical interresidual interaction energy has been successfully used to find key residues in proteins (Bendová-Biedermannová and Vondrášek, 2008).

There are many ways how to construct an IEM depending on property one wishes to mine. Protein has to be fractionated into residues appropriately represented by sets of atoms. Then interaction energy for every pair is calculated. Various residue representations have been used depending on investigated properties. Computational methods also vary from very expensive QM to empirical force field calculations.

2.5 Hydrophobic core

Many definitions of hydrophobic core can be found in literature. HC concept is usually challenged by the mapping of the following aspects:

1. **Thermodynamic** aspect - focused on the contribution of HC to protein stability.
2. **Structural** aspect - focused on HC-protein structure relationship. A part of this concept is that the whole protein structure might be determined by composition and architecture of the HC.
3. Nucleation event - it is supposed that nucleation is the most important and earl;iest event in the process of PF. Its composition tightly connected with HC has high impact on **kinetics** and a foldability of protein.
4. HC amino acid compositions are highly conserved during the **evolution**. This makes possible to evaluate simultaneously the importance of the sequence, fold and the function.
5. Biological importance

All the aspects having influence on the definition of HC can be briefly summarized in two major categories. The first category defines HC as all hydrophobic residues in protein interior. The second category defines HC as a set of residues forming spatial unit of protein where residues interact substantially higher with each other than in the rest of the protein.

2.5.1 Interior of globular proteins

Well documented feature of globular proteins is HC composed of hydrophobic residues and exposition of charged and polar groups on the surface. Statistically, there is about 40% of hydrophobic residues in globular proteins. Expected size of a protein interior increases with domain size and reaches about 40% for medium-sized domains. Packing density of protein interiors is similar to that in organic crystals (Tsai and Gerstein, 1999).

The question is what drives a protein into the ordered form? Hydrophobic effect which is supposed to be the main player in PF provided the adjective of the hydrophobic core conception. Because the hydrophobic residues almost exclusively present in a core, it is supposed that reluctance of water to be in contact with these residues is at least a consequence of water affinity towards the polar and charged residues on the surface. On the other hand, it could be the effective packing and high interaction energy between side chains of the HC composing amino acids which keeps the core compact. It was observed that HC of thermophilic proteins

differs from their mesophilic counterparts mainly involving larger or more numerous aromatic residues in core.

2.5.2 Key residues and residues with strongest interaction energy

Generally, residues having important role in proteins are usually called key residues. They can hold important structural information, perform enzymatic function, contribute much to protein stability or lower kinetic barrier of folding to native state. Study of relations and overlapping of key residues of different definitions attempts to shed light to theory behind their properties. In this work, we study protein energetics, so we try to develop a method of identification residues with highest importance for protein energetics and estimate their importance.

The identification of key residues is usually done by mutational studies, distance-based contact theory, Gaussian network model analysis (Bahar *et al.*, 1997; Haliloglu *et al.*, 1997) and lattice simulations (Prudhomme and Chomilier, 2009). Identification of the key residues using IEM analysis was first proposed by Chen and Xiao (2009). Use of IEM faces problems of computational cost and low accuracy of IE evaluation. To our best knowledge, none have performed IEM based key residues identification on a representative set of structures.

3 Methods

3.1 Representative structure set selection

Throughout the whole study we use sets of X-ray and NMR determined protein structures from PDB database. Representative sets of structures for the study were chosen in order to represent complete structure diversity of the database. Structures had to meet the following criteria:

1. Structures that contain only protein molecules (no DNA/RNA)
2. No ligands in structure
3. Structures of maximum 2 protein chains
4. Removed structures having sequence identity higher than 95 %.
5. pH of the crystallization - lower than 7.6
6. Only mainly α or mainly β proteins according to CATH[9] classification were taken into the sets
7. Size limit - Structures containing more than 1700 heavy atoms were removed.

Only small to medium sized proteins were considered in this study. We intended to take into account a wide variety of sequences. Structures resolved at pH higher than 7.6 were omitted because of problems with histidine protonation.

Calculations on 2 distinctive structural sets give us opportunity to investigate secondary structure content as an effector of the protein behavior. It also makes possible to establish new characterization of secondary-tertiary structure relationship based on energy balance. The selection resulted in 618 structures with mainly α secondary motifs and 833 structures with that of mainly β.

3.2 Structure preparation

Downloaded structures from database were checked for including any incomplete amino acid which were removed. In case of more structures of the same protein only the first one was used. Only side-chain heavy atoms geometries were extracted and missing hydrogen were added according to Cα FF concept and then optimized.

[9] http://www.cathdb.info/

For optimization we used GROMACS molecular simulation package with steepest descent algorithm employed.

Algorithm 1 Steepest descent

Input: starting configuration $r_i(0)$
starting maximum displacement $h_i(0)$ set to 0.01 nm
1. calculation of potential gradient (forces) for every particle
2. calculation of new positions:

$$\mathbf{r}_i(n+1) = \mathbf{r}_i(n) + \frac{\mathbf{F}_i(n)}{\max(|\mathbf{F}_i(n)|)} h_i(n)$$

here $\mathbf{F}_i(n)$ denotes force applying on i^{th} particle at n^{th} step, $\max(|\mathbf{F}_i(n)|)$ means the largest of the absolute values of the force components, $\mathbf{r}_i(n)$ is position of i^{th} particle at n^{th} step
3. if energy was decreased, $h_i(n+1) = 1.2h_i(n)$ else $h_i(n+1) = 0.2h_i(n)$
4. continue from 1. until converged

3.3 IEM calculation

IEMs were calculated using parm03 empirical potential (Duan and Kollman, 2003) with following modifications. Backbone atoms of each residues were replaced by a methyl group on Cα carbon atom. It has been proved that interaction energies calculated using this potential are very close to that calculated using benchmark QM calculations (Berka and Vondrášek, 2009).

Molecule was fractionized into side-chains of residues represented by corresponding molecules. For every residues pair A-B, list of all atom pairs (bipartite graph) - one belonging to residue A, the other to the B and only non-bonding energy terms were evaluated according to Equations. 4 and 5 for each pair. The atom pairwise contributions were summed up to provide IE value for a residue pair. Home made Python program was employed for this purpose utilizing basic libraries only.

3.4 Surface area calculation

To find surface residues we used algorithm proposed by Conolly (1983) implemented in GROMACS package as "g_sas" routine. Probe was set to 1.6 Å, residue was considered to be on the surface if the SAS was more than 20% of whole surface.

3.5 Basic statistics

3.5.1 Least squares

Slope of linear function from 2-dimensional data points $[x_j, y_j]$ can be evaluated as (11):

$$S = \frac{n\sum_j x_j y_j - \sum_j x_j \sum_j y_j}{n\sum_j x_j^2 - \left(\sum_j x_j\right)^2} \tag{11}$$

3.5.2 Correlation coefficient

Correlation of 2 related data sets can be characterized by correlation coefficient calculated as follows (12):

$$\rho(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{D(x)}\sqrt{D(y)}} \tag{12}$$

where $\mathrm{cov}(x,y)$ is covariance of quantities x and y calculated according to Equation 14 and $D(x)$ is standard deviation of x calculated according to Equation 13.

$$D(x) = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}} \tag{13}$$

$$\mathrm{cov}(x,y) = \sum_i \frac{(x_i - \bar{x})(x_i - \bar{x})}{n} \tag{14}$$

We evaluate $\bar{x}$ (mean value) as the arithmetic average of x_i; n is the number of data.

$$\bar{x} = \sum_i \frac{x_i}{n}$$

4 Results and discussion

Final aim of the proposed work is a definition of the method leading to HC identification. We use set of energy based analyses of protein structures to set up this method on reasonable physical basis.

4.1 Distributions of IEs in proteins

Understanding HC energetics requires knowledge of energy proportions of relevant residues in proteins. The question is if all residues contribute similarly to the stability of the protein. We expect that classification of residues according their sum of IEs with the others we could determine an importance of particular residue.

Calculation of high number of structures allows statistical processing of obtained data (with reasonable convergence of averages) and therefore solid base for conclusions. We calculate IEM for each structure to characterize fully the protein belonging to a set. To characterize SC-SC interactions in proteins, we plotted distributions of IEs in all proteins of the set. First group of distributions characterizes energetic importance of particular residues in one protein. The distributions provide valuable information about internal energetic proportions of proteins. The second group characterizes SC-SC interactions in all proteins and provide an estimate of general feature of all proteins.

4.1.1 IEs in proteins

We face serious problem caused by overestimation of the electrostatic term in calculated IEM. Simple screening by dielectric constant is not the correct solution to the problem. Moreover, if we focus on the HC there are no charged residues at all. Hence, we take only uncharged residues into account.

Total stabilization of one residue (residue IE - RIE) can be obtained summing all values in particular row / column of IEM. The procedure sorting residues according their RIE provides a contribution of SC to the stability of the folded state. The graph in Figure 4 illustrates distribution of RIEs for a mutant of FADD death affector domain protein (PDB ID 1A1W).

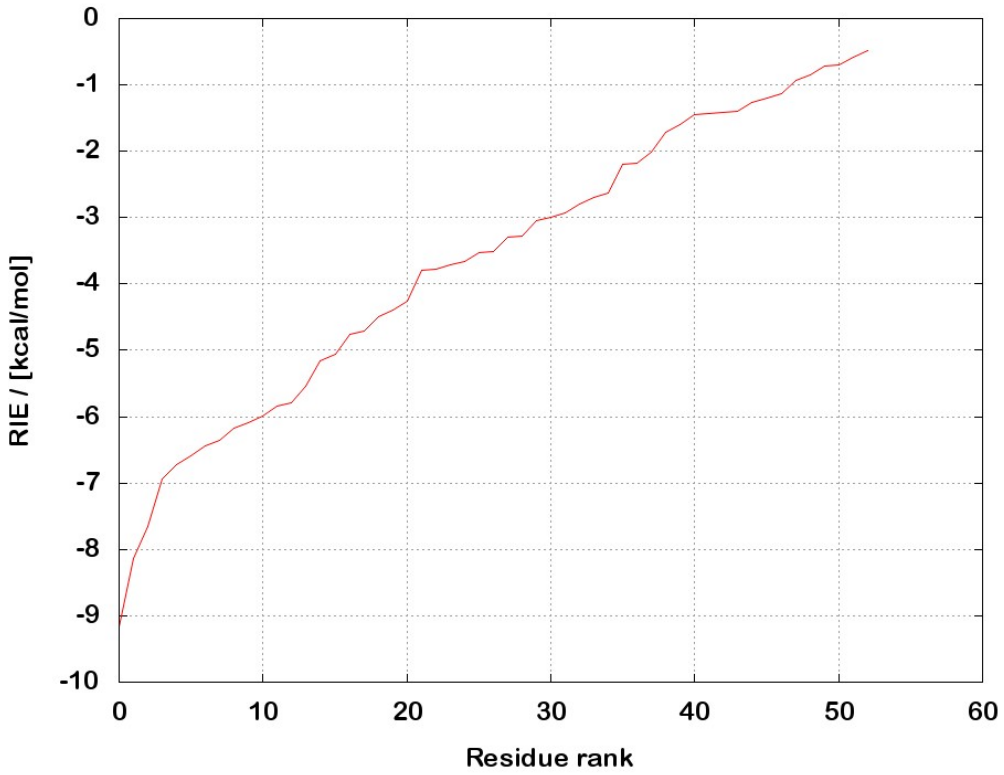


Figure 4: Distribution of RIE in F25Y mutant of FADD death affector domain (PDB ID 1A1W)

It is apparent the decrease of residues with higher stabilization energy. 37 % of residues with highest RIE contribute to the overall energy by 63 % of the total stabilization. First N residues are the key residues because they contain considerable part of the total energy.

Roughness of the curve is because of the small number of plotted residues. We can determine an average of the distributions for all proteins in our 2 sets (mainly α, mainly β). We obtained the results shown on Figure 5.

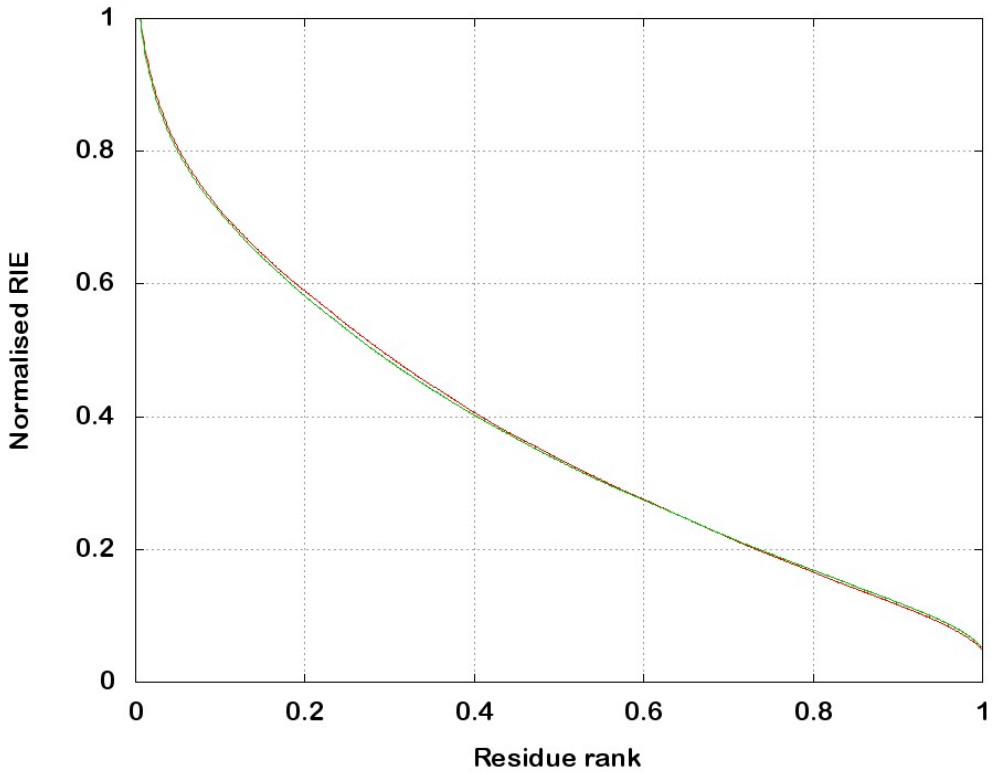


Figure 5: Normalized average RIE distributions for mainly α and mainly β sets

Difference between the curves is negligible. They can be fitted by 6-parametric function - sum of two exponentials and a linear function (15) and the expression fits the experimental data extraordinary well.

$$E(r) = 0.173147 \cdot e^{-44.1123 \cdot r} + 0.355419 \cdot e^{-3.83228 \cdot r} - 0.457555 \cdot r + 0.514057 \qquad (15)$$

We suppose that each component of distribution has its physical meaning whose revelation remains for future investigations (see Figure 6).

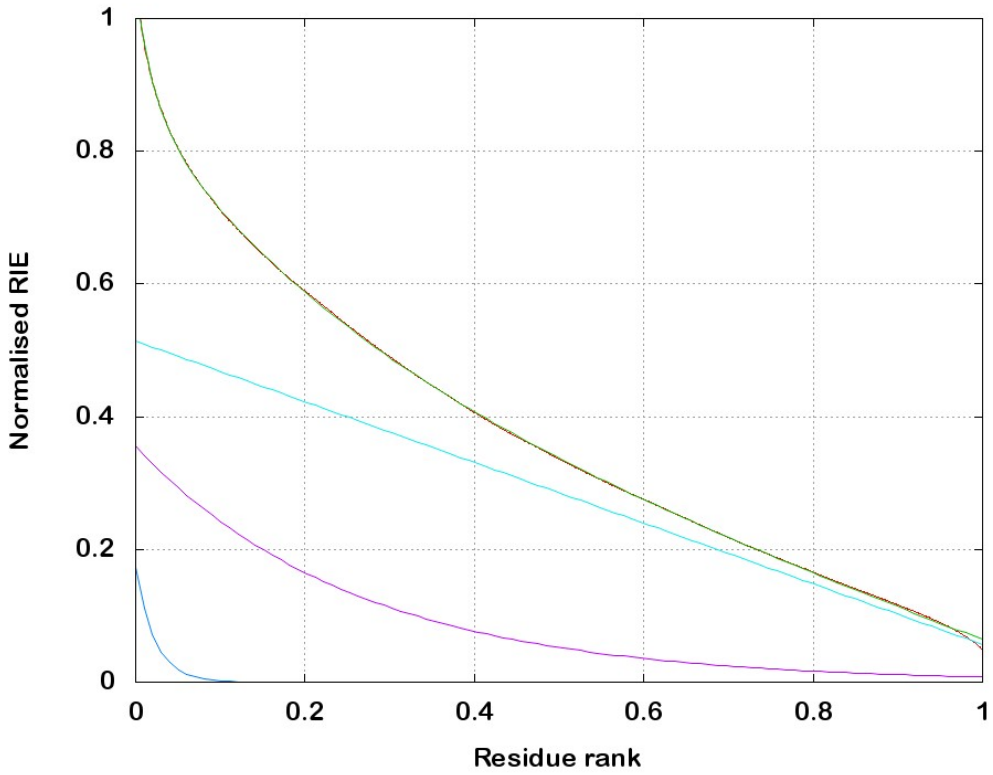


Figure 6: Normalized average distribution of representative set of proteins with mainly α secondary structure and its decomposition into three terms. Experimental data are denoted red, the fit is green color. Light blue line best fits the right sector of experimental curve. Middle region is represented by sum of light blue and magenta curve. The far left blue curve represents low occurrence region of residues with highest RIE.

4.1.2 IEs for particular AA pairs

To identify important IE region for specific AA pair, it is useful to know the specific pair distribution in all proteins. The knowledge of the mean or the median IE for such distribution is a characteristic helpful to explain particular importance of the residue pair. The identified key residues are supposed to bind stronger than the rest of AA of the same kind in a protein. Hence, we need an estimate of IE of medium contact.

We can plot distributions of IEs for every AA pair. Each one of 210 distributions (for every representative set) was constructed as a specific interaction of the studied AA pair from all IEMs in the protein set. Contact for interacting AA was defined by IE threshold -0.5 kcal/mol (see next chapter). Distribution of IEs of phenylalanine in contact with all 20 AA residues for set of all mainly α protein is on Figure 7.

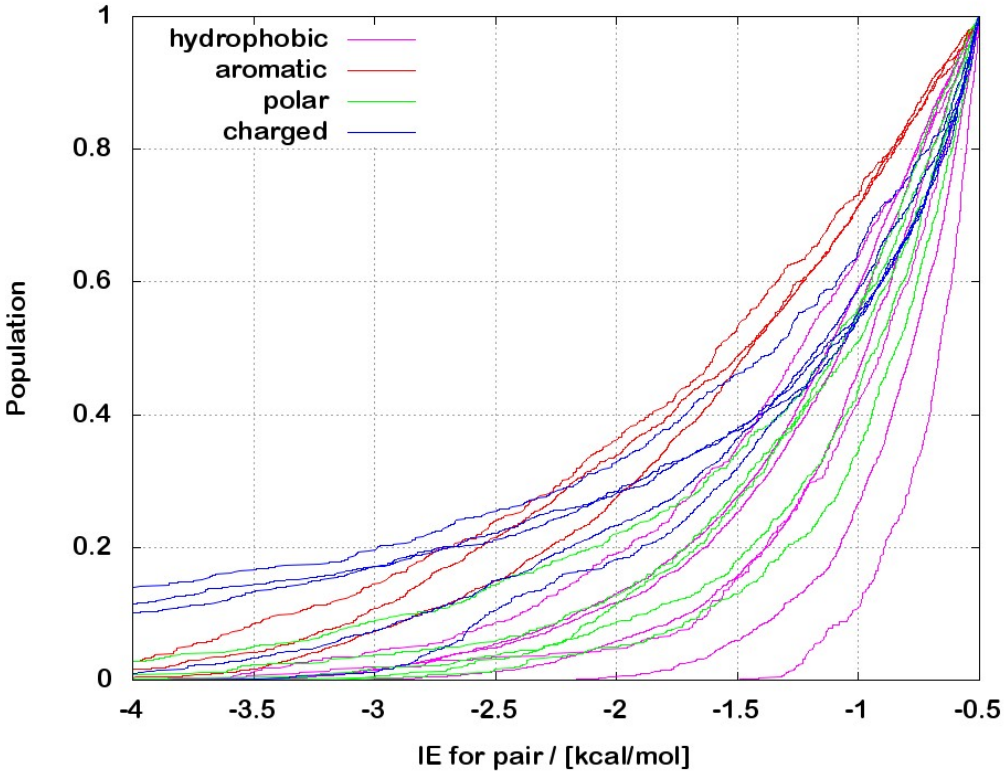


Figure 7: Distribution function of pairwise IEs of PHE with other residues. Hydrophobic residues: ALA, GLY, VAL, LEU, ILE, CYS, MET. Aromatic residues: PHE, TYR, TRP. Polar residues: SER, THR, PRO, ASN, GLN. Charged residues: HIS, LYS, ARG, ASP, GLU.

The shape of the distribution (Figure 7, compare with Trebbi and Zerbetto, 2005) depends on character of the contact and the curves can intersect each other. Nevertheless, curves

representing interactions of same type of AA do not intersect each other. It means that they are only shifted according to interaction strength without changing the shape.

The medians tell us about ability of specific AA interaction to stabilize a protein. We use them to introduce a concept of relative residue interaction energy (RRIE). It is defined as a sum of IE of all contacts for a specific residue from which is IE of mean contacts subtracted. This value represents the efficiency of residue stabilization in protein structure.

Symmetric 20×20 matrices of medians obtained for α an β set (Tables 2 and 3) show that the strongest interactions (with exception of the charged residues interaction) in proteins come from aromatic-aromatic contacts. The values for polar residues show that only small number of them are connected by HBs. The values in the matrices also support a previous findings that TRP-PRO interactions belong to strongest ones (Biedermannova *et al.*, 2008). Surprisingly, the biggest difference between studied sets of proteins comes from CYS-TRP interaction which is very selective and shows stabilization only if sulphur interacts with π electrons of TRP (Figure 8).

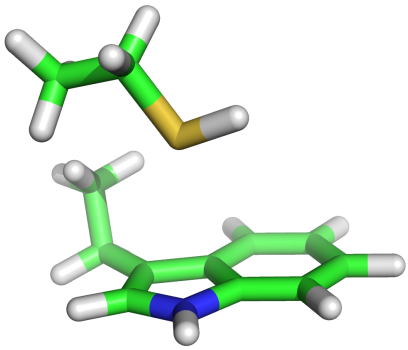


Figure 8: Strong TRP-CYS interaction

By the comparison of the matrices we can conclude that there is no preference in side-chains to form particular type of secondary structure.

These matrices can be compared to those proposed by Jernigan and Miyazawa (1996). Their pairwise potentials represent free energies of contacts and were determined by statistical processing of contact occurrence and are successfully used for lattice simulations. In contrast to their results, we identified apparent preferences for aromatic residues. The difference between values in both representation might account for hydrophobicity and compensation contributions.

Table 2: Median IE of residue pairs in proteins with mainly α secondary structure in kcal/mol.

	ASP	GLU	ARG	LYS	HIS	GLN	ASN	THR	SER	MET	CYS	ALA	GLY	VAL	LEU	ILE	PRO	TRP	PHE	TYR
ASP	n	n	9.40	9.33	6.45	0.91	0.90	0.92	0.94	0.81	0.83	0.88	0.62	0.86	1.04	0.90	0.99	0.98	1.13	0.97
GLU		n	9.20	9.15	6.16	0.91	0.89	0.95	0.94	0.84	0.81	0.86	0.62	0.84	1.05	1.00	1.07	1.01	1.29	1.02
ARG			n	n	n	0.96	0.96	0.94	0.98	0.89	2.35	0.83	0.66	1.18	1.02	0.93	1.15	0.97	1.03	1.04
LYS				n	n	0.91	0.96	0.89	0.91	0.85	2.17	0.80	0.63	1.03	0.92	0.86	1.05	0.93	0.87	0.92
HIS					n	0.87	0.90	0.93	0.93	0.89	1.47	0.83	0.64	1.16	1.06	1.15	1.11	0.99	1.19	1.26
GLN						1.21	1.36	1.14	1.06	0.99	1.01	0.75	0.63	0.99	1.08	1.01	1.01	1.51	1.25	1.28
ASN							1.31	1.03	1.02	1.08	1.04	0.71	0.61	0.86	0.94	0.92	0.97	1.25	1.00	1.15
THR								0.83	0.89	0.93	0.89	0.67	0.60	0.82	0.83	0.84	0.93	1.13	0.95	1.03
SER									0.89	0.87	0.85	0.63	0.55	0.75	0.78	0.76	0.85	1.18	0.89	0.94
MET										1.10	0.96	0.72	0.61	0.94	1.04	1.03	0.93	1.38	1.34	1.29
CYS											1.02	0.65	0.58	0.76	0.85	0.84	0.83	0.99	0.93	1.03
ALA												0.61	0.52	0.68	0.71	0.70	0.71	0.85	0.76	0.82
GLY													0.50	0.62	0.60	0.64	0.59	0.73	0.68	0.67
VAL														0.81	0.93	0.90	0.88	1.15	1.01	1.09
LEU															1.03	1.01	0.97	1.29	1.17	1.25
ILE																1.00	0.99	1.27	1.16	1.17
PRO																	0.90	1.57	1.24	1.30
TRP																		1.33	1.66	1.50
PHE																			1.45	1.47
TYR																				1.56

Table 3: Median IE of residue pairs in proteins with mainly β secondary structure in kcal/mol.

	ASP	GLU	ARG	LYS	HIS	GLN	ASN	THR	SER	MET	CYS	ALA	GLY	VAL	LEU	ILE	PRO	TRP	PHE	TYR
ASP	n	n	9.27	9.29	6.39	0.91	0.90	0.90	0.94	0.86	0.87	0.89	0.61	0.95	1.03	0.97	0.92	0.86	1.10	0.98
GLU		n	9.02	9.19	6.24	0.91	0.93	0.98	0.96	0.84	0.78	0.89	0.62	0.83	1.11	1.04	0.94	1.03	1.24	1.02
ARG			n	n	n	0.93	0.95	0.88	0.90	0.91	2.24	0.81	0.67	1.24	0.99	0.91	1.05	0.94	1.05	1.12
LYS				n	n	0.90	0.96	0.88	0.90	0.83	2.19	0.77	0.62	1.12	0.89	0.90	0.91	0.85	0.95	1.00
HIS					n	0.86	0.98	0.91	0.89	0.93	1.50	0.82	0.66	1.08	1.02	1.11	1.24	1.13	1.13	1.22
GLN						1.25	1.25	1.17	1.09	1.07	0.85	0.78	0.66	0.96	1.03	1.05	1.04	1.27	1.14	1.19
ASN							1.24	1.11	1.05	0.95	0.81	0.73	0.63	0.89	0.96	0.90	1.03	1.14	1.06	1.16
THR								0.82	0.88	0.93	0.85	0.67	0.59	0.81	0.85	0.80	0.90	0.94	0.89	0.97
SER									0.89	0.87	0.97	0.65	0.56	0.78	0.80	0.76	0.85	1.06	0.87	0.94
MET										1.12	0.98	0.79	0.65	0.92	1.10	1.05	1.00	1.36	1.27	1.22
CYS											0.93	0.71	0.56	0.77	0.95	0.87	0.80	1.93	0.93	0.95
ALA												0.60	0.53	0.70	0.72	0.71	0.71	0.81	0.78	0.83
GLY													0.50	0.60	0.62	0.61	0.60	0.79	0.69	0.76
VAL														0.82	0.93	0.89	0.83	1.10	0.99	1.02
LEU															1.02	0.99	0.99	1.32	1.08	1.18
ILE																0.98	0.95	1.17	1.11	1.09
PRO																	0.94	1.62	1.21	1.41
TRP																		1.53	1.32	1.45
PHE																			1.31	1.30
TYR																				1.46

4.2 Contact definition based on IEM concept

By evaluation of IEM we shift from representation of protein in Euclidean space to interaction energy space. Residues with strong IE are close to each other in the energy space, so clustering algorithm can be applied. Similarly, transformation of IEM to CM requires a threshold of a measure in energy space.

Contact order is a quantity characterizing importance of particular residue contribution to the compactness and stability of the whole protein. It can be obtained from contact matrix by summation of all contacts in particular row / column. It is closely related to its counterpart in IEM - RIE. Here we show that definition of contact is robust and we can determine its optimum value reliably.

4.2.1 Residue interaction energy and contact order

IE per CO is higher than the contact definition and therefore each contact contributes to RIE by IE at least by the contact definition value. To determine, how the average contact contributes to protein stability, we plotted CO against RIE for every protein in both sets (instance on Figure 9).

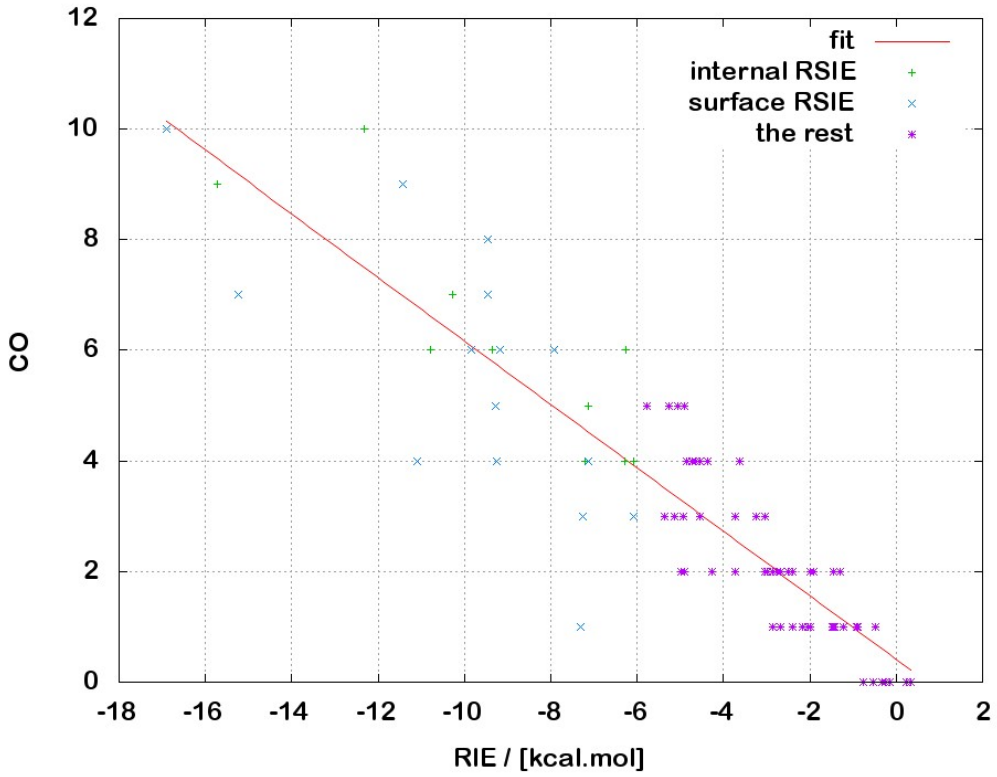


Figure 9: CO as a function of IE for Barnase, PDB ID 1BNI. Residues that belong to upper 37 % sorted according to RIE are imprinted with green color if they are situated in protein interior, blue if they are on surface. Other 63 % of residues are denoted purple.

We found that there is linear dependence between CO and RIE. Correlation coefficient is higher than 0.7 for all graphs. Inaccuracies come partially from discrete nature of CO. Residues bound by hydrogen bonds have generally lower CO and can be found at the bottom left part of the data chart. We evaluate the slope by least squares method (Equation 11).

4.2.2 Optimum contact definition

There is an interesting fact of the linear dependence between CO and RIE which very conserved among the residues. We tried to find contact definition providing highest conservation of slopes of CO-RIE dependence that would be robust and transferable. CM for each protein for particular contact definition was constructed and CO was plotted against RIE. Statistical processing of all the slopes led to standard deviation of the slopes (Equation 13) and represents the measure of their transferability.

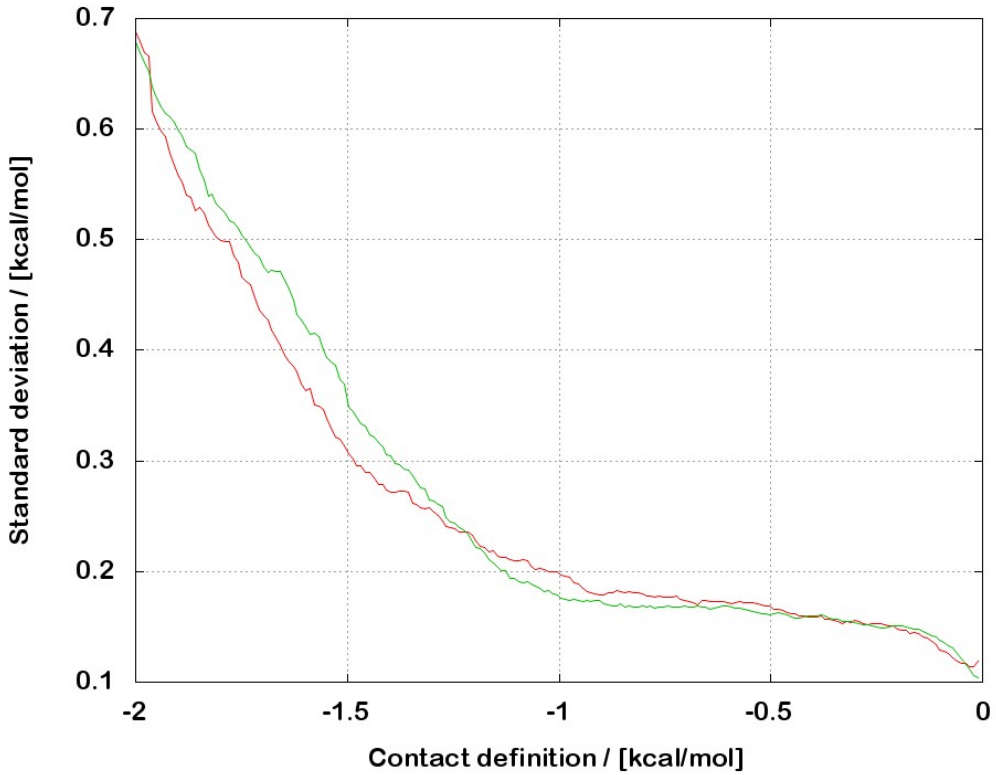


Figure 10: Transferability of CO per residue represented by the standard deviation of slopes depending on contact definition. Curve for α set is denoted red, β set green.

Figure 10 shows that energy of contact definition between 0 and -1 kcal/mol is transferable. Small shift in the contact definition affects equally average RIE per CO for all proteins. The average of correlation coefficient of IE-CO dependence versus contact definition was plotted. (Figure 11).

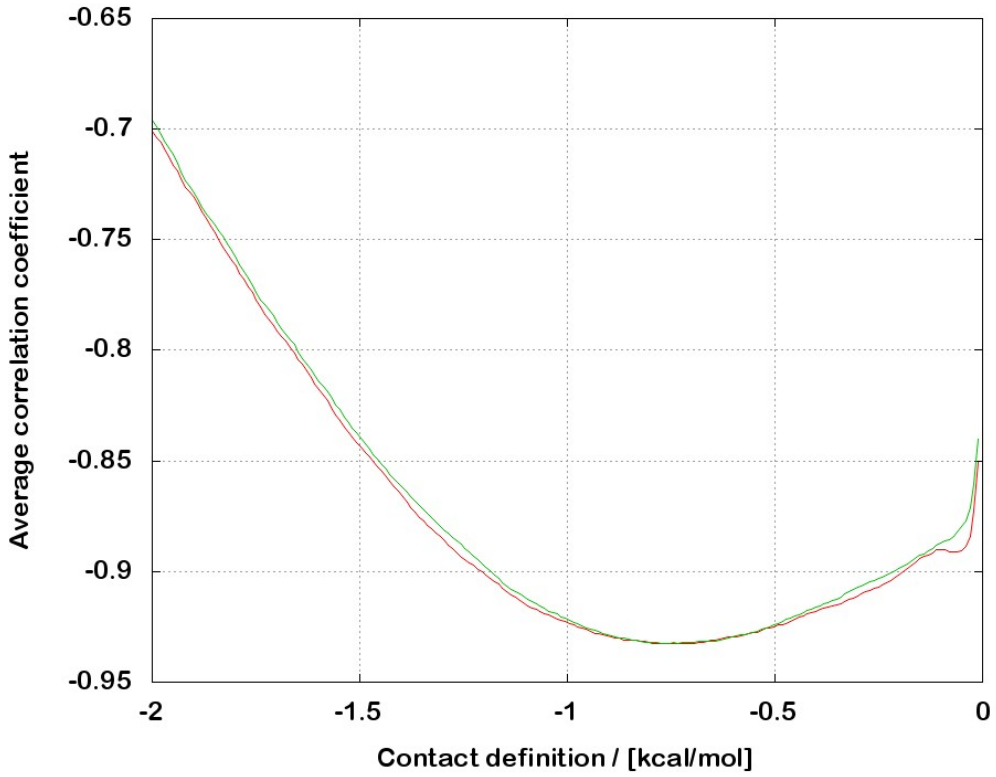


Figure 11: Normalized standard deviation of slopes dependence on contact definition. Curve for α set is denoted red, β set green.

There is a shallow minimum in the graph (best correlation) for interval (-1.0 , -0.5) kcal/mol. This value is very close to energy of 2 degrees of freedom (kT) at 300 K, which is approximately 0.6 kcal/mol. Robustness of contact definition can be characterized by plotting of average CO-IE dependence slope against contact definition (Figure 12). We can identify a steady point for values around -0.5 kcal/mol.

To synthesize the data we compromise between the best transferability at 0 kcal/mol, the best correlation at about -0.7 kcal/mol and the best robustness at -0.5 kcal/mol. Contact definition at 0 kcal/mol does not provide any information at all and the transferability slowly decreases with lowering the threshold. The correlation does not change much around -0.7 kcal/mol. High robustness of the method prevails over correlation and finally, we recommend **-0.5 kcal/mol** as the best contact definition in IEM approach.

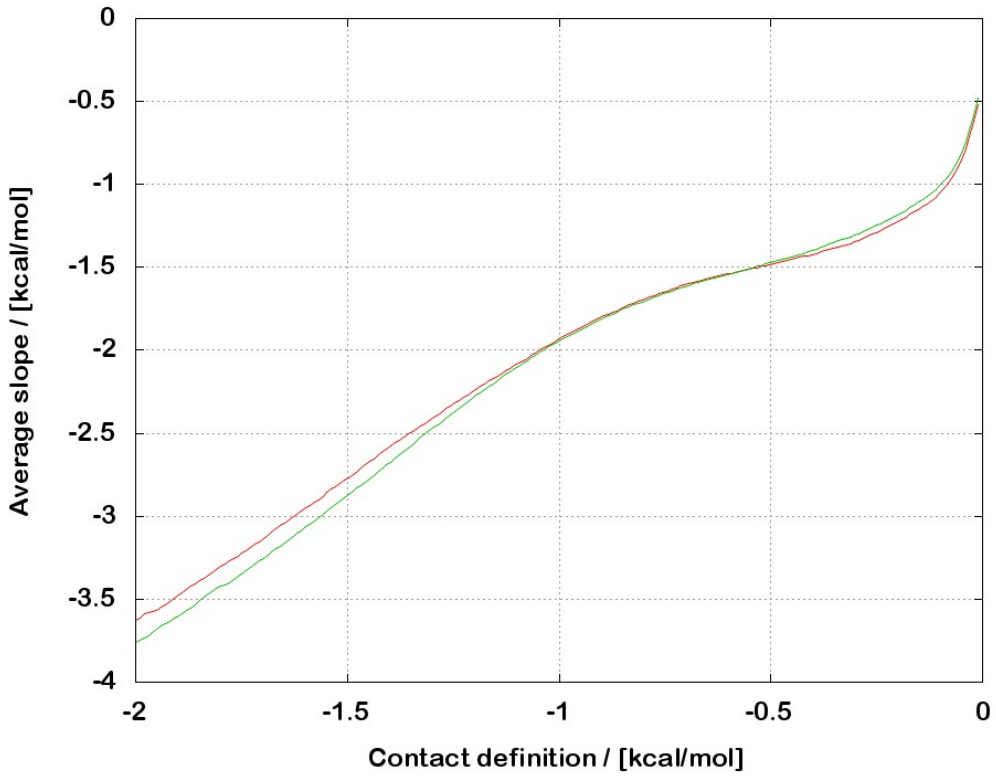


Figure 12: Plot of average interaction energy increase with contact against contact definition. Curve for α set is denoted red, β set green. Closenes of the curves corresponds to the minimum in the previous graph. Highest robustness of the contact definition is represented by steady point at conact definition at -0.5 kcal/mol.

4.3 Core definition

There is no exact and transferable definition of hydrophobic core. We understand a core as a set of strongly bound residues which significantly contribute to overall IE and have high impact on protein structure and energetics. Development of convenient methods for HC identification is one of the results of this work. In this chapter, we present four different methods which can be applied to find key residues in protein structure having importance in localization of HC. All the methods are described in details below.

4.3.1 Using RIE, CO and RRIE

Formerly introduced residue energetic quantities (RIE, CO, RRIE) can help us to measure an energetic contribution of particular residue to protein stability. High RIE does not mean that residue necessarily increases the protein stability significantly. Energy can come from one strong hydrogen bond which can be also compensated in the unfolded states. CO does not say anything about energy importance of the contacts and if the contact definition is too vague it becomes only a dependent on level of exposure. RRIE tells us how strong is particular interaction compared to other interactions of this type. However, some contacts are generally stronger than others by definition. Our methods are based on sorting residues in protein according to CO, RIE and RRIE. Critical step is to choose appropriate number of strongest residues and we present a few methods to overcome this problem.

4.3.2 Number of residues in the HC

To reliably identify HC we need a method that provides a number of residues which is transferable, robust and limited. Graph clustering method yields exact number of residues, yet there is no single dependence of the number on a protein length. Unfolding enthalpy, sum of interaction energies in proteins, number of all contact, surface area and other important energy characteristics depend on the protein length. We consider dependence of number of core residues on the protein size. The final goal of this work is our proposal of 4 methods for HC identification described below.

1. **Fixed number of key residues.** It has been recently proved that identification of limited number of key residues is sufficient for forming of a protein structure (Vendruscolo, Paci, Dobson and Karplus, Three key residues form a critical contact network in a protein folding transition state, 2001b). It has been shown that strength of interactions between few key residues and their neighbors are the stability determinants of proteins. They might be used to calculate stability of mutants and evaluation of their free energy difference compared to the wild type. This method is size-independent and provides values within a range of 1-20 kcal/mol.

2. **Fixed ratio between key and all residues**. As we showed earlier, distribution of RIE is conserved among proteins. If we define key residues by this way the steep exponential function component must characterize them mainly. Sum of RIE, CO or RRIE for these residues increases linearly with protein size and can be correlated to quantities with this property. Residues are sorted according to residue energetic quantities and only first M residues are considered. The number M is proportional to protein length N and is defined as $M = r \cdot N$. Choosing proper r will be described below.

3. **Surface area exclusion**. Supposing that definition of the HC residues accept a condition that the residue must be buried inside a protein, we can combine previous method with surface area calculation. We developed a method based on RIE sorting and surface area calculation which in 3 steps iterates to final set of residues. Demonstration of this method is a subject of following chapter.

4. **All residue energetic quantities**. This method keeps fixed ratio of key and all residues, but the sorting method is different. The algorithm takes in the beginning only the residues with highest contribution and adds in every step the residues by applying criteria of CO,RIE and RRIE simultaneously. Fortunately, the residues that have IE with only few other residues (usually HB) are removed from the set. On the other hand, it can discriminate AAs that are generally important to protein energetics as it reduces their IEs by mean IE of this type interaction.

4.3.3 Core selection demonstration

Here we present step-by step selection of core residues oxidized microsomal rabbit cytochrome b5 (PDB ID 1DO9) using Surface area exclusion method. Procedure can be illustrated as follows (see Figure 13):

1. Input of the method is a protein structure (Figure 13a).
2. Cα representation is utilized. The backbone is removed and hydrogens are added and optimized (Figure 13b).
3. IEM of uncharged residues is calculated, residues are sorted according to their RIE and selection of AAs is made. Only first M residues are considered, where

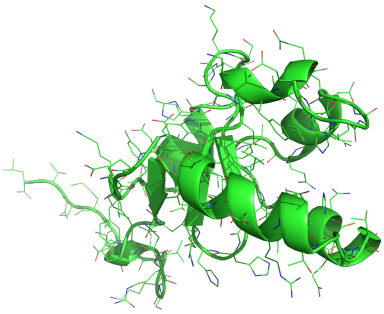
$$M = r \cdot N \tag{16}$$

   and

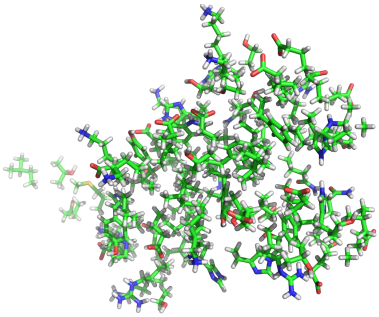
$$\sum_{i}^{M} RIE_i = (1-r) \sum_{i}^{N} RIE_i \tag{17}$$

   Here N is number of residues entering step 3 (excluding charged residues; Figure 13c).
4. Solvent accessible surface area is calculated and surface residues are excluded (Figure 13d).
5. Distribution of remaining residues determines number of key residues according to Equation 16 and 17 (Figure 13e).
6. Output of the method is a set of key residues (Figure 13f).

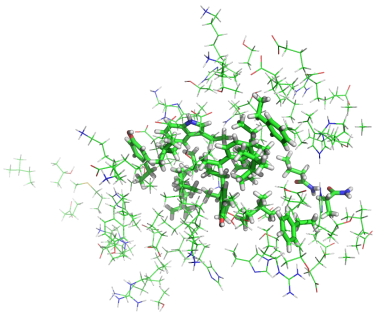
This method is fully automatic, utilizing home made Python program and GROMACS package for optimization and SAS calculation. Main pitfall of this method is high sensitivity to parameters in the 4^{th} step (probe diameter and SAS ratio).



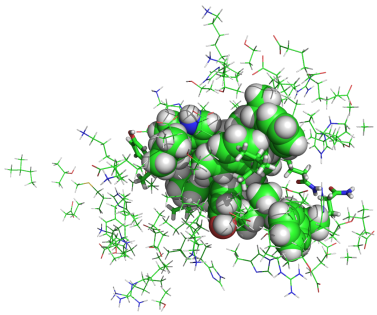
(a) Input geometry - structure of oxidized microsomal rabbit cytochrome b5 (PDB ID 1DO9) without hydrogens



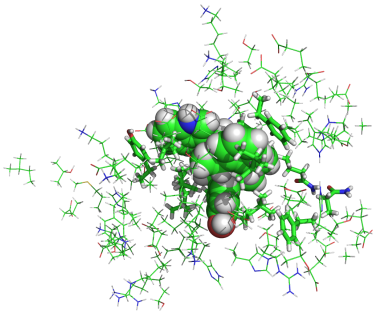
(b) Structure with optimized hydrogens in Cα representation



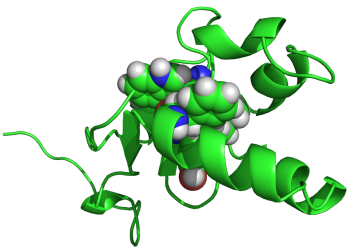
(c) First distribution-based selection



(d) Surface residues excluding



(e) Second selection based on distribution



(f) Final structure. Core residues are pictured by spheres

Figure 13: Illustration of core selection algorithm

5 Conclusion

We offer comprehensive and throughout study of non-bonding interactions between uncharged residues in globular proteins. Main conclusions of this work can be summarized by these statements:

We characterized proteins by universal distribution of amino acid side-chain energy contribution to overall protein stabilization. To our best knowledge, this is the first quantitative evaluation of the enthalpic term to the stability. We utilized sufficient computational accuracy method developed in the supervisor's group which was proved to be reliable in protein side-chain interaction energy evaluations. The number of structures used (about 1,500 protein structures $\rightarrow$ 300,000 residues $\rightarrow$ 60 million interactions) significantly diminishes statistical error. We proposed 6-parameters functional form for this distribution. We reached almost ideal correlation between the experimental and the analytical form.

We characterized all possible pairs of interacting amino acid side chains by distributions. Median of these distributions led to typical energy value for each pair suggesting the importance of the interaction type for protein stabilization. It can be used to determine whether a contact is in its minimum or not.

The medians in proteins with mainly α and mainly β secondary structure are similar. Average energy distribution in proteins does not depend on secondary structure at all. We therefore conclude that preference to form particular secondary structures is probably not in side-chain interactions of composing amino acid residues.

We proposed new definition of contact based on interaction energy matrix. We proved that our definition is robust, transferable and can guarantee the best correlation between interaction energy and contact order. Moreover, our calculations identified the optimum for the contact around -0.5 kcal/mol, a little below energy kT.

Based on knowledge of energy proportions in proteins we proposed sophisticated and transferable methods for HC residues definition. One of the method is described in details. We have a plan to set up a web service on the Institute's page utilizing our methods as well as quantitatively investigate available structural data and characterize core properties in details.

Comparison of the HC stabilization energy with Gibbs free energy of unfolding for ProTherm protein set did not provide any significant correlation of these two quantities. We think that to be able to compare results of molecular modeling with experimental data, new theoretical model is required.

References

Anfinsen, C.B. et al., 'Principles that govern the folding of protein chains', *Science*, vol. 181, no. 96, 223 (1973).

Bahar, I, Atilgan, a R and Erman, B, 'Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.' *Folding and Design*, vol. 2, no. 3, 173–81 (1997).

Baldwin, R.L., 'Temperature dependence of the hydrophobic interaction in protein folding', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 21, 8069 (1986).

Baldwin, R.L., 'Energetics of protein folding', *Journal of Molecular Biology*, vol. 371, no. 2, 283–301 (2007).

Baldwin, R.L. and Muller, N., 'Relation between the convergence temperatures protein unfolding', *Biochemistry*, vol. 89, no. August, 7110–7113 (1992).

Ball, P., 'Water as an active constituent in cell biology', *Chemical Reviews*, vol. 108, no. 1, 74–108 (2008).

Bashford, D. and Case, D.A., 'Generalized born models of macromolecular solvation effects.' *Annual Review of Physical Chemistry*, vol. 51, 129–52 (2000).

Bastolla, U., Porto M. Roman H.E. and Vendruscolo, M., 'Principal eigenvector of contact matrices and hydrophobicity profiles in proteins.' *Proteins*, vol. 58, no. 1, 22–30 (2005).

Bendová-Biedermannová, L., Hobza P. and Vondrášek, J., 'Identifying stabilizing key residues in proteins using interresidue interaction energy matrix.' *Proteins*, vol. 72, no. 1, 402–13 (2008).

Berka, K., Laskowski R. Riley K.E. Hobza P. and 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*, vol. 5, no. 4, 982–992 (2009).

Best, R.B., Buchete N. and Hummer, G., 'Are current molecular dynamics force fields too helical?' *Biophysical Journal*, vol. 95, no. 1, L07–9 (2008).

Biedermannova, L., Riley, K.E., Berka, K., Hobza, P. and Vondrasek, J., 'Another role of proline: stabilization interactions in proteins and protein complexes concerning proline and tryptophane', *Physical Chemistry Chemical Physics*, vol. 10, no. 19, 2581–3 (2008).

Bussi, G. and Parrinello, M., 'Accurate sampling using Langevin dynamics', *Physical Review E*, vol. 75, no. 5, 056707 (2007).

Chalikian, T.V., 'Structural Thermodynamics of Hydration', *The Journal of Physical Chemistry B*, vol. 105, no. 50, 12566–12578 (2001).

Chan, H.S., 'Modeling protein density of states: additive hydrophobic effects are insufficient for calorimetric two-state cooperativity.' *Proteins*, vol. 40, no. 4, 543–71 (2000).

Chen, H., Liu Y. Zhou Z. Hu L. Ou-Yang Z. and Yan, Jie, 'Temperature dependence of circular DNA topological states', *Physical Review E*, vol. 79, no. 4, 041926 (2009).

Connolly, M. L., 'Analytical molecular surface calculation', *Journal of Applied Crystallography*, vol. 16, no. 5, 548–558 (1983).

Cooper, A., 'Heat capacity effects in protein folding and ligand binding: a re-evaluation of the role of water in biomolecular thermodynamics.' *Biophysical Chemistry*, vol. 115, no. 2-3, 89–97 (2005).

Dadarlat, V.M. and Post, C.B., 'Contribution of charged groups to the enthalpic stabilization of the folded states of globular proteins.' *The Journal of Physical Chemistry. B*, vol. 112, no. 19, 6159–67 (2008).

Dill, K.A., Bromberg S. Yue K. Fiebig K.M. Yee-D.P. Thomas P.D. and Chan, H.S., 'Principles of protein folding - A perspective from simple exact models', *Protein Science*, vol. 4, 561–602 (1995).

Dill, K.A., 'Dominant forces in protein folding', *Biochemistry*, vol. 29, no. 31, 7133–7155 (1990).

Dill, K.A., 'Additivity principles in biochemistry.' *The Journal of Biological Chemistry*, vol. 272, no. 2, 701–4 (1997).

Ding, F. and Dokholyan, N.V., 'Simple but predictive protein models.' *Trends in Biotechnology*, vol. 23, no. 9, 450–5 (2005).

Duan, Y., Wu C. Chowdhury S. Lee M.C. Xiong-G. Zhang W. Yang R. Cieplak P. Luo R. Lee T. Caldwell Ja. Wang J. and Kollman, P., 'A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations.' *Journal of Computational Chemistry*, vol. 24, no. 16, 1999–2012 (2003).

Fogolari, F., Brigo A. and Molinari, H., 'The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology.' *Journal of Molecular Recognition*, vol. 15, no. 6, 377–92 (2002).

Gromiha, M.M., 'Revisiting "reverse hydrophobic effect": applicable only to coil mutations at the surface.' *Biopolymers*, vol. 91, no. 7, 591–9 (2009).

Guerois, R., Nielsen J.E. and Serrano, L., 'Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations', *Journal of Molecular Biology*, vol. 320, no. 2, 369–387 (2002).

Halgren, Thomas a., 'Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94', *Journal of Computational Chemistry*, vol. 17, no. 5-6, 553–586 (1996).

Haliloglu, Turkan, Bahar, Ivet and Erman, Burak, 'Gaussian Dynamics of Folded Proteins', *Physical Review Letters*, vol. 79, no. 16, 3090–3093 (1997).

Hermans, J., 'Hydrogen bonds in molecular mechanics force fields', *Advances in Protein Chemistry*, vol. 72, no. 05, 105–119 (2006).

Hildebrand, J.H., 'Is there a "Hydrophobic Effect"?' *Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, no. 1, 194–194 (1979).

Honig, B. and Yang, A.-S., 'Energetics of protein structure', *Advances in Protein Chemistry*, vol. 46, 27. (1995).

Horn, J.R., Russell D. Lewis E.A. and Murphy, K.P., 'Van't Hoff and calorimetric enthalpies from isothermal titration calorimetry: are there significant discrepancies?' *Biochemistry*, vol. 40, no. 6, 1774–8 (2001).

Hummer, G., Garde S., Garcia, A.E., Paulaitis M.E. and Pratt, L.R., 'The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins', (1998).

Kabakçioğlu, A., Kanter I. Vendruscolo M. and Domany, E., 'Statistical properties of contact vectors', *Physical Review E*, vol. 65, no. 4, 1–7 (2002).

Kannan, N. and Vishveshwara, S., 'Aromatic clusters: a determinant of thermal stability of thermophilic proteins.' *Protein Engineering*, vol. 13, no. 11, 753–61 (2000).

Karplus, M. and Weaver, D.L., 'Protein folding dynamics: the diffusion-collision model and experimental data.' *Protein Science: A Publication of the Protein Society*, vol. 3, no. 4, 650 (1994).

Kauzmann, W., 'Some Factors in the Interpretation of Protein Denaturation', vol. 14 of *Advances in Protein Chemistry*, pp. 1 – 63 (1959).

Liu, L. and Guo, Q., 'Isokinetic Relationship, Isoequilibrium Relationship, and Enthalpy-Entropy Compensation', *Chemical Reviews*, vol. 101, no. 3, 673–696 (2001).

Makhatadze, G. I. and Privalov, P. L., 'Energetics of protein structure', *Advances in Protein Chemistry*, vol. 47, 307. (1995).

Matulis, D. and Bloomfield, V.A., 'Thermodynamics of the hydrophobic effect. II. Calorimetric measurement of enthalpy, entropy, and heat capacity of aggregation of alkylamines and long aliphatic chains.' *Biophysical Chemistry*, vol. 93, no. 1, 53–65 (2001).

Mirsky, A.E. and Pauling, L., 'On the structure of native, denatured, and coagulated proteins', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 22, no. 7, 439 (1936).

Mittag, T. and Forman-Kay, J.D., 'Atomic-level characterization of disordered protein ensembles.' *Current Opinion in Structural Biology*, vol. 17, no. 1, 3–14 (2007).

Miyazawa, S. and Jernigan, R.L., 'Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.' *Journal of Molecular Biology*, vol. 256, no. 3, 623–44 (1996).

Miyazawa, S. and Kinjo, A.R., 'Properties of contact matrices induced by pairwise interactions in proteins', *Physical Review E*, vol. 77, no. 5, 051910 (2008).

Müller-Dethlefs, K. and Hobza, P., 'Noncovalent interactions: a challenge for experiment and theory.' *Chemical Reviews*, vol. 100, no. 1, 143–68 (2000).

Onufriev, A., Bashford D. and Case, D.A., 'Modification of the Generalized Born Model Suitable for Macromolecules', *The Journal of Physical Chemistry B*, vol. 104, no. 15, 3712–3720 (2000).

Pace, C.N., Shirley B.A. McNutt M. and Gajiwala, K., 'Forces contributing to the conformational stability of proteins', *The FASEB Journal*, vol. 10, no. 1, 75 (1996).

Porto, M., Bastolla U. Roman H. and Vendruscolo, M., 'Reconstruction of Protein Structures from a Vectorial Representation', *Physical Review Letters*, vol. 92, no. 21, 1–4 (2004).

Privalov, P.L., 'Stability of proteins', *Advances in Protein Chemistry*, vol. 33, 167–241 (1979).

Prudhomme, N. and Chomilier, J., 'Prediction of the protein folding core: application to the immunoglobulin fold.' *Biochimie*, vol. 91, no. 11-12, 1465–74 (2009).

Qiu, D., Shenkin P.S. Hollinger F.P. and Still, W.C., 'The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii', *The Journal of Physical Chemistry A*, vol. 101, no. 16, 3005–3014 (1997).

Rathore, N., Knotts T.A.IV and de Pablo, J.J., 'Density of states simulations of proteins', *Chemical Physics*, vol. 118, no. 9, 4285–4290 (2003).

Robertson, A.D. and Murphy, K.P., 'Protein Structure and the Energetics of Protein Stability', *Chemical Reviews*, vol. 97, no. 5, 1251–1268 (1997).

Sharp, K., 'Entropy-enthalpy compensation: fact or artifact?' *Protein Science*, vol. 10, no. 3, 661–7 (2001).

Shortle, D., 'The denatured state (the other half of the folding equation) and its role in protein stability.' *The FASEB Journal*, vol. 10, no. 1, 27–34 (1996).

Trebbi, B., Fanti M. Rossi I. and Zerbetto, F., 'Intraresidue distribution of energy in proteins.' *The Journal of Physical Chemistry. B*, vol. 109, no. 8, 3586–93 (2005).

Tsai, J., Taylor R. Chothia C. and Gerstein, M., 'The packing density in proteins: standard radii and volumes.' *Journal of Molecular Biology*, vol. 290, no. 1, 253–66 (1999).

Vassura, M., Margara L. Di Lena P. Medri F.-Fariselli P. and Casadio, R., 'Reconstruction of 3D structures from protein contact maps.' *Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, vol. 5, no. 3, 357–67 (2008).

Vendruscolo, M., Kussell E. and Domany, E., 'Recovery of protein structure from contact maps', *Folding and Design*, vol. 2, no. 5, 295–306 (1997).

Vendruscolo, M., Najmanovich, R. and Domany, E., 'Protein Folding in Contact Map Space', *Physical Review Letters*, vol. 82, no. 3, 656–659 (1999).

Vendruscolo, M. and Paci, E., 'Protein folding: bringing theory and experiment closer together', *Current Opinion in Structural Biology*, vol. 13, no. 1, 82–87 (2003).

Vendruscolo, M., Paci, E., Dobson, C.M. and Karplus, M., 'Three key residues form a critical contact network in a protein folding transition state', *Nature*, vol. 409, no. 6820, 641–645 (2001).

Vendruscolo, M., Subramanian B. Kanter I. Domany E. and Lebowitz, J., 'Statistical properties of contact maps', *Physical Review E*, vol. 59, no. 1, 977–984 (1999).

Vondrášek, J., Bendová L. Klusák V. and Hobza, P., 'Unexpectedly strong energy stabilization inside the hydrophobic core of small protein rubredoxin mediated by aromatic residues: correlated ab initio quantum chemical calculations.' *Journal of the American Chemical Society*, vol. 127, no. 8, 2615–9 (2005).

Vondrášek, J., Kubař T. Jenney F.E. Adams M.W. Kozísek-M. Cerný J. Sklenář V. and Hobza, P., 'Dispersion interactions govern the strong thermal stability of a protein.' *European Chemistry Journal*, vol. 13, no. 32, 9022–7 (2007).

Widom, B., Bhimalapuram P. and Koga, K., 'The hydrophobic effect', *Physical Chemistry Chemical Physics*, vol. 5, no. 15, 3085 (2003).

Yue, K., Fiebig K.M. Thomas P.D. Chan H.S. Shakhnovich-E.I. and Dill, K.A., 'A test of lattice protein folding algorithms', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 1, 325 (1995).

Zwanzig, R., Szabo A. and Bagchi, B., 'Levinthal ' s paradox', *Illinois Research*, vol. 89, no. January, 20–22 (1992).

List of abbreviations

AA - Amino Acid

BB - Backbone

CATH - Class - Architecture - Topology - Homologous Superfamily Protein Structure Classification

CBS - Complete Basis Set

CCSD(T) - Coupled Cluster with Single, Double and Perturbative Triple Excitations

CM - Contact Matrix

CO - Contact Order

DNA - Deoxyribonucleic Acid

DOS - Density of States

DSC - Differential Scanning Calorimetry

FF - Force Field

HB - Hydrogen Bond

HC - Hydrophobic Core

HI - Hydrophobic Interaction

H-S - Enthalpic-Entropic

IE - Interaction Energy

IEM - Interaction Energy Matrix

MC - Monte Carlo

MD - Molecular Dynamics

MM - Molecular Mechanics

NMR - Nuclear Magnetic Resonance

PDB - Protein Data Bank

PDB ID - Protein Data Bank Identification Code

PES - Potential Energy Surface

PF - Protein Folding

PFP - Protein Folding Problem

QM - Quantum Mechanics

RIE - Residue's Interaction Energy

RNA - Ribonucleic Acid

RRIE - Relative Residue Interaction Energy

RSIE - Residues with Strongest Interaction Energy

SA - Surface Area

SAS - Solvent Accessible Surface

SB - Salt Bridge

SC - Side-Chain

vdW - van der Waals

Abbreviations of amino acids

ALA - Alanine

ARG - Arginine (protonated)

ASN - Asparagine

ASP - Aspartate (deprotonated)

CYS - Cysteine

GLN - Glutamine

GLU - Glutamate (deprotonated)

GLY - Glycine

HIS - Histidine (protonated)

ILE - Isoleucine

LEU - Leucine (protonated)

LYS - Lysine

MET - Metionine

PRO - Proline

PHE - Phenylalanine

SER - Serine

THR - Threonine

TRP - Tryptophan

TYR - Tyrosine

VAL - Valine