Charles University in Prague

Faculty of Science

Chemistry:

Modeling of Chemical Properties of Nano- and Biostructures

Bc. Boris Fačkovec

Intramolekulární a intermolekulární interakce v proteinech

Intramolecular and intermolecular interactions in proteins

DIPLOMA THESIS

Supervisor: RNDr. Jiří Vondrášek, CSc.

Prague 2012

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In ..... date ..... signature of the author

Název práce: Intra a intermolekularni interakce v proteinech

Autor: Boris Fačkovec

Katedra: Katedra fyzikální a makromolekulové chemie

Vedoucí diplomové práce: RNDr. Jiří Vondrášek, CSc.

Abstrakt: Volná energie sbalení proteinu představuje pozoruhodnou rovnováhu mezi stabilizujícími a destabilizujícími nekovalentními interakcemi. V této práci stabilizující energii rozkládáme na fyzikálně smysluplné příspěvky, které by jsme následne dokázali složit do konzistentního transferabilního obrazu teplotní stability. Empririckým potenciálem vypočteme interakční energie mezi fragmenty klasifikovanými na základe jejich fyzikálních vlastností na skupine 1200 neredundantních struktur z PDB databáze. Výsledkem práce je lepší pochopení vztahu mezi interakčními energiemi vypočenými metodami teoretické chemie a příspěvkami jednotlivých interakcí na této rovnováze.

Klíčová slova: protein, aminokyselina, interakční energie, stabilita proteinů, termodynamika proteinů, molekulární modelování, bioinformatika, biofyzika

Title: Intramolecular and intermolecular interactions in proteins

Author: Boris Fačkovec

Department: Department of Physical and Macromolecular Chemistry

Supervisor: RNDr. Jiří Vondrášek, CSc., Insitute of Organic Chemistry and Biochemistry, AS CR, vvi.

Abstract: Folding free energy of a protein is a delicate balance between stabilizing and destabilizing non-covalent itneractions. In this work, we decompose folding free energy into physically meaningful contributions, in which we aim to find general trends. Empirical potential is used to calculate interaction energy between all protein fragments, which are classified based on their dominant term in multipolar expansion. Calculations are done using 1200 non-redundant structures from PDB database. Based on general trends found in interactions between these fragments, we attempt to better understand relationships between interaction energies calculated using computational chemistry methods and their corresponding free energy contributions on stabilization.

Keywords: protein, amino acid, interaction energy, protein stability, protein thermodynamics, molecular modeling, bioinformatics, biophysics

Contents

Preface

Protein folding problem has been attracting researchers for more than 80 years. Despite incredible work done by theoretical and experimental groups, we still do not understand balance of the driving forces of protein folding. It seems now, that we have fundamental understanding and that only quantitative performance of our models is not sufficient for the problem to be solved. Probably most promising research direction are studying energy landscape topology, identification of errors in interaction energy calculations and decomposition of stability.

Although the thesis is called "Intramolecular and intermolecular interactions in proteins", only intramolecular interactions in globular proteins are addressed, since protein-protein interactions have been studied by Jiří Kysilka, PhD student in the group of Dr. Vondrášek. In our group, interaction energies between amino acids were calculated at highest level of accuracy few years ago. In this time, there is only limited space for improvement of description of particular pairwise interactions. More challenging and urgent problem became the connection of the interaction energies and their free energy contribution to protein stability.

The main contribution of this work is a novel treatment of solvent effects and development of stability model and method for optimization of its parameters.

Introduction presents an essential view on protein folding problem integrating statistical mechanics, biophysical experimentation and molecular modeling. Attention was paid to discuss in details terms ambiguously defined in literature. Phenomena studied by one of the mentioned disciplines were formulated in a way understandable by researchers from different fields.

1. Introduction

Proteins are linear biomacromolecules synthesized in ribozomes by linking amino acids by peptide bonds. The can fold into stable 3D structure at certain conditions. Foldable proteins have 2 forms

Definition 1. *Native state of a protein is an ensemble of microstates which are very similar in structure. This state is the most stable at native conditions.*

Definition 2. *Denatured state of a protein is an ensemble of microstates which are higher in energy but much more numerous than those of native state ensemble. Denatured state is more stable than native state in presence of denaturant, which can be chemical agent .*

In order to prevent ambiguities this work strictly distincts between denatured and unfolded state.

Definition 3. *Unfolded state is in this work defined as a random coil - a theoretical construct of heteropolymer chain in which non-covalent interactions between constituent amino acids are negligible.*

In the first subsection of Introduction, only certain observations easy to interpret without complex models are briefly overviewed. Emphasis is put on stability studies, but other relevant experimental observations are briefly mentioned. In the next section, the facts are complemented by most influential theories of protein folding which formed the field. As the most important result in this work is represented by the stability model, previous stability models and additivity principles are discussed. Subsection 1.4 provides three different views on intramolecular interactions - view of statistical physics, quantum chemistry and structural biology. Link between interactions and protein stability adressed by previous studies is referenced throughout the whole work. As we propose new definitions of residue-residue contact and interaction energy matrices are used throughout the work, section 1.5 introduces to the matrix representation of protein organization.

1.1 Experimental studies of protein structure and stability

1.1.1 Biological context

Proteins constitute about 40 % of dry weight of human organism [1], in which they perform most important tasks beacuse of their structural variability, their functional and binding specificity and for their ability to adopt special properties like enzymatic activity or even fluorescence. Proteins are evolutionary optimized for general well-being of an organism, i.e. not causing illnesses by misfolding, being able to resist short- or long-term adverse conditions and catalyze reactions to increase adaptability and compatibility of the whole metabolic network of an organism to its environment. For successful accomplishment of the mentioned properties, proteins need to have stability in some specified thermodynamic interval and to fold in order of nanoseconds to minutes.

Proteins are synthesized in ribosomes in process of translation where a selected linear information from DNA is (after transcription and splicing) translated into sequence of amino acids. Simultaneously with synthesis of the chain, the synthesized N-terminus of the new protein starts to fold into 3D structure [2]. It is generally accepted that function of a protein is determined by its structure. However, about 30% of proteins are unstructured [3] [4]. These intrinsically disordered proteins have were overlooked by protein biophysics community in the past but recently have been becoming increasingly popular[5]. The process of folding is in cells sometimes promoted by chaperones inhibiting misfolding of nascent polypeptides. In eukaryotic organisms, translation is usually followed by posttranslational modification changing chemical character of proteins. After this process, loss of ability to refold is usual and the structure with minimum energy can change substantially .

Protein folding problem in the natural environment is a complex problem of molecular biology. Number of relevant factors of such studies is immense. The cellular environment can be imagined as a dense soup with concentration of proteins in cell being in order of 300 g/l with no aggregation. Effects of macromolecular crowding on protein structure and function have been recently extensively reviewed by Zhou [6] and Elcock [7].

1.1.2 Calorimetric studies

Necessary reduction of protein folding problem came to a simplified formulation which remains a challenge for physicists:

Definition 4. *Protein folding problem concerns about thermodynamics and kinetics of transformation of a protein chain composed of 20 types of amino acids from denatured to native state in a buffered water solutions of low protein and buffer / salt concentration.*

In this thesis, only foldable globular proteins are studied; disordered, membrane and fibrilar proteins are not subject of this study. In 1961, Anfinsen proposed a hypothesis based on his in vitro refolding studies of ribonuclease [8].

Hypothesis 1. *(Anfinsen) Structure of a protein is uniquely determined by its sequence and environmental conditions (solvent composition, temperature etc). The native state structure represents global free energy minimum ensemble at these conditions.*

Pitfalls of the original formulation by Anfinsen are discussed in works of Govindarajan [9] and Ben-Naim [10]. Throughout the rest of this work, we assume validity of thermodynamic hypothesis. We can now define thermodynamic stability of a protein.

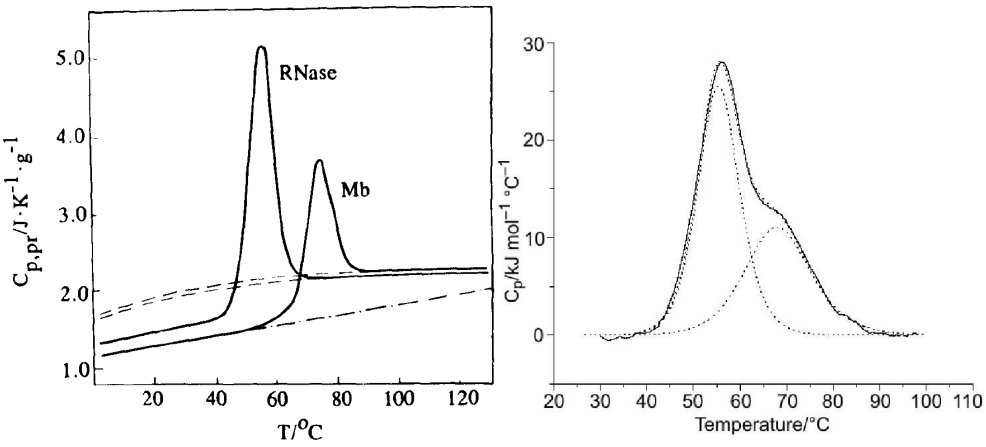
Definition 5. *Protein stability is defined as the Gibbs free energy difference between native and denatured state of a protein at defined conditions (temperature, pressure, solvent composition, pH etc.). Negative value means higher stability of native state.*

In the rest of the text, shorter term 'free energy' will be used instead of Gibbs free energy (or free enthalpy). The difference between Helmholtz free energy and

Gibbs free energy is in liquids such negligible that the two terms can be used interchangeably.

In 1976, Pfeil and Privalov proved [11] that calorimetrically determined enthalpy, entropy and free energy represent real thermodynamic potentials specifying the states of a protein. Differential scanning calorimetry (DSC) is therefore the source of experimental data with most straightforward interpretation, so it is no surprise that DSC has become standard methodology of biophysics. Heat capacity of a small sample of dissolved protein is recorded as a function of temperature. The methods is considered to be very precise; the least precise variable is probably protein concentration [12] (error is estimated to be under 2%). A typical output curve of DSC experiment of protein denaturation is shown on Figure 1.1.

Figure 1.1: Left - curve of DSC measurement of pancreatic ribonuclease A (RNase) and sperm whale myoglobin (Mb). Peak maxima correspond to melting temperatures. Dashed lines denote extrapolated heat capacities. Adapted from ref. [13]. Right - DSC experiment of defatted bovine serum albumine, a protein consisting of 2 domains. Adapted from ref. [14]



Protein presents a thermodynamically discrete macroscopic system i.e. it can be divided into domains each having only 2 accessible states [15]. Temperature dependence of stability of 1 domain can be well described by 3 parameters.

$$\Delta G = \Delta H - T\Delta S = \Delta H_0 + \int_{T_m}^{T} \Delta C_p dT + T\left(\frac{\Delta H_0}{T_m} + \int_{T_m}^{T} \frac{\Delta C_p}{T} dT\right) \qquad (1.1)$$

$\mathbf{T_m}$ - unfolding (or melting) temperature or transition midpoint is temperature at which Δ G = 0 is the highest point of the DSC curve. T_m can be easily measured by other techniques and its value is known for most of proteins.

$\mathbf{\Delta H_0}$ - enthalpy of unfolding is the heat release accompanying unfolding at temperature T_m. Unfolding of a protein is always endothermic process, i.e. energy must be supplied to increase entropy. It can be determined by DSC as area under the peak in C_p/T plot. An alternative method of determination of Δ H can be derived from van't Hoff equation

Table 1.1: Correlation between thermodynamics quantities of globular proteins and chain length. First two columns contain correlated quantities (N - protein size = number of amino acids, $\Delta H, \Delta S, \Delta C_p$ and T_m are defined above. Numbers in parentheses after quantities are temperatures at which the quantities were measured. Data in line 1 to 5 are obtained on set of 65 proteins [12], data in line 6 and 7 are obtained using datase of 3224 proteins[18]. Measurements at inconsistent conditions. Data in columns 8 and 9 on set of 19 proteins [19]

N	$\Delta H(374K)$	0.922
N	$\Delta H(333K)$	0.789
N	$\Delta S(385K)$	0.920
N	$\Delta S(333K)$	0.759
N	ΔC_p	0.862
$\Delta H(298K)$	298 K $\Delta S(298K)$	0.991
$\Delta H(298K)$	$\Delta G(298K)$	0.600
T_m	$\Delta G(298K)/N$	0.830
T_m	$\Delta H(298K)/N$	0.810

$$\Delta H = k_B T_m^2 \frac{\mathrm{d}\ln K}{\mathrm{d}T}|_{T_m} \tag{1.2}$$

where k_B is Boltzmann constant, and K (equilibrium constant) is ratio of concentration of folded and unfolded form. Comparison of calorimetric ΔH and van't Hoff ΔH can be used to assess cooperativity of folding [16].

$\mathbf{\Delta C_p}$ - heat capacity difference between native and denatured state. In 1979, Privalov [17] discovered that linear extrapolation of experimental data yields convergence of both specific enthalpy and entropy of unfolding to common values at approximately the same temperature. Independent studies of liquid hydrocarbons dissolution showed that at this temperature (112 degC) entropy of hydrocarbon fransfer to water becomes zero.

In 80's and 90's, extensive calorimetric measurements of globular proteins were performed. Experimental correlation between thermal data and protein chain length can be found in Table 1.1.

Briefly, folding enthalpy very well correlates with protein size. Each residue contributes almost 2 kcal/mol (about 8 kJ/mol) to folding enthalpy. Stability of all proteins lies between 0 and -20 kcal/mol despite good correlation of unfolding ethalpy and protein size. This inconsistency is due to strong compensation of enthalpy term by opposing entropy. The enthalpy-entropy compensation [20, 18, 21] is about 91% [18]. Energy of contributing interactions is two orders of magnitude higher than free energy of folding at standard conditions. Proteins can be denatured by heat or cold.

1.1.3 NMR and X-ray

In 1958, Kendrew et al published first X-ray structure of protein [22]. It was found that unlike DNA, protein chains are folded into globules, with hydrophobic residues being buried inside and hydrophillic ones being exposed on protein

surface. On average, approximately 83% of hydrophobic residues are buried compared to 63% for polar and 54% for charged ones [23]. This fact indicated that hydrophobic effect plays significant role in protein folding. Packing of hydrophobic residues in protein interiors has ratio of void space similar to aliphatic crystals rather than liquids. Dense packing of hydrophobic core has been identified as a determinant of protein stability.

Secondary structure had been discovered even before first X-ray structure of protein appeared [24]. Nowadays, structures are classified hierarchically by two major classification schemes (CATH [25]) and SCOP [26]). An interesting fact is that although the number of structures determined per year continually increases, the number of new folds discovered per year decreases since 2004 (SCOP) or 2007 (CATH).

Other general structural features have been observed in proteins and their occurence has been correlated with thermal stability, mostly by comparison of structures of homologues proteins with significantly different thermal stability. In these studies, an important assumption is made:

Hypothesis 2. *Native state of a protein can be characterized by one structure.*

Crystallography provides an averaged structure which is probably the best represenative one. Some examples of common structural features in proteins are pairing of oppositely charged amino acid sidechains [27, 28], and aromatic clusters [29]. Structural features like polar contacts or volume have been correlated with protein stability. Some structural studies can provide information about denatured state. It has been shown [30, 31] that hydrophobic clusters can be present in denatured state of proteins.

Native structures are essential for most of the theoretical protein thermodynamics studies. As an efficient application of NMR methods is limited to small proteins, X-ray crystallography remains the main source of data. The assumption that structures of proteins in crystals are identical to those in solutions is widely accepted and firmly based [32]. Unfortunately, X-ray crystallography cannot always determine rotameric position of asparagine, glutamine and histidine which are ex post modeled [33]. Studies by Higman et al. [34] assessing X-ray structures by NMR experiments show that in the studied protein, at ξ_2 or ξ_3 in at least one glutamine or asparagine residue (respectively) is incorrect. Quality of structure of flexible regions is usually very poor. Although methods for structure determination are subject of continual development determination of a protein native structure remains an expensive and long process.

1.1.4 Biophysical and chemical studies

Apart from high temperatures, unfolding of a protein can be caused by chemical agents called denaturants. Concentrated acids were the first known denaturants. High stability dependence on pH led into conclusions that ion pairing is the main force responsible for protein structure . Other denaturants (guanidinium chloride, urea, etc.) are used to measure stability of proteins by measuring concentration ratio of denatured and native state using spectral probes [35] and extrapolating to zero concentration. Slopes of stability dependence on denaturant concentration, called m-values [36], determined by this method well correlate with solvent

accessible surface area difference of native and unfolded state. Recent theories about chemical denaturation suggest that denaturant change structure and lower free energy of denatured state rather than increasing energy of native state. Therefore, one should be careful when considering m-values to derive models of protein thermal stability [37, 38].

Hydrodynamic methods like dynamic light scattering are used for determinations of gyration radius. Scaling of hydrodynamic radius with chain length is compared with theoretical predictions based on simplified models of denatured or unfolded state. Pressure perturbation calorimetry [39] enabling measurement of temperature dependence of thermal expansion coefficient [40, 41] helps to understand hydration of folded and unfolded proteins.

Mutational studies are invaluable tools of studying effects of single amino acid sidechain on protein stability or folding rate. Alanine screening measures contribution of a sidechain to stability by recording stability change upon its replacement by alanine. Double mutant cycles are used to measure interaction between two sidechains.

Atomic force microscopy enables studying thermodynamics and kinetics of mechanical unfolding of a sigle molecule. Such experiments can be directly reproduced by simulations. Pathways of protein folding have been studied by NMR [42] and protein folding process can be observed real-time in vivo using fluorescent labeling [43]. Information about transition state ensemble can be obtained by ϕ and ψ value analyses [44].

1.1.5 Availability of experimental data

Up to now, more than 80,000 structures have been deposited to PDB publicly open protein database [45] and the number of structures still increases. About 24,000 protein structures out of 80,000 represent non-redundant proteins (70% sequence identity removed, 16,000 if 30% removed), 11,500 of which have been resolved by X-ray crystallography at resolution better than 2 Å.

Thermodynamics data are collected in ProTherm [46] database. Although determination of native structure is more laborous than DSC measurements, full thermodynamics characterization of proteins is spare. Out of about 25,000 entries in ProTherm database, only about 100 represent ΔG values measured by DSC on proteins with structure deposited in PDB at consistent conditions (pH 7). Similarly, about 100 entries represent consitently measured ΔH by DSC on proteins with publushed structure. Most of the data represent stabilities of substitution mutants, so structure must be modeled for theoretical studies. Stability changes upon amino acid replacement for same systems measured at same conditions published by different research groups are not equal, but correlate with Pearson coefficient 0.86 [47]

1.2 Theoretical investigations of protein structure and stability

1.2.1 Framework of protein modeling

Models of proteins have been developed simultaneously with experimental studies to interpret observations and to answer some theoretical questions. Development in integration of theory and experiments is reviewed in references [48, 49, 50]. Aims of the protein models were usually to explain folding times which are surprisingly low [51], to find the dominant force of protein folding or to study phase diagrams of proteins. For structure prediction, models leading to energy function for structures or to protein folding kinetics are of principal interest. The latter interest is motivated by the fact that finding a global minimum on energy landscape is NP-complete problem [52], though the nature can fold a protein in miliseconds to seconds. Structure prediction algorithm might efficiently simulate natural folding process.

Protein folding occurs on a broad scale of times and lengths. Proteins are composed of tens to thousands of amino acids, i.e. hundreds to tens of thousands of atoms. Diamaters of folded proteins vary from units to tens of nanometers. Helix-coil transition occurs in order of microseconds, folding in order of milicesonds to seconds. The limitations of model complexity imposed by protein size and complexity of their configurational space severely limits the maximum accuracy. It is in principle possible to run accurate quantum mechanics simulations, where nuclei are treated as quantum objects. However, even a single point quantum mechanics calculations are prohibitively expensive. All-atom empirical force field model is the highest accuracy of microstate description that allows sufficient sampling to some extent. Folding simulations from extended state to native state have been performed only for few small proteins. On the other side, the most simplified models are simply theories with no analytical solution. Their solution can be found by simple simulations, e.g. averaging of all states of a HP lattice model.

Pure theory with analytical solutions to model equations is rare because there is myriad of possible sequences, each having its own properties. General polymer theory is well developed but applicability on proteins is limited. Analytical model of entropy for a cross-linking polymer was proposed by Vorov et al [53]. Some purely theoretical models for proteins have been proposed by Dill's group. Wide variety of simplified model was reviewed [54]. Models for entropy of protein ensembles is digestedly reviewed in review by Dill and Stigter [55]. Statistical mechanics of warm and cold unfolding has been studied by Hansen et al. [56]. Excellent review of protein models and theories of protein folding, their assumptions and experimental validations was written by Shakhnovich [57]. More recent review on theoretical studies of proteins focused on molecular transfer model and unfolding was written by Thirumalai et al. [58]. Another review on the topic focused on Markov state model has been recently published by Bowman et al. [59].

Models of proteins can be characterized by 3 main features.

1. Space in which a protein resides. Continuous space is more realistic while lattices enable sampling of all possible protein configurations.

2. Particle representation of the protein chain. In all-atom models, each nucleus is represented by one particle, in coarse-grained models, small groups of atoms are represented by beads. A residue is in many models represented by 1 or 2 spheres or a points on a grid. Each amino acid can have its own paremeters or amino acids can be grouped like in HP model[60], where there are only 2 types of amino acids.

3. Potential energy form or force field assigns energy to a given protein structure based on mutual position of residues and displacement from equilibrium positions. In lattice models, residue-residue interaction energy is evaluated based on connectivity on lattice while in continuous models geometric distance or mutual orientation are considered.

Direct simulations of protein folding using all-atom models are rare. Implicit solvent models are usually applied to decrease computational cost. Recently, 1 ms simulation of BPTI in explicit solvent [61] was performed and boldly presented. Large independent simulation running on different computers was pioneered by folding@home project. Other related notable projects are bluegene, poem@home, protein structure initiative and dynameomics.

Standard view on protein emphasizes native states and usually simplifies solvation and denatured state. Therefore, these two important but less developed aspects of protein modeling are discussed in separate sections.

1.2.2 All-atom models of proteins

Despite the technical and methodology development, accurate quantum dynamics studies are limited to systems with few electrons. Born-Oppenheimer approximation, i.e. treating nuclei as classical objects, allows substantial simplification with reasonable error. Electronic structure problem for given coordinates of nuclei is mostly solved by variational or perturbation wavefunction theories or densitiy functional theories. Wide scale of methods has been developed [62] but generally, the time needed for accurate calculations increases at least with third power of number of atoms or base functions. Accurate methods with linear scaling derived from the mentioned methods imposing a distance constraint on orbital correlation are in vigorous development and might be in near future useful for protein modeling.

In proteins, stability in order of tens kcal/mol is a balance of thousands of residue-residue and residue-solvent interactions whose strength is in order of units of kcal/mol. Since random error propagates with square root of number of residues and systematic error linearly with number of residues each interaction must be calculated to level of sub-chemical accuracy and systematic errors eliminated to level of 'sub-sub-chemical' accuracy (0.01 kcal/mol)[63]. Such accuracy is unreachable even by very expensive benchmark ab initio quantum mechanics calculations. Therefore, empirical force fields seem to be good tradeoff between accuracy and speed. They preserve most important features, like mutual orientation , but in the same time can be calculated in a short time allowing sampling of millions of microstate in hours of computational time. Review of force fields used for all-atom protein simulations can be found in [64]. Potential energy is composed of additive contribution of the form

$$U = U_{bonds} + U_{angles} + U_{torsions} + U_{coulomb} + U_{vdW} \tag{1.3}$$

$$U_{bonds} = \sum_{bonds} \frac{1}{2} k_b (l - l_0)^2 \tag{1.4}$$

$$U_{angles} = \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 \tag{1.5}$$

$$U_{torsions} = \sum_{torsions} \frac{1}{2} V_t [1 + \cos(n\omega - \gamma)] \tag{1.6}$$

$$U_{coulomb} = \sum_{i=1}^{i<N} \sum_{j=i+1}^{j<N+1} 332 \frac{q_i q_j}{r_{ij}} \tag{1.7}$$

$$U_{vdW} = \sum_{i=1}^{i<N} \sum_{j=i+1}^{j<N+1} 4 \cdot \epsilon_{i,j} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{1.8}$$

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 cosines.

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. Is partial charges q_i are given in units of elementary charge, $U_{coulomb}$ is in kcal/mol.

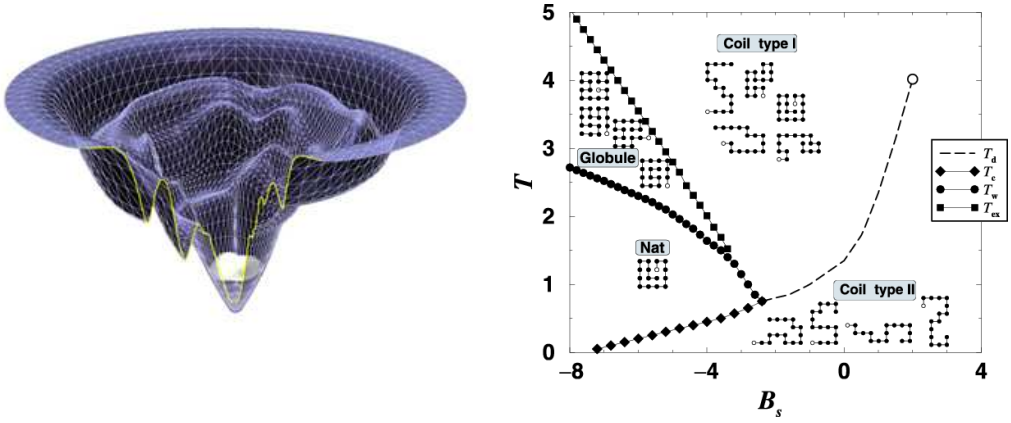
Empirical force fields fail to describe bond breaking, so the protonation state of each basic or acidic site must be determined before simulation. Dispersion energy is surprisingly well described in force fields by Lennard-Jones potential [65], whereas non-additive induction forces remain a challenge. Until polarizable force fields are sufficiently developed, induction energy significantly contributing to amide hydrogen bonding will remain the weak point. It seems that backbone parameters are biased towards helical conformations [66] as non-bonding backbone parameters are often fitted to reproduce secondary structure [67].

Empirical force fields has been used for simulations of protein folding [68] and dynamics, free energy perturbation calculations of stability change upon amino acid replacement [69]. Native state ensemble can be very well sampled [70] but sampling of all-atom models of denatured state remains a challenge. It is also worth to mention popular model systems. Some proteins are extraordinarily suitable for simulations for their small size or available mutational data, for example BPTI, villin headpiece, SH3 domain, lysozyme, GroEL, rubredoxins etc.

1.2.3 Simplified models of proteins

By simplified models I hereby mean all models which cannot reproduce all the desired features of all-atom models. One approach of their construction is coarse-graining, a process in which atoms are grouped to be represented by one sphere or elipsoid. Coarse-grained force fields like MARTINI [71] are widely used for simulations of membrane proteins , biomolecular complexes and for sampling denatured state ensemble.

Figure 1.2: Left - schematic picture of folding funnel. Adapted from ref. [80]. Right - phase diagram of protein constructed using lattice model. Adapted from ref. [81]



Many valuable simple models provided deeper insight into protein folding. They fill the gaps of time scales unreachable by experiments and link experiments and more accurate simulations to theory. Go model [72] is often called a perfect gas model for protein folding. Smooth energy landscape where reaction coordinate of protein folding can be easily defined. In 90s, lattice models were extensively used to study whole configrational space of proteins. Phase diagrams and folding funnels of proteins (see Figure 1.2) were studied. Distance constraint model (DCM) was constructed to provide a rapid and accurate estimate of conformational entropy with minimal computational cost [73, 74]. Excellent review by Kolinski and Skolnick [75] summarizes the most significant outputs of reduced protein models. Simple or minimalist models of proteins are also reviewed in [54, 76, 77, 78, 79].

1.2.4 Solvation

In vitro experiments which are usually benchmarks for theoretical models are done in low-concentrated solutions of buffers in water. Interactions contributing to thermodynamics of protein folding can be therefore divided into 3 groups

- protein-protein interactions
- protein-solvent interactions
- solvent-solvent interactions

It is generally accepted, that solvent plays important or even critical role in protein folding. Many puzzles of protein thermodynamics result from contraintuitive features of solvent effects. For its importance, water is sometimes called the 21st amino acid [82].

Mechanism of solvation of polar species is more intuitive than solvation of non-polar ones. Free energy of polar particle transfer from vacuum to water is almost equal to enthalpy. Its because of strength of electrostatic interactions between solvent and partially strongly charged solute, which exceed water-water interactions. Unlike polar hydration, hydrophobic hydration causes density of water to decrease and heat capacity of the system to increase [1]. Entropy of hydrophobic hydration is 0 for many solutes at about 400 K [83]. Hydrophobic effect has been a subject of debates since its discovery [84, 85, 86, 87]. In 1959, Kauzmann proposed that it is the driving force of protein folding. This indication of biological relevance contributed to considerable attention it attracted. Entropic contribution to hydrophobic effect seems to prevail for small and enthalpic contribution for large solutes. Recent studies suggest that shape of the solute is also important [88].

In unfolded state, only protein-solvent non-covalent interactions, backbone torsion and some self-avoiding residue-residue potential, like hard spheres, play role. Free energy of burying a hydrophobic group has been studied extensively experimentally and theoretically . Usually independence of hydration of backbone and sidechain is supposed. However, it has been found that even group additivity is unjustified in this case[89]. While hydration energies of sidechain analogues are measured to high level of accuracy, data for capped amino acids are spare.

Polarizable all-atom water is the most precise used model of solvent in simulations. However, as mentioned above, simulations using polarizable force fields are rare. Many water models for empirical potential simulations have been developed. Their main parameter is number of sites for partial charges which corresponds to their computational cost and boundaries of their quality from 3-site models (TIP3P, SPC/E) to n-site models (TIP4P, TIP4P/2005; TIPnP, n=5,6...). Further simplification leads to coarse-grained water models.

Explicit modeling significantly increases number of degrees of freedom. Therefore, simplified model have been devised. In review [90], Warshel et al. present hierarchy of solvation models. At the top of the hierarchy, the most accurate and expensive microscopic models are followed by simplified microscopic and macroscopic are at the bottom. In 1976, Warshel and Levitt proposed modeling solvent by explicit grid of Langevin dipoles [91] (LD model). Poisson-Boltzmann (PB) models solve Poisson-Boltzmann equation, which is simply linearized Poisson equation [2] assumed Boltzmann's distribution. They are strongly dependent on dielectric constant used. Even more simplified treatment of electorstatics, Generalized Born (GB) model, was proposed in 1997 [92]. Actually, GB model stands for simple Coulomb law with distance-dependent dielectric constant [90]. The relationship between simplified microscopic models (like LD or BDL models) and macroscopic models (PB, GB) was established by Papazyan and Warshel [93]. Briefly, macroscopic models disregard structure of water and even extended Poisson model cannot describe physical discreteness of the solvent. Explicit and implicit solvend models are often combined to improve description of solvation shell without significantly increasing degrees of freedom. In such treatments, small amount of water molecules representing few hydration shells are added and

[1] Great source of strange properties of water and hydration is website of London South Bank University "water structure and science" created by Dr Martin Chaplin

[2] Poisson equation is an alternative formulation of Coulomb law.

reaction field represents the bulk water.

Compensation of intramolecular interactions by protein-solvent interactions is crucial for descriptions of Free energy of an interaction is usually calculated using a simple solvent model (GB) or just a dielectric constant for scaling electrostatics. Optimum value of dielectric constant has been proposed as 2 in the case induced diploes are included implicitly and about 40 if Dielectric constant is often assumed to be a universal quantity representing interaction energy decrease However, scaling of a residue-residue interaction by a dielectric constant is equivalent to assumption that residues are fully solvated in Moreover, the solvation is macroscopic, which means that not only group additivity applies, but also

1.2.5 Denatured state

Definition of protein stability implies that two states of proteins should be investigated and thermodynamically characterized. However, low attention was paid to structure and energetics of denatured state as it was usually identified with unfolded state. Unfolded state can be very well described by polymer theories [94] because residue-residue interactions are negligible and solvation of amino acids is independent.

Hypothesis 3. *Denatured state ensemble is identical to unfolded state ensemble.*

While structure of native state can be experimentally determined, structure and therefore energetics of denatured state ensemble is poorly understood [95]. Initial approximations of denatured state by unfolded were supported by scaling of hydrodynamic radii of urea and GdCl unfolded proteins with chain length. However, mechanism of denaturation by chemical agents seems to be much different from thermal denaturation. Apart from that, the highly concentrated solutions of denaturants are far from natural environment of proteins. Several studies have shown that denatured state is much more compact than unfolded state [96, 97]. NMR methods suggest that denatured state contains hydrophobic clusters [31] and large amounts of residual secondary structures [98]. Hypothesis 3 can be therefore confidently refused.

Modeling denatured state ensemble consisting of huge number of poorly folded structures remains a challenge for molecular modeling. First, it seems that backbone parameters of current force fields are biased towards helical structures and would probably fail to describe protein structures being far from folded ones [66]. Second, ensemble constitutes a complex subset of configurational space , which is very difficult to sample. Simplified models have been used to study compact denatured states [99] and recently to study dominant forces in denatured state ensembles [100].

Denatured state did not receive attention it deserved due to difficulties in its proper modeling. Moreover, denatured state is irrelevant for structure prediction as it is same for all decoys as well as for the native state.

1.3 Stability models

1.3.1 Driving force of protein folding

In 1959 in his seminal paper [84][3], Kauzmann suggested that hydrophobic effect might be this force. Strong enthalpy-entropy compensation observed in proteins is common for solvation of hydrophobic species. This theory was also supported by urea denaturation experiments of Tanford [102] which supported theory of identification of unfolded and denatured states and by other properties of proteins like Privalov puzzle and burial of hydrophobic residues observed in experimentally determined native structures.

In 1984, Dill [60] derived a model based on the abovementioned assumption, which predicted stability as a function of fraction of hydrophobic residues. According to the model, at least 42% of residues in a protein must be hydrophobic in order to the protein be stable.

However, it has been found that hydrophobic effect is not the only stabilizing force, indeed its contribution to protein stability was estimated to 60% by Pace et al. [103]. Recent studies show that loss of van der Waals interactions upon unfolding might be even more important than the effect of hydrophobic surface exposition [104, 105, 106, 107]. Comparison of fusion enthalpy of benzene or propane (up to 1 kcal/mol) is not far from the values of average enthalpy contribution per residue (up to 2 kcal/mol). Compactness of denatured states undermines theories emphasizing role of hydrophobic effect and supports the view of protein unfolding as a hydrophobic crystal fusion rather than oil droplet dissolution.

1.3.2 Free energy partitioning

In early 90's models of stability decomposition into contributions of amino acids were proposed. Such models assume additivity of free energy.

Hypothesis 4. *Stability can be decomposed into free energy contributions of groups af amino acids.*

The hypothesis 4 is valid only if energy contributions are independent, i.e. if sum of interaction energies of the components is zero in each microstate (Equation 1.9).

$$E(x_1, x_2, ...) = \sum_i E_i(x_i) \tag{1.9}$$

Therefore, even though additivity of potential energy perfectly applies for a single microstate, which is the case of empirical potentials, additive treatment of ensemble average quantities like free energy and even enthalpy introduces error. While entropy is almost negligible in thermochemistry, entropy is very important in protein folding. Additivity of potential energy can be safely used for studying native state if the hypothesis 2 is accepted. Justification of free energy decomposition in protein stability studies has been vigorously debated [108, 109]. Dill suggested that the non-additivity problem is particularly significant for energy component decomposition [110].

[3] according to Pace [101] the most important paper ever published on protein stability

As mentioned before, each residue contributes approximately 2 kcal/mol to folding enthalpy [12]. Also heat capacity change upon unfolding can be well decomposed into contributions of amino acids (see Table 1.1). Such partitioning is problematic for protein stability as it disregards residue-residue interactions. Since value of typical interaction energy is the same magnitude as protein stability, an average contribution of amino acids to stability would be inaccurate. Such contributions for an amino acid type would involve

- average free energy of burial from denatured state to native state, provided that residues are more exposed to water in denatured state
- average free energy contribution of residue-residue interaction of this amino acid in denatured state
- average free energy contribution by conformational entropy of folding
- average free energy contribution of residue-residue interaction of this amino acid in native state

Mutation studies have shown that such contribution correlates with buried area [111]. Remarkable model by Ghosh and Dill [112] enables stability prediction from sequence.

1.3.3 Contribution of interactions in native state

1-body decompositions cannot discriminate native state from decoys, since they neglect interactions in native state. Considerable non-additivity in double mutant cycles [113] also corroborates significance of residue-residue contacts. Therefore, much attention was paid to stabilization by intramolecular interactions. The methods mostly identify structural biology features and assign them universal values established by previous studies. One of the first free energy partitionings by Ponnuswamy and Gromiha [114] comprises hydrophobic, electrostatic, hydrogen bonding, disulfide, van der Waals, entropic and non-entropic (in denatured state) free energy contributions (Equation 1.10).

$$\Delta G = G_F - G_U = (G_{hy} + G_{el} + G_{hb} + G_{ss} + G_{vw})_F - (G_{en} + G_{ne})_U \qquad (1.10)$$

G_{hy} was calculated from solvent accessible surface areas of amino-acids calculated by previous study [115] and solvation parameters. Each identified surface ion pair [116] was assigned 1 kcal/mol and each buried one 3 kcal/mol. G_{hb} was calculated from protein chain length assuming number of hydrogen bonds is 0.73 times number of residues and each hydrogen bonds was expected to contribute 1 kcal/mol. Each disulfide bridge was assigned stabilization of 2.3 kcal/mol, G_U was calculated from chain length and G_{hb} as $1.2N + 0.5G_{hb}$. van der Waals energy was calculated from chain length as $G_{vw} = 8.885 + 0.1413N$ kcal/mol. Average contribution of hydrogen bonding, hydrophobic burial and van der Waals forces was similar, while contribution of electrostatics was small. Fitting 6 parameters to 14 experimental values led expected correlation.

More recent study integrating stabilizing energy contributions by Pace et al. [117] also relies on structural biology definitions of interactions.

Table 1.2: A rough estimate of the contribution of various forces to the conformational stability of RNase T1 [117]

	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**

Assuming that denatured state is identical to unfolded, one can easily improve description of folding enthalpy by adding contribution of native state itneractions. Lazaridis, Archontis and Karplus [118] decomposed stabilizing enthalpy into contribution using physics-based empirical force field. Nowadays, stability models are studied by groups developing energy functions for protein structures.

1.3.4 Energy functions for structure prediction

Potential energy function gives energy of one microstate. From this value, free energy of ensemble of states close to the representative one can be calculated only by sampling over the ensemble. Such sampling is extremely computationally expensive, especially if explicit water model is used. In structure prediction, energy function needs to be evaluated fast from a single structure. Free energy function enabling comparison of any pair of ensembles by comparison of representative structures are called energy functions for protein structure (EF). Energy functions for structure prediction must deal with the mentioned error propagation problem but need not deal with denatured state as all decoys and the native state share the same denatured state ensemble.

The energy fucntions can be classified into 4 groups

- Physical effective energy functions (PEEF) are based on fundamental analysis of the interactions stabilzing proteins Their parameters have physical meaning and do not depent on any data set.
- Statistical effective energy functions (SEEF) are derived from known protein structures. They are less sensitive to errors of atom displacement.
- Empirical effective energy functions (EEEF) are fitted to experimental stability measurements. They are fitted to data they are intended to reproduce.

For structure prediction, sampling algorithm applying an energy function is needed. Energy functions are tested at biennally CASP competitions. Rosetta program [119], which is particularly successful in these competitions, appeared in 1999 at IIIth CASP. In Rosetta, simulation annealing in torsion space representation is performed to sample structures evaluated using statistical potential. Review of global optimization methods was published by Wales and Scheraga [120]. Review of structure prediction methods with emphasis on energy functions has been published by Prentiss et al [121].

1.3.5 Energy functions for mutants

Energy functions for predction of stability change upon amino acid replacement must deal with change both in native and denatured state but do not face the problem of enormous error propagation. They can be classified into 4 groups [122]:

- First principle methods. Free energy is calculated using detailed atomic models.
- Statistical potential.
- Force fields combined with empirical parameters fitted to experimental data. These methods are most relevant to the work presented in this thesis.
- Machine learning methods.

The mentioned energy functions have been integrated to various software packages or web services. Eris [123] by Dokholyan group uses Medusa force field and was tested against 595 mutants. In Medusa force field, free energy change upon mutation is calculated as a sum of 8 energy terms including van der Waals interactions, sidechain and backbone hydrogen bonding, solvation energy and internal degrees of freedom. Electrostatics is omitted. Scoring function of Eris comprises 8 empirical parameters. FoldX is based on FoldX energy function [124] comprising 10 empirical parameters.

Performance of stability predictors upon mutation is evaluated in recent article by Potapov et el. Prediction performances assessed by Potapov et al. [47] can be underestimated; for example FoldX stability predictions disregard pH. If data with different pH were omitted, correlation would probably be better. Also as mentioned above, Pearson correlation coefficient of experimental data published by different groups is 0.86. Therefore, best Pearson correlation coefficient of predicted and experimental data is about 0.55.

1.4 Intramolecular interactions in native state

1.4.1 Covalent and non-covalen interactions in native state

Protein folding is driven by weak interaction interactions, particularly non-covalent interactions and torsional strains. Strong interactions, called also stiff degrees of freedom (SDoFs), confine configurational space of all the protein ensembles into a narrow complex subspace, which is difficult to sample. There are three types

of SDoFs in non-polarizable empirical force fields. First, Lennard-Jones repulson term defines the shape of residues ("lego set"). Second, covalent bonds which remain stable and their lengths oscillate a little around their equilibrium values. Nevertheless, their energy increase with a disturbance is such stiff that they do not significantly contribute to energy.

However, covalent sidechain bridges are interesting for protein stability research since they strongly bridge long-ranged residues (in sequence), so protein is no longer a linear polymer. Anfinsen in his experiments [8] showed that proteins can fold to correct native structures even if S-S bridges are broken. Effect of S-S bridges on protein stability is still not understood. Thornton suggested that they contribute 2.3 kcal/mol [125] to protein stability per S-S bridge. Beside disulfide bonds, other covalent sidechain bridges has been found for example sulfilimine bonds which have been recently studied by Oncak et al. [126].

Ionic bonds are so strong in terms of bonding energy that vaporization temperatures of ionic species are very high. Therefore, one may expect that pairs of ions with opposing charges will contribute to stiff, or strong, degrees of freedom. However, as mentioned above, solvation decreases actual bonding energy of ionic species to low values. Second, energy of ionic bond decreases slowly with distance which allows more flexibility and increases opposing entropic effect.

Pitzer (torsional) strain can be classified as soft degree of freedom. Rotamers are direct consequence of torsional energy barriers which are usually in order of 5 kcal/mol.

1.4.2 Distance scaling

From statistical mechanics point of view, the most important features of an interaction are additivity distance scaling Distance scaling simplifies PES, characterizes density of states, and therefore effect of entropy. Distance dependence of non-covalent interaction energy can be expanded into polynome in $\frac{1}{r}$

$$IE = \sum_{i=1}^{\infty} a_i \, r^{-i} \tag{1.11}$$

Non-covalent interactions can be naturally classified into short-ranged and long-ranged. Contribution of distant short-ranged interactions vanish

$$4\pi \int_0^{\infty} IE \; r^2 dr < \infty \tag{1.12}$$

which is true for all interactions $IE \propto r^{-\mathrm{n}}$ for $n > 3$. Long-ranged interactions ($\mathrm{n} \leq 3$) vanish only in presence of opposite interactions, for eqample in ionic crystals. These sums converge slowly in original (or direct or standard 3D) space but rapidly in reciprocal space. Therefore, Ewald summation [127] and augmented Ewald summation [128] is used in molecular modeling for treatment of electrostatics. Errors of these methods are discussed in [129] where correction formulas for charge distribution are proposed and in [130] and [131].

1.4.3 Classification in quantum chemistry

Interaction energy of systems A and B si defined as

Definition 6. *Interaction energy between systems A and B is defined as*

$$E_{A...B} = E_{AB} - (E_A + E_B)$$

where E_{AB} is energy of system involving both system A and B, E_A energy of system A and E_B energy fo system B.

This definition serves also as a standard scheme for its calculation. Quantum chemistry methods calculate ground state energy of the system defined by positions of nuclei (Born-Oppenheimer approximation), number of electrons and their spin. The energy is in order of 38 a.u. per heavy atom (carbon), while interaction energies are in order of 10^{-3} a.u. per heavy atom and stabilities of proteins are in order of 10^{-5} a.u. Interaction energies of pairs of small biologically relevant molecules can be calculated using "golden standard" of quantum chemistry, i.e. the best feasible method which is believed to reach level of sub-chemical accuracy (0.1 kcal/mol interactions), CCSD(T)/CBS. [4] Computational cost of these interaction energy calculations increases with 7th power of number of basefunctions (equivalent to number of heavy atoms). Calculations involving up to 30 heavy atoms are feasible [62].

An alternative method of calculation is Symmetry-adapted perturbation theory (SAPT) in which interaction Hamiltonian is treated as perturbation of sum Hamiltonian. Interaction energy is obtained as energy perturbation. First perturbation order represents electrostatic energy and second order represents induction and dispersion energy.

$$E_{IE} = E^{(1)}_{elst} + E^{(1)}_{exch} + E^{(2)}_{ind} + E^{(2)}_{disp} + \delta HF \tag{1.13}$$

SAPT calculations [132] are very computationally expensive compared to their accuracy (7th power with number of basefunctions). Therefore, DFT-SAPT method [133] based on density functional theory (DFT) was proposed, which scales with 5th power of number of basefunctions. From quantum chemistry point of view, each interaction can be decomposed and classified according to its dominant term.

1.4.4 Classification in biology

In biology, some types intramolecular interactions between amino acids have been given names. Oppositely charged ion pairs separated less than about 4 Å are called salt bridges. Hydrogen bond is defined by IUPAC as an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation [134]. In proteins, it can be realized between backbones, polar sidechains and charged sidechains (pairwise, i.e. 5 possibilities, since charged-charged sidechain form salt bridges). Another type of interactions, cation - π between charged and aromatic sidechains are recognized by wide community. Other non-specific dispersive interactions are called van der Waals bonds. Non-specific electrostatic interactions remain nameless for biologists.

[4] Coupled cluster of second order and perturbative treatment of triple excitations. For accurate calculations, zero point vibration energy must be calculated, basis set superposition treated and energy extrapolated to complete basis set limit.

In 2009, Berka [135] decomposed representative sidechain interaction energies using SAPT method. He showed that induction and dispersion energy significantly contribute especially to hydrogen bonds. It is therefore worth to discuss quality of their description in cheap methods.

For studies of biomolecules, sampling is so important that decrease of description quality for each microstate is tolerated, since entropy term estimate from single structure using rigid rotor harmonic oscillator (RRHO) approximation is insufficient. It has been found that empirical force fields are in surprising agreement with benchmark (CCSD(T)) calculations[65]. If one is not interested in quantum effects (ZPVE, tunneling, resonance), empirical force fields provide good approximation of true potential.

Reluctance of biomolecular modeling community to polarizable force fields is an unfortunate paragon of conservative attitude in science. Computational cost increase of in simulations using always stable predictor-corrector integrator [136, 137] with introduction of polarization through Drude model would be much smaller than computational cost increase with introduction of explicit solvent. Also implementation would be straightforward and no iteration would be needed.

1.4.5 Contribution of interactions to stability

Free energy of an interaction is a collective property and cannot be determined from single microstate. However, methods of estimation of entropic term (for example RRHO) have been proposed to prevent computationaly expensive sampling. Free energy contribution of a native state residue-residue interaction can be experimentally determined by double mutant cycles.

Relationship between interaction energy and free energy contribution of an interaction to protein stability remains unsolved. Most usual approach in structural biology is definition of interaction type based on some orientational and distance thresholds and assigning a uniqe constant value to each interaction type [138, 117]. Interaction of the same type, which is stronger in terms of interaciton energy retains same strength in terms of free energy. Weighting of these arbitrary values by distance or mutual contact surface area [139]. Another approach is using a dielectric constant which is equivalent to defining function for free energy as

$$FE = IE_{LJ} + \frac{IE_{el}}{\varepsilon_r} \tag{1.14}$$

where FE is free energy of an interaction, IE_{LJ} is its Lennard-Jones component, IE_{el} its Coulombic component and ε_r dielectric constant of environment. Dielectric constant close to 80 (value for water) describes interaction which is fully solvated after breakage, so no persistence of contacts in denatured state is assumed. If effect of denatured state is not included, an improved method of entropy estimate would well stand for IE $\rightarrow$ FE mapping. It is plausible to assume that interaction with lower interaction energy value is more stabilizing than another interaction of the same type (for example arginine - glutamate) with higher interqaction energy value.

Hypothesis 5. *Transformation of interaction energies pertaining to the same class into their free energy contributions to protein stabilization is monotonic. In*

mathematical expression:

$$IE_1 \geq IE_2 \Rightarrow FE_1 \geq FE_2$$

where IE_1 and IE_2 are pairwise interaction energies and FE_1 and FE_2 are corresponding free energy contributions to protein stability.

which is a weaker assumption than using dielectric constant. Monotonicity of IE $\rightarrow$ FE mapping is surely approximative but widely accepted.

1.5 Mathematical representation of protein structure and stability

1.5.1 Representation of protein organization by matrices

Internal organization of a protein of N residues can be well represented by N $\times$ N matrix having in each field a_{ij} an association measure of ith and jth residues in sequence. The association measure is a real number which can vary from distance between Cα atoms, through correlation of residues' motion in a simulation to interaction energy, depending on the property studied. Such matrix is symemtric and represents a graph, vertices corresponding to residues and edges corresponding to their interactions. Advantage over simple listing or summing interactions is inclusion of sequence information, therefore vertices cannot be arranged in different order. Such graphs are small-world networks, known from other fields of physics [140].

It is generally accepted that proximity in 3D structure implies interaction in energy and function means. However, such assumption is unjustified. Distance is actually a very rough measure of communication of a residue pair. For identification of most stabilizing residues or allosteric communication network studies, an actual energetic measure might be more beneficial than just mutual distance. Introduction of physical character of residue-residue interactions is promising for example in revealing relationship between protein flexibility and stability. Possible applications to protein stability studies are discussed in [141].

Residue-residue pairwise association measures are usually simplified to contacts, boolean quantities being 1 if the measure exceeds some arbitrary predefined threshold and being 0 otherwise. The loss of accuracy is compensated by 2 advantages. First, contact matrices are sparse, the number of contacts scales roughly linearly with number of residues. Second, contacts are additive, whereas mutual distance is not additive and interaction energies between 2 hydrophobic residues and between 2 charged residues are not comparable. Contacts are often used to describe topology of proteins by identification of "long-range" contacts. This range means sequential distance and has nothing common with long-ranged interactions defined in section 1.4.2. Plaxco and Baker [142] have found correlation between topology measured by contact order (Equation 1.15) and

$$CO = \sum^{N} \frac{\Delta S_{ij}}{L\ N} \tag{1.15}$$

where ΔS_{ij} is separation of residues i and j (in contact), N is number of contacts and L is number of residues. Threshold for contact definition has been

optimized by Gromiha [143] to maximize the correlation. Other measures of protein topology have been proposed [140].

It is worth to remind that in all mentioned studies, only one structure representing native state ensemble is studied as validity of hypothesis 2 is assumed.

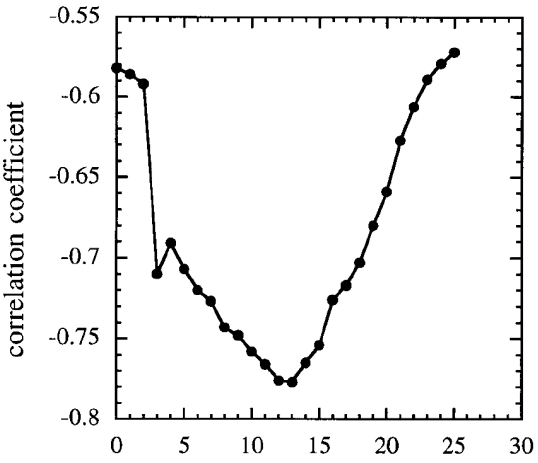
1.5.2 Geometry representation

Distance matrices (DMs) are intended to represent protein geometry in a way useful for particular studies. DMs are standard structure representation in structure determination studies by NMR. Contact maps derived from DMs have been extensively used in structural biology and bioinformatics for structure alignment by DALI [144], MatAlign, by combinatorial extension of the optimal path or using contact map overlap[145]. Distance-based contacts are also essential for fold recognition [146].

There are many posible definitions of distance-based contacts. First, there is a variety of measures of proximity of two amino acids. Geometry of each residue is usually reduced to one point and distance between such two points is measured. Such approach is faster than averaging distances between all pairs of atoms and no advantage of the latter has been found. The point representing a residue can be for example Cα atom, Cβ atom or average position of all heavy atoms. Second, cutoff value is arbitrary. The following definitions are used.

- 5.4 Å separation of Cα atoms reproduces length of one turn in alpha-helix.
- 6 Å is separation at which occurence of oppositely charged ion pairs becomes uncorrelated [147].
- 4 Å is separation of atoms at which most Lennard-Jones interactions cease.
- 11 Å separation between Cβ atoms is optimal for reproduction of structure from contact maps [148].
- 8 Å separation of Cα atoms is reported by Gromiha and Selvaraj [143] as optimal definition of contact for folding rate prediction (Figure 1.5.2).
- 7 Å separation between Cα atoms was used in elastic network model studies by Kundu et el [149].
- 8 Å, 10 Å and 21 Å cutoffs are used in automated class assignment [146] depending on the types of interacting secondary structures.

Figure 1.3: Optimization of cutoff value to maximize correlation between folding rate (logarithm of folding rate) and long-range contact order. Abscissa - Cα distance cutoff in Å. Ordinate - correlation coefficient. Adapted from ref. [143].



Vendruscolo and Domany [150] studied protein dynamics in contact map space and came to conclusion that searching contact map space is more efficient than searching space of possible as changing few contacts in contact matrix corresponds to a large move in conformational space[151, 152]. Contact map can be also reduced to vector from which protein structure can be recovered [153]. Formulation of protein folding as mapping from sequence space to space of principal eigenvectors of contact maps is probably its most simplified formulation. [5] The authors (Vendruscolo and Domany) predict that itroduction of an energy function discriminating the native state from decoys might make the contact map search applicable for structure prediction. However, such potentials are very difficult to construct. As mentioned above, contribution of an interaction highly depends on proximity and mutual orientation of the pair of residues. Error of energy assigned to a contact, which completely neglects orientation and distance, is such high that error of its sum is magnitudes higher than Khatun et al. argue that it is impossible to reach experimental accuracy using simple contact potentials [154].

1.5.3 Energy representation

Physical interactions between residues can be introduced by three different ways. First, normal mode analysis [155] of proteins can provide measure of dynamic correlation for each residue pair. Flexibility of proteins can be studied by well established and simple Gaussian network model [156] or by Distance constraint model [73]. Mean square fluctuations calculated using GNM surprisingly well correlate with experimental temperature factors [149]. Flexibility of proteins is of interest not only for allostery studies but also stability studies. For structure

[5] mapping from space of N-digit numbers in base 20 to space of N-dimensional vectors of positive real numbers. Representation of structure by 2N-dimensional real vector of ϕ and ψ angles is inconvenient since any physically meaningful energy function is extremely sensitive to errors in some torsional angles.

prediction, fast configurational entropy estimates of higher accuracy than simple RRHO [157] are desirable. GNM can be used to identify residues critical for conformational transitions in protein structures [158].

Second approach of including energetics is based on quasi-chemical approximation. The potential proposed by Miyazawa and Jernigan [159, 160] comprises 210 parameters - energies corresponding to probabilities of proximity of particular pairs of residues. They are well suited for lattice models.

Third approach is calculation of physical interaction energy between each pair of residues. In 2008, our group [161] introduced interaction energy matrix (IEM) concept for identification of most stabilizing residues. Interaction energies between sidechains were calculated using GB solvent model. Sum of interaction energies of one residue with all the others was proposed as a measure of its contribution to overall stability of Trp-cage miniprotein. Key residues for stability are the ones with highest interaction with all the other residues. Unfortunately, we still do not know, how to exploit sequence information contained in IEMs to improve estimate of free energy from interaction energy.

Methods of computational chemistry are well developed and enable calculations of interaction energies on level of sub-chemical accuracy (0.1 kcal/mol). Berka et al calculated benchmark interaction energies of representative interactions between sidechain analogues and decomposed them using DFT-SAPT [135, 162]. It seems that fragmentation of a protein and then calculation of pairwise interaction energies of contacting fragments could lead to accurate description of protein energetics scaling roughly linearly with protein size [163]. Unfortunately, utilization of quantum mechanics methods requires artifical fragmentation schemes like Cα representation of sidechains [135, 162]. Contribution of 3-body interactions could also be high enough to introduce substantial errors. Moreover, entropy estimates using RRHO approximation are inaccurate for biomolecules. As mentioned before, even sub-chemical accuracy is insufficient for accurate calculation of protein stability. Nevertheless, identification and diminishing the most significant errors can in future lead to accurate energy functions for protein structures.

1.6 Aims of the thesis

The main aims of this thesis are to

1. propose treatment for proper modeling of solvent effects on intramolecular interactions
2. characterize distributions of residue-residue interaction energies, eventually propose a model for the distributions
3. propose a unified treatment for all interactions in interaction energy matrices
4. discuss contribution of particular interaction ekergies to protein stability
5. study relationship between sidechain interaction energies and secondary structure

1.7 Organisation of the thesis

Thesis is based on 2 attached papers (Appendix 1 and Appendix 2) and yet unpublished data which are in stage of preparation for publishing. Results published in the 2 papers are not duplicated in the Results section. Instead, they are complemented, summarized and re-evaluated or just briefly summarized. Methods section contains methodology details of work presented in the Results section. Methodology of work presented in attached papers and only reviewed in Results section can be found in the corresponding papers.

- Paper in Appendix 1 was published in form of peer-reviewed open access book chapter [164]. In the paper, we present results of our studies of interaction energy distributions. Proper modeling of solvent effects is addressed, relationship between interaction energies and secondary structure content is studied.
- Paper in Appendix 2 was submitted to Journal of Physical Chemistry B on 4th April 2012. In the paper 2, we propose new definition of residue-residue contact based on interaction energy calculations. Thresholds were set independently for each interaction energy type. Classification of sidechains is justified and a little different from those in Paper 1.
- The core of Results section is model of stability decomposition into 1-body and 2-body free energy contributions, which has not been prepared in the form of paper yet.

2. Methods

2.1 Structures of proteins

2.1.1 Structure set selections

Structures were selected from PDB open database to represent wide variety of structural information. For all the studies presented in sections 3.1 to 3.6, the same basic structure set of 1358 protein structures was used. We selected only protein molecules with one chain, no ligands, resolved by the X-ray crystallography method at a minimum resolution of 2.0 Å. We also omitted structures with a 70% sequence identity and higher. 1531 structures were returned by "advanced search" in PDB database (download Jan 31, 2011). Unfortunately, inconsistences in structures such as missing backbone atoms or more residues forced us to omit 173 structures. We selected 1358 structures, in which we had higher trust. The characteristics of the structure set are illustrated in Figure 1 of Appendix 1.

In Appendix 1, we also address the question of the residue selectivity for secondary structure motifs. We therefore constructed additional structure sets from structures of the mentioned set of 1358 structures based on their size and secondary structure content. As we have found that average RIE strongly depends on protein size, the set for secondary structure - RIE relationship had to be consistent in chain length distributions. We selected 99 structures to each set, their properties are summarizes in Appendix 1, section 2.1.1. Structure sets for size - RIE relationship studies were not equal in number of structures, see Appendix 1, section 2.3.

2.1.2 Structure preparation

Amino acids lacking sidechain heavy atoms were turned into glycines. If backbone atoms were missing, the structure was ommited (therefore only 1358 structures). Hydrogens were added by *pdb2gmx* procedure implemented in GROMACS package at pH 5.5 (histidines were always double protonated and therefore charged) and optimized as int the whole structure (not pairwise).

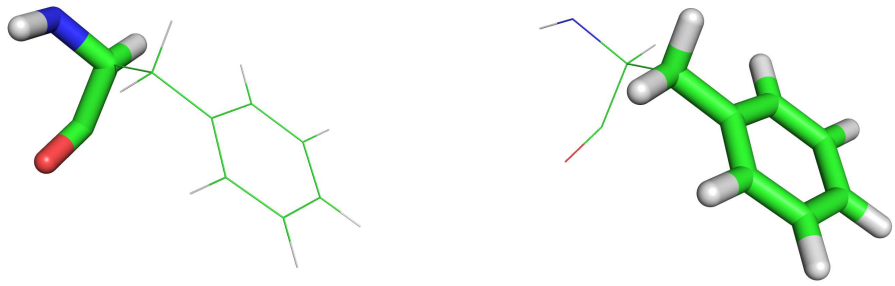
2.2 Energy calculations

2.2.1 Fragmentation

In previous works by Berka et al. [135, 162] and in my bachelor thesis[165], Cα representation was widely used. In this representation, only sidechain interaction is interesting and backbone atoms are replaced by methyl groups, so methane represents glycine, ethane represents alanine and so on. This fragmentation is used in Appendix 1 to show that OPLS RIE distribution are in perfect agreement with Amber03 distributions. Interaction energy calculation using this fragmentation is straightforward and discussed in [135]. Pitfall of this fragmentation is arbitrary Cα modification of existing force field and disregarding backbones interactions. The fragmentation enabled comparison of force fields with higher level methods.

In Appendix 1, backbone interactions were included. A protein containing N residues of which M (M < N) were glycins was fragmented into N backbone and N-M sidechain fragments, since glycine possesses only backbone atoms.

Figure 2.1: fragmentation of amino acids into backbone (BB, left) and sidechain (SC, right). Cα atom is assigned to backbone.



In Appendix 1, sidechain fragments were classified as follows. charged sidechains (abbreviated to CH - asp, glu, lys, arg, his), polar sidechains (abbr. PO - asn, gln, thr, ser) and non-polar sidechains (abbr. NP - ala, leu, ile, val, pro, cys, met, phe, tyr, trp). In Appendix 2, fragmentation scheme was slightly changed to reflect similarities of interaction energy distributions. The only change was classification of tyr and trp as polar (instead of non-polar in Appendix 1).

2.2.2 Interaction energy matrix

Interaction energy was calculated between each pair of fragments which were not covalently bound as sum of Lennard-Jones and Coulombic terms (Equations 2.1 and 2.2) for each pair of atom, one from each residue. A protein having N residues, of which M are glycines has $2N^2 - 4MN + M^2/2 - 3N + 3/2M + 1 - O$ interaction energies, where O is number of S-S bridges. No covalent bonding terms were used.

$$U_{coulomb} = \sum_{i=1}^{i<N} \sum_{j=i+1}^{j<N+1} 332 \frac{q_i q_j}{r_{ij}} \tag{2.1}$$

$$U_{vdW} = \sum_{i=1}^{i<N} \sum_{j=i+1}^{j<N+1} 4 \cdot \varepsilon_{i,j} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.2}$$

where r_{ij} is distance between atoms, σ_{ij} and ε_{ij} are calculated for every pair from particular atom parameters ε and σ. If partial charges of atoms is given in units of elementary charges and distance is given in nm, $U_{coulomb}$ is in kcal/mol.

Parameters were taken from GROMACS implementation of OPLS [166] force field. The only change made was in proline, where partial charge of N atom was changed from -0.14 to -0.07 and of CD atom changed from to maintain electroneutrality of fragments. To examine robustness against force field used, we performed the calculations in Appendix 2 using CHARMM27 [167] force field, which also enables fragmentation with electroneutral (or properly charged in case of charged

sidechains) fragments. Amber force field was not used in this fragmentation since any attempt to neutralize fragments leads to significant modification of force field.

The classification of the amino-acid atoms in four groups resulted in ten types of mutual interactions - BB-BB, BB-CH, BB-PO, BB-NP, CH-CH, CH-PO, CH-NP, PO-PO, PO-NP, NP-NP. Interaction energies were collected in 10 separate interaction energy matrices of size $N \times N$. It is guaranteed that no interaction energy is counted twice, so the sum of all of the matrices provides the interaction energy between the corresponding residues, but the matrices are symmetric and each interaction energy is twice recorded in corresponding matrix.

2.2.3 Residue interaction energy

In order to compare the residual energy content, we have introduced a residue interaction energy (RIE) characteristic for each residue. The definition and calculation is similar as in [161] with the only difference that our RIE values are classified into 7 types. In other words, NP-NP RIE of a residue is sum of all numbers in line (or column) corresponding to the residue in NP-NP matrix. It is by definition 0 for all polar and charged amino acids. Therefore, sidechain-sidechain IEMs contain many zero values, while matrices containing interactions of backbones have exactly zero values only for covalently bound fragments.

RIE distributions can be done for set of all residues from all the proteins or for one protein only. In Appendix 1, mostly the latter is presented. It is also averaged throughout the set of proteins to get average distribution of RIE. Details of averaging can be found in section 2.1.4 of Appendix 1. Calculation of RIE distributions and HCIE curves (histograms of IE distributions multiplied by IE) is described in detail in Appendices 1 and 2 respectively.

2.2.4 Solvent accessible surface area

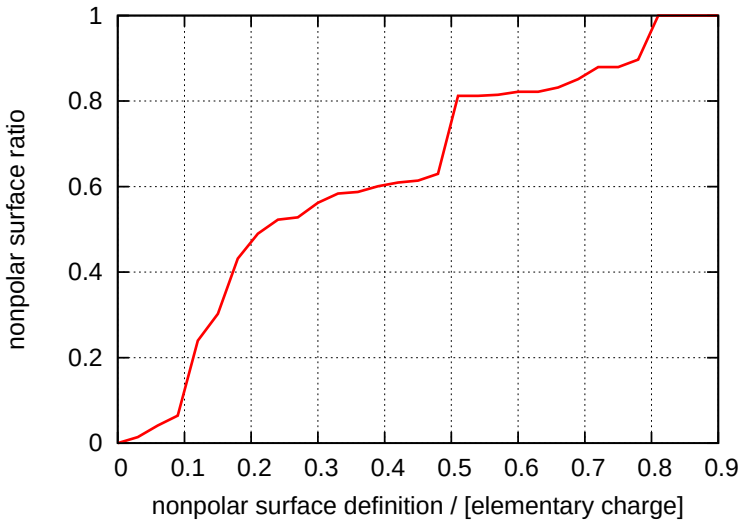
Stability decomposition model described in section 3.6 uses a simplistic macroscopic solvation model. Surface was classified as non-polar, polar and charged based on partial charge of atom. Non-polar surface was defined as surface of atoms with absolute value of partial charge between 0 and 0.35 [1], polar surface between 0.35 and 0.65 and very polar (or charged) surface over 0.65. These values were chosen as the values dividing the distribution of surface polarity (Figure 2.2.4) into well-defined regions. The calculation is performed using g_sas routine implemented in GROMACS package.

2.2.5 Molecular dynamics

For examination of hypothesis 2, structure of Fusarium Solani Cutinase (PDBID 3QPA) was selected for molecular dynamics study of IEMs in native state ensemble. 3QPA is very suitable, since it is resolved at 0.85 Å resolution, contains both helices and β-sheets and all types of residue-residue interactions. Position of hydrogens in the raw structure were optimized using OPLS/AA force field in explicit solvent (SPC/E).

[1] if not stated, charges are given in units of elementary charge, i.e. 1.602E-19 C.

Figure 2.2: Average distribution of surface polarity of the proteins in our structure set. As non-polar surface definition shifts to higher partial charge (parameter of g_sas routine), its area grows.



Then, the structure was equilibrated for 200 ps and then simulated for 3.8 ns using leap-frog integrator implemented in GROMACS package. Temperature was kept using V-rescale thermostat by Bussi et al. [168] and pressure using Berendsen barostat. Periodic boundary conditions were used and Particle Mesh Ewald treatment for long-range electrostatics.

All equilibrations and consecutive simulations were performed at 6 different temperatures. For each run, 190 IEMs were calculated on snapshots (no optimization) from simulation after every 20 ps. IEM was also calculated for single X-ray structure for comparison.

2.3 Parametrization of models

Parametrization of a model of the form

$$\mathbf{A}\ \mathbf{p} = \mathbf{0} \tag{2.3}$$

comprising n parameters (scaling factors) to best reproduce m experimental data (A is of size $m \times n$, $m >> n$) is non-trivial. Least squares method cannot be used as overdetermined system has single solution p = 0. If none of the contributions is dominant, setting one of the parameters to 1 biases the result. We are interested only in relative numbers of scaling factors.

Here we propose solution to the problem using Courant-Fischer-Weil min-max theorem, from which it follows that expression

$$\frac{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \tag{2.4}$$

reaches minimum in $\mathbf{x}$ if $\mathbf{x}$ is eigenvector of $\mathbf{A}$ corresponding to lowest eigenvalue. Particularly in calculations presented in section 3.6, matrix $\mathbf{A}$ of size

1188×33 is multiplied by its transpose to yield matrix $\mathbf{B}$ of size 33×33, from which $\mathbf{x}$ is calculated as the eigenvector of size 33 corresponding to lowest eigenvalue. Calculation were done using Octave interpreted language, routine $eig()$ implemented in Octave was used for calculation of eigenvectors.

3. Results

3.1 Classification of amino acids

In this work we propose a natural fragmentation of all-atom force field model of globular protein. Amino acids are divided into backbones and sidechains and sidechains are classified into 3 groups - charged, polar and non-polar. Aim of the sidechain classification is grouping interactions with similar properties, such as distance scaling (and therefore entropy contribution) and similar interactions with water.

We assume validity of monotonicity of IE $\rightarrow$ FE mapping for IEs in particular class. Despite lacking structural knowledge of denatured state ensemble and disregarding dynamics and cooperativity of bonds, such hypothesis seems close to reality. Validity of this hypothesis is assumed in variety of works. It is usually assumed, that all fragments should be in one group with a slightly modified scaling scheme (Equation 1.14).

Backbones were selected because their interactions form special structural paterns and hydrogen bonds between backbone atoms are abundant in comparison to sidechain polar-polar hydrogen bonds. Classification of sidechain fragments into groups was examined in Appendix 2, where similarity in distribution of interaction energies was required.

3.2 Residue interaction energy

In Appendix 1, we propose reisude interaction energy (RIE) as a measure of contribution of an amino acid to overall stability of the protein. Residue interaction energy (RIE) is a property of 1 amino acid which can be calculated from the corresponding IEMs as the sum of all IEs in which the amino acid takes part. The method is similar to the work of Bendova et al. [161]. The difference is in treatment of solvent by classification of RIEs in accord with 10 types of contributing interactions resulting in 7 RIE values for each amino acid. The only exception is glycine which lacks sidechain and therefore has only 4 RIEs.

Distribution of RIE can be easily characterized by median or average. Unfortunately, we could not propose a physical model of distributions since interactions should be scaled and all types merged, since their distribution is not independent. At least, we can propose mathematical description of the RIE distribution curves involving few parameters (see [165]). From distributions of RIE, it can be concluded that there are residues with significantly strong residue-residue interactions which significantly contribute to stabilization.

Following our hypothesis that IEs of particular type are comparable and can be summed, RIE is a measure of contribution of particular amino acid to stability of the protein. RIE distribution profiles in proteins are studied.

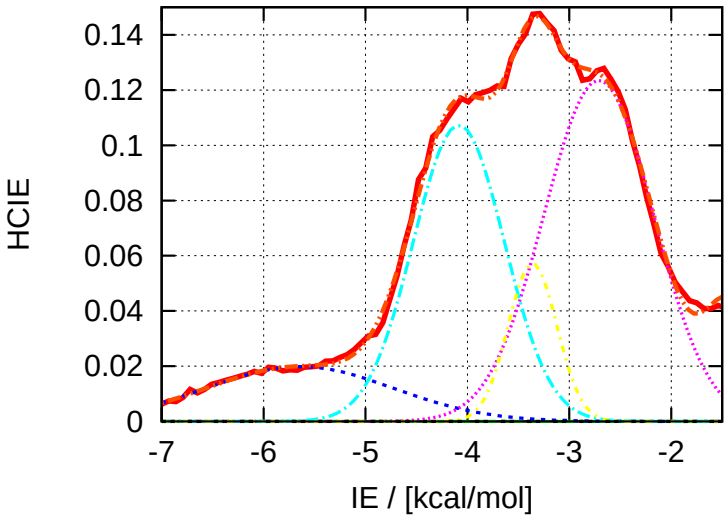
3.3 Domain size

Domain is thermodynamically defined as an independently folding subunit of a protein (see section 1.1.2). We define domain as a protein subunit with its own hydrophobic core. Hydrophobic cores of globular proteins are bounded by backbones or polar sidechains. Since the size of globular cores is limited, large proteins must contain more cores. In Appendix 1, we present 2 simple models of globular hydrophobic cores. Average RIE of non-polar residues increases with protein size, since the number of non-polar residues located at core boundaries becomes lower relative to number of residues inside the core, having more contacts. This increase is limited by maximum size of one hydrophobic core. We estimated size of the domain to about 110 amino acids.

3.4 Interaction energy distributions

In Appendix 2, we present distributions of residue-residue interaction energies. As number of residue pairs grows with square number of residues in a protein, most of their interaction energies are zero. Therefore, average or median of interaction energies has no physical meaning. We have suggested an alternative representation of IE distributions, HCIE curves. Their construction is discussed in Methods section of Appendix 2 and their properties are reviewed in the beginnings of the Results and Discussion section of Appendix 2. A HCIE curve is equivalent to IE histogram multiplied by IE (Figure 3.1).

Figure 3.1: IE histogram and HCIE of backbone-backbone interacitons with fitted particular Gaussians. Only region of productive contacts was selected.



We have found that HCIE can be well approximated by sum of Gaussians multiplied by IE. Largest error of such description can be found in the region of diverging zero contacts ($IE \to 0$).

3.5 Definition of residue-residue contact

Geometric definition of residue-residue contact disregards strength of interaction between residues. Success of models decomposing stability into 1-body contributions neglecting native state interactions [112] indicates that energy is so balanced that only significantly strong interactions play role. In Appendix 2, we propose such definition based on HCIE curves as an interaction energy value dividing regions of significantly strong (we call them productive) contacts from the weak (we call them bulk) interactions.

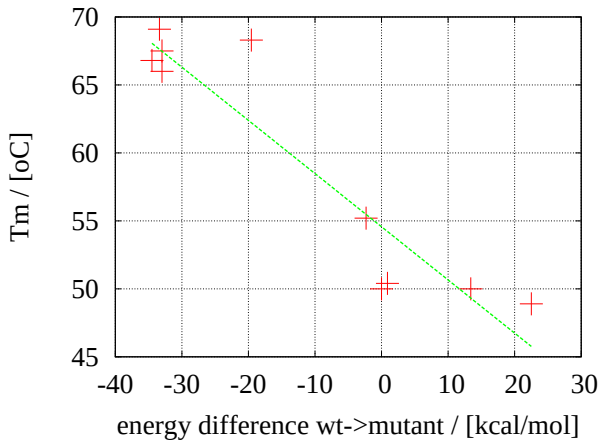
In addition to assuming validity of hypotheses 1, 2, 4 and 5, we assume that peaks in HCIE curves correspond to interaction patterns and they can be used to discriminate between productive and bulk interactions. A HCIE curve contains multiple stationary points, so we employed additional criteria to identify the one corresponding to optimum contact definition. Contact number (average number of contacts per residue) for productive contacts was expected to be more restrictive than the one for geometric contacts. Therefore, optimum contact definition were expected to result in contact numbers in order of units. At most 3 stationary points were possible candidates for a productive contact definition (see Appendix 2), so we suppose that the selection of the contact definition resulting in contact numbers consistent with current understanding of contacts However, there is a possibility that our view was biased by current recognition of residue-residue interaciton strength.

We have shown that using different force field (CHARMM) leads to very similar contact matrices. From comparison of contact matrix similarities at fixed average contact numbers (Table 2 in Appendix 2), it is obvious that contact definition based on interaction energies is less ambiguous than any geometry-based definition. We have also shown that thermostable proteins contain higher number of contacts defined based on interaction energy than do mesostable proteins. Contact definition might find its application in flexibility models, identification of aminoacids key to stability and in identification of domains as clusters of productive contacts. The view of cooperatively folding domains as clusters of energy-based contact is plausible because of entropic effects connected with non-covalent bonding of coupled bonding sites [169].

3.6 Interaction energy balance in proteins

This project was motivated by successful application of the interaction energy matrix concept on prediction of the melting temperatures for sequence variants of haloalkane dehalogenases. based on change of summed interaction energies of uncharged sidechains calculated in native state structures. We could not include charged residues, since their interaction energies are incomparable to interaction energies between uncharged residues and from the beginning, we did not trust using dielectric constant.

Figure 3.2: Correlation between melting temperatures and interaction energy differences between mutants and wild original haloalkane dehalogenase.



The correlation was, surprisingly, better than the one obtained acknowledged web applications (e.g. FoldX). This result might be only a unique property of the tested system. However, interaction energy matrix approach to stability prediction from structure has a lot of space for further improvement.

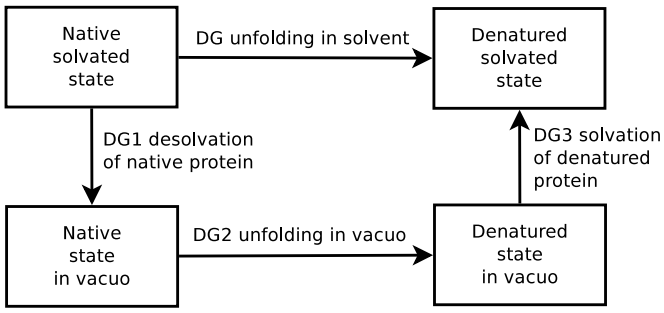
3.6.1 Derivation of the model

We hereby propose a model of protein stability decomposition into separate free energy terms based on physical considerations. The model assumes additivity of free energy contributions (hypothesis 1.9) and linear dependence of force field interaction energy of a residue-residue interaction and its free energy contribution to overall stability. Free energy of folding is formally expanded into 1-body and 2-body terms.

$$\Delta G_{fold} = \sum_{i=1}^{N} \Delta G_i^{(1)} + \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta G_{ij}^{(2)} + \delta \Delta G^{(3)} \tag{3.1}$$

where N is number of residues in the protein. 1-body terms represent stability contributions of single amino acids and 2-body terms represent pairwise stability contributions. These terms are calculated from native structure. Thermodynamic cycle (Figure 3.3) provides a physically intuitive picture of stability contributions.

Figure 3.3: Thermodynamic cycle divides protein unfolding into 3 steps - desolvation of native state, unfolding *invacuo* and solvation of denatured state.



- ΔG_1 - Solvation free energy in the native state.

  Using this macroscopic solvation model, contribution is **1-body** and can be approximated by a linear function of and non-polar solvent accessible surface areas of the native state.

- ΔG_2 - Folding free energy in vacuo.

  1-body free energy contribution comes from configurational entropy change and internal energy change, e.g. torsional strain.

  2-body free energy contribution is a difference of residue-residue interaction energies between native and denatured state corrected for entropic effects. Assuming that proportion of persistence of interactions of a particular type in denatured state is universal for all proteins, this free energy contribution can be calculated as sum of interaction energies of that type multiplied by a universal scaling factor.

- ΔG_3 - Solvation free energy of the denatured state.

  Denatured state is approximated by an unfolded state with presence of residual native state contacts. Macroscopic model of solvation leads to **1-body** contributions of amino acids to solvation of unfolded state.

To summarize, 1-body contributions comprise solvation of the native and denatured state and configurational entropy. The latter two can be calculated as sums of amino acids of particular type multiplied by universal scaling factors representing their characteristic solvation energies and configurational entropies. In addition to the mentioned, 1-body contribution of an amino acid also involves average interaction of the amino acid in the native state. Solvation of the native state can be calculated as a linear transformation of surface areas (see methods section 2.2.4). 2-body interactions comprise residue-residue interactions reduced by their persistence in denatured state and opposing entropic effect. The reduction is modeled by linear transformation of interaction energies to free energy contributions, which is quite stronger assumption than hypothesis 5. Stability of a protein can be written as

$$\Delta G_{fold} = \sum_{i=1}^{20} a_i \; n_i + \sum_{j=1}^{10} IE_j \; b_j + \sum_{k=1}^{3} SAS_k \; c_k \tag{3.2}$$

where i's denote amino acid types, j's denote interaction energy types and k's surface types. n_i is number of residues of type i (dimensionless), IE_j sum of interaction energies of type j (in kcal/mol) and SAS_k area of native state surface of type k (in Å^2). There are 33 scaling factors, 20 a_i's (in kcal/mol), 10 b_j (dimensionless) and 3 c_k (in kcal/(molÅ^2)).

3.6.2 Interpretation of the scaling factors

Scaling factors calculated using Courant-Fischer-Weyl theorem (see section 2.3) to best represent stabilites of 1188 proteins are summarized in Table 3.1. First, it is worth to remind that their values are relative to each other, therefore one scaling factor must be set to arbitrary value. Largest 2-body scaling factor was set to 1. Such selection of scale leads to 1-body scaling factors in the order of magnitude of solvation free energies (units to decades of kcal/mol). Expected values of 2-body scaling factors are between 0 and 1, since they originate in entropic compensation of interaction energies and persistence of interactions in denatured state. Both the mentioned effects are destabilizing while interaction energies are stabilizing. In Table 3.1, destabilizing terms are negative and stabilizing terms positive.

Highest stabilizing contribution (42%) comes from solvation of native state. This result might seem surprising, but average value of this free energy contribution (700 kcal/mol) accords with usual solvation free energies of proteins. Contribution of surface solvation to stability also decreases with increasing protein size, as expected. High contribution comes also from interactions between backbones and non-polar sidechains. This term is independent of protein size and seems to be a result of overestimated Lennard-Jones parameters of backbones. As this type of interactions seems to be non-specific, their effect on fold selection should be lower.

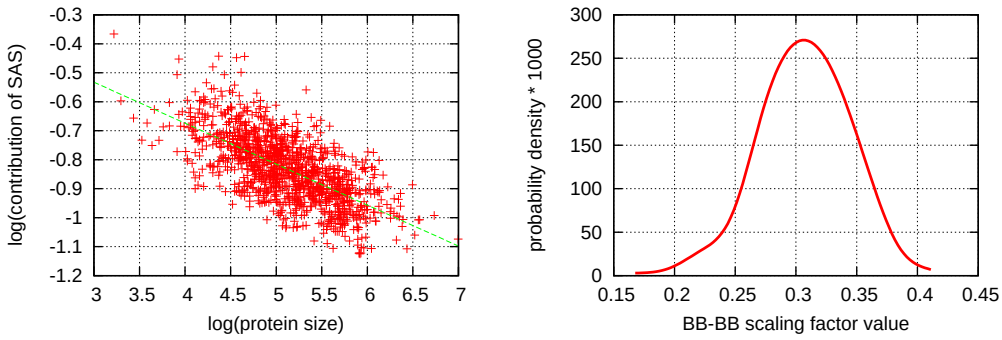
Surprisingly low are scaling factors of charged residues. If solvation was the main force compensating stabilization by interactions of charged sidechains, high compensating values of both 1-body and 2-body contributions would be expected. Three interpretations of the surprisingly low values obtained by model parametrization appear. First, energy of charged interactions might be incorrectly calculated at force field level. Inaccurate distributions of partial charges on charged sidechains can damage quality of close-ranged interaction energies between charged sidechains. However, long-range interactions between charged residues remain unaffected, since Second, hypothesis 2 can be invalid for charged interactions. In each protein, dynamics of resulting in disagreement between native state ensemble average sum of charged-charged interactions and the sum calculated for single structure from PDB.

Third explanation seems to be the most interesting one. In Appendix 2, we show that 92% of energy contribution by attractive interactions between charged sidechains is compensated by repulsive interactions. Low values of scaling factors corresponding to charged residues suggest that very similar compensation might take place in the denatured state, so that positions of charged sidechains are correlated to decrease energy. In other words sort of Madelung constant of native and denatured state is very similar. This effect can be confirmed by interaction energy matrix analyses of protein unfolding simulations.

Table 3.1: Scaling factors obtained using the prioposed stability model. Table is divided into 2 subtabler, 1-body contributions are in the left one and 1-body native state solvation contributions in the right subtable at the bottom. 2-body contributions are stabilizing and can be found in the right subtable (first ten rows). First column of each subtable is name of the scaling factor, second is its value, third is its standard deviation and fourth columns contain contribution of the corresponding scaling factor to protein stability. Destabilizing cintributions are marked by minus sign and in the left subtable. Units are as mentioned above.

scaling factor	value	stdev in %	contrib in %	scaling factor	value	stdev in %	contrib in %
GLY	-4.9	2.4	-4.0	BB-BB	0.31	8.9	16.2
ALA	-6.6	1.8	-5.8	BB-CH	0.10	22	3.4
VAL	-7.3	1.8	-6.1	BB-PO	0.57	7.1	9.2
ILE	-7.4	1.9	-4.8	BB-NP	0.78	7.0	17.6
CYS	-7.8	2.2	-1.3	CH-CH	0.012	86	0.7
LEU	-8.2	1.6	-8.6	CH-PO	0.051	51	0.5
MET	-9.2	1.9	-1.7	CH-NP	0.15	24	0.7
PHE	-10.3	1.7	-4.7	PO-PO	0.75	8.8	1.5
PRO	-16.7	1.8	-8.8	PO-NP	1.00	6.5	5.1
THR	-7.9	1.6	-4.9	NP-NP	0.16	29	1.1
SER	-8.2	1.5	-5.5				
TYR	-13.8	1.3	-5.4	SASNP	0.43	13	6.3
ASN	-14.8	1.0	-7.3	SASPO	6.2	4.3	6.8
TRP	-16.1	1.4	-2.5	SASCH	19.9	1.0	30.9
GLN	-17.0	0.90	-7.4				
LYS	-4.6	2.7	-3.5				
HIS	-5.3	2.9	-1.3				
ASP	-5.9	3.3	-3.9				
GLU	-6.6	3.2	-5.3				
ARG	-12.3	1.3	-7.2				

Figure 3.4: Left - correlation between contribution of native state solvation to stabilization and protein size in log scale is decent with slope 0.14. Right - histogram of scaling factors of backbone-backbone 2-body interactions.



3.6.3 Reliability of the model

Transferability of the model was tested by its parametrization on a set of randomly selected structures from the large set. Scaling factors fitted to 600 structures differed by at most 1 standard deviation (Table 3.1) from the scaling factors presented in Table 3.1. Robustness of the model was tested by introduction of vector of random numbers into matrix **A** (see methods). Vector corresponding to minimum eigenvalue becomes meaningless and the vector corresponding to second lowest eigenvalue becomes the vector of scaling factors. and contribution of the introduced random vector to stability is close to 0.

In Table 3.1, a correlation between value of scaling factor and its deviation can be noticed. This effect can be attributed to reduced importance of small scaling factors in the expression 2.4. The deviation can be artificially reduced by multiplication of all the values in corresponding column of matrix **A** (see method section 2.3) by arbitrary real number $0 < r < 1$, which causes increase of scaling factor by a factor of $1/r$. Such treatment keeps values of all scaling factors almost unaffected, while deviation of other scaling factors is increased.

The model is surprisingly well balanced. While average sum of stabilizing terms is 1700 kcal/mol, average absolute value of protein stability is 30 kcal/mol, which means decent 97.3% compensation of stabilizing and destabilizing terms. The number of parameters can be optically reduced by grouping 1-body contributions with similar values without notable change in model performance. Therefore, we can say that such performance can be reached by our model using about 20 empirical parameters.

3.6.4 Applications

The proposed model involves compensation by denatured state, so it is not directly suitable for construction of energy functions for structure prediction. However, the parametrization procedure can be applied to database of decoys as well. We plan to parametrize the model on experimental stability change amino acid substitution. The model also simplifies identification of amino acids mutation of

which might increase stability, since residues can be ranked according to their stabilization effect from most destabilizing to most stabilizing ones.

The model also attempts to solve the question of stabilization of proteins by particular types of interactions. Understanding forces contributing to stability is crucial for systematic improvement of energy functions. The presented model is not purely empirical; particular contributions have physical meaning which can be further studied.

3.7 Dynamic properties of interaction energy sums

Applicability of hypothesis 2 is interesting for all structural studies. We have undertaken a short study to examine dynamical behavior of interaction energy sums, since we assume validity of hypothesis 2 in most of the presented work. Description of native state by simulation is more realistic. Moreover, it might be useful in prediction of entropy and average IEs from single structure In Table 3.2, interaction energy sums calculated for a single X-ray structure are compared to average sums obtained from IEM analysis of snapshots from simulation (see methods section 2.2.6). Differences between static and dynamic description are small in case of interaction energy types with high scaling factors in Table 3.1.

Table 3.2: Comparison of dynamic and static interaction energy matrix description of protein 3QPA. First column contains interaction energy types, second column contains sum of interaction energies of that type calculated using a single structure, third contains average sum of theinteraction energies thourgout simulation and fourth gives standard deviation of value in the third column. All values are in kcal/mol.

IE type	X-ray	dynamics	stdev
BBBB	-668	-719	12
BBCH	-957	-586	20
BBPO	-351	-293	14
BBNP	-358	-349	7.2
CHCH	-1813	-1450	58
CHPO	-336	-212	15
CHNP	-82.9	-77.5	4.8
POPO	-42.2	-43.4	5.2
PONP	-93.7	-78.7	3.3
NPNP	-99.1	-96.3	3.3

4. Conclusion

In the presented thesis, connection between pairwise interaction energies and their free energy contribution to stability was studied.

Aims listed in section 1.6 were addressed as follows

1. Solvent effect and compensation of interaction energies by persistence in denatured state is modeled by fragmentation of proteins and classification of fragments based on their multipolar character. Instead of using an empirical dielectric constant, we propose comparison of interaction energies with the found characteristic values or linear scaling of interaction energies using the proposed empirical aprameters.

2. In Paper 1 and Paper 2, distributions of residue interaction energies and interaction energies (respcetively) are studied. Interaction energy histograms has been found to be composed of Gaussian-shaped peaks corresponding to particular interaction patterns. Rationalization of this distribution was given in Paper 2.

3. In Paper 2, we present a method of constructing matrices of productive contacts from native state geometry. We hypothesize that such contacts are additive quantities. We have not succeed in exploiting the sequence information contained in contact matrices yet.

4. We have proposed a model integrating all interaction energies into protein stability. The model is well balanced and its further improvement is possible.

5. In Paper 1, we have found no correlation between secondary structures and sidechain interaction energies. The correlation was found only between secondary structure content and interactions between backbones, which is not surprising. Secondary structure propensities could not be correlated with positions of peaks in BB-BB RIE distributions.

There is a lot of space for future studies. First, we want to launch a web application for calculation of interaction energy matrices and construction of our contact matrices. We hope that results of our work will be useful for structural biologists in finding structural and functional features in protein structures. Second, we want to study entropic effect of contact separation in sequence. Third, we plan to use the energy balance in construction of energy functions for prediction of stability change upon amino acid replacement. Last, but not least, we want to modify the methodology to avoid problems arising from difference between native state ensemble and single X-ray structure from database.

Bibliography

[1] RM FORBES, AR COOPER, and HH MITCHELL. THE COMPOSITION OF THE ADULT HUMAN BODY AS DETERMINED BY CHEMICAL ANALYSIS. *JOURNAL OF BIOLOGICAL CHEMISTRY*, 203(1):359–366, 1953.

[2] Gong Zhang and Zoya Ignatova. Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *CURRENT OPINION IN STRUCTURAL BIOLOGY*, 21(1):25–31, FEB 2011.

[3] JJ Ward, JS Sodhi, LJ McGuffin, BF Buxton, and DT Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *JOURNAL OF MOLECULAR BIOLOGY*, 337(3):635–645, MAR 26 2004.

[4] AL Fink. Natively unfolded proteins. *CURRENT OPINION IN STRUCTURAL BIOLOGY*, 15(1):35–41, FEB 2005.

[5] Vladimir N. Uversky and A. Keith Dunker. Understanding protein nonfolding. *BIOCHIMICA ET BIOPHYSICA ACTA-PROTEINS AND PROTEOMICS*, 1804(6):1231–1264, JUN 2010.

[6] Huari-Xiang Zhou. Protein folding in confined and crowded environments. *ARCHIVES OF BIOCHEMISTRY AND BIOPHYSICS*, 469(1):76–82, JAN 1 2008.

[7] Adrian H. Elcock. Models of macromolecular crowding effects and the need for quantitative comparisons with experiment. *CURRENT OPINION IN STRUCTURAL BIOLOGY*, 20(2):196–206, APR 2010.

[8] C ANFINSEN and E HABER. STUDIES ON REDUCTION AND REFORMATION OF PROTEIN DISULFIDE BONDS. *JOURNAL OF BIOLOGICAL CHEMISTRY*, 236(5):1361–&, 1961.

[9] S Govindarajan and RA Goldstein. On the thermodynamic hypothesis of protein folding. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 95(10):5545–5549, MAY 12 1998.

[10] Arieh Ben-Naim. Pitfalls in Anfinsen's thermodynamic hypothesis. *CHEMICAL PHYSICS LETTERS*, 511(1-3):126–128, JUL 26 2011.

[11] W PFEIL and PL PRIVALOV. THERMODYNAMIC INVESTIGATIONS OF PROTEINS .1. STANDARD FUNCTIONS FOR PROTEINS WITH LYSOZYME AS AN EXAMPLE. *BIOPHYSICAL CHEMISTRY*, 4(1):23–32, 1976.

[12] AD Robertson and KP Murphy. Protein structure and the energetics of protein stability. *CHEMICAL REVIEWS*, 97(5):1251–1267, JUL-AUG 1997.

[13] GI Makhatadze and PL Privalov. Energetics of protein structure. In *ADVANCES IN PROTEIN CHEMISTRY, VOL 47*, volume 47 of *ADVANCES IN PROTEIN CHEMISTRY*, pages 307–425. ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495, 1995.

[14] A Michnik. Thermal stability of bovine serum albumin DSC study. *JOURNAL OF THERMAL ANALYSIS AND CALORIMETRY*, 71(2):509–519, 2003.

[15] PL PRIVALOV and KHECHINA.NN. THERMODYNAMIC APPROACH TO PROBLEM OF STABILIZATION OF GLOBULAR PROTEIN STRUCTURE - CALORIMETRIC STUDY. *JOURNAL OF MOLECULAR BIOLOGY*, 86(3):665–684, 1974.

[16] H Kaya and HS Chan. Polymer principles of protein calorimetric two-state cooperativity. *PROTEINS-STRUCTURE FUNCTION AND GENETICS*, 40(4):637–661, SEP 1 2000.

[17] PL Privalov. *ADVANCES IN PROTEIN CHEMISTRY*, 33:167–241, 1979.

[18] L Liu, C Yang, and QX Guo. A study on the enthalpy-entropy compensation in protein unfolding. *BIOPHYSICAL CHEMISTRY*, 84(3):239–251, MAY 15 2000.

[19] S Kumar and R Nussinov. How do thermophilic proteins deal with heat? *CELLULAR AND MOLECULAR LIFE SCIENCES*, 58(9):1216–1233, AUG 2001.

[20] JD DUNITZ. WIN SOME, LOSE SOME - ENTHALPY-ENTROPY COMPENSATION IN WEAK INTERMOLECULAR INTERACTIONS. *CHEMISTRY & BIOLOGY*, 2(11):709–712, NOV 1995.

[21] DM Ford. Enthalpy-entropy compensation is not a general feature of weak association. *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 127(46):16167–16170, NOV 23 2005.

[22] JC KENDREW. ARCHITECTURE OF A PROTEIN MOLECULE. *NATURE*, 182(4638):764–767, 1958.

[23] GJ LESSER and GD ROSE. HYDROPHOBICITY OF AMINO-ACID SUBGROUPS IN PROTEINS. *PROTEINS-STRUCTURE FUNCTION AND GENETICS*, 8(1):6–13, 1990.

[24] AE Mirsky and L Pauling. On the structure of native, denatured, and coagulated proteins. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 22:439–447, 1936.

[25] CA Orengo, AD Michie, S Jones, DT Jones, MB Swindells, and JM Thornton. CATH - a hierarchic classification of protein domain structures. *STRUCTURE*, 5(8):1093–1108, AUG 15 1997.

[26] AG MURZIN, SE BRENNER, T HUBBARD, and C CHOTHIA. SCOP - A STRUCTURAL CLASSIFICATION OF PROTEINS DATABASE FOR THE INVESTIGATION OF SEQUENCES AND STRUCTURES. *JOURNAL OF MOLECULAR BIOLOGY*, 247(4):536–540, APR 7 1995.

[27] A Szilagyi and P Zavodszky. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *STRUCTURE*, 8(5):493–504, MAY 15 2000.

[28] Anna V. Glyakina, Sergiy O. Garbuzynskiy, Michail Yu. Lobanov, and Oxana V. Galzitskaya. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *BIOINFORMATICS*, 23(17):2231–2238, SEP 1 2007.

[29] N Kannan and S Vishveshwara. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *PROTEIN ENGINEERING*, 13(11):753–761, NOV 2000.

[30] D NERI, M BILLETER, G WIDER, and K WUTHRICH. NMR DETERMINATION OF RESIDUAL STRUCTURE IN A UREA-DENATURED PROTEIN, THE 434-REPRESSOR. *SCIENCE*, 257(5076):1559–1563, SEP 11 1992.

[31] H Schwalbe, KM Fiebig, M Buck, JA Jones, SB Grimshaw, A Spencer, SJ Glaser, LJ Smith, and CM Dobson. Structural and dynamical properties of a denatured protein. Heteronuclear 3D NMR experiments and theoretical simulations of lysozyme in 8 M urea. *BIOCHEMISTRY*, 36(29):8977–8991, JUL 22 1997.

[32] HBR COLE, SW SPARKS, and DA TORCHIA. COMPARISON OF THE SOLUTION AND CRYSTAL-STRUCTURES OF STAPHYLOCOCCAL NUCLEASE WITH C-13 AND N-15 CHEMICAL-SHIFTS USED AS STRUCTURAL FINGERPRINTS. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 85(17):6362–6365, SEP 1988.

[33] IK MCDONALD and JM THORNTON. SATISFYING HYDROGEN-BONDING POTENTIAL IN PROTEINS. *JOURNAL OF MOLECULAR BIOLOGY*, 238(5):777–793, MAY 20 1994.

[34] VA Higman, J Boyd, LJ Smith, and C Redfield. Asparagine and glutamine side-chain conformation in solution and crystal: A comparison for hen egg-white lysozyme using residual dipolar couplings. *JOURNAL OF BIOMOLECULAR NMR*, 30(3):327–346, NOV 2004.

[35] CN PACE and TE CREIGHTON. THE DISULFIDE FOLDING PATHWAY OF RIBONUCLEASE-T1. *JOURNAL OF MOLECULAR BIOLOGY*, 188(3):477–486, APR 5 1986.

[36] JK MYERS, CN PACE, and JM SCHOLTZ. DENATURANT M-VALUES AND HEAT-CAPACITY CHANGES - RELATION TO CHANGES IN ACCESSIBLE SURFACE-AREAS OF PROTEIN UNFOLDING. *PROTEIN SCIENCE*, 4(10):2138–2148, OCT 1995.

[37] Qian Wang, Alexander Christiansen, Antonios Samiotakis, Pernilla Wittung-Stafshede, and Margaret S. Cheung. Comparison of chemical and thermal protein denaturation by combination of computational and experimental approaches. II. *JOURNAL OF CHEMICAL PHYSICS*, 135(17), NOV 7 2011.

[38] Jeremy L. England and Gilad Haran. Role of Solvation Effects in Protein Denaturation: From Thermodynamics to Single Molecules and Back. In Leone, SR and Cremer, PS and Groves, JT and Johnson, MA, editor, *ANNUAL REVIEW OF PHYSICAL CHEMISTRY, VOL 62*, volume 62 of *Annual Review of Physical Chemistry*, pages 257–277. ANNUAL REVIEWS, 4139 EL CAMINO WAY, PO BOX 10139, PALO ALTO, CA 94303-0897 USA, 2011.

[39] LN Lin, JF Brandts, JM Brandts, and V Plotnikov. Determination of the volumetric properties of proteins and other solutes using pressure perturbation calorimetry. *ANALYTICAL BIOCHEMISTRY*, 302(1):144–160, MAR 1 2002.

[40] A. Cooper, D. Cameron, J. Jakus, and G. W. Pettigrew. Pressure perturbation calorimetry, heat capacity and the role of water in protein stability and interactions. *BIOCHEMICAL SOCIETY TRANSACTIONS*, 35(Part 6):1547–1550, DEC 2007.

[41] Alekos D. Tsamaloukas, Neena K. Pyzocha, and George I. Makhatadze. Pressure Perturbation Calorimetry of Unfolded Proteins. *JOURNAL OF PHYSICAL CHEMISTRY B*, 114(49):16166–16170, DEC 16 2010.

[42] RL BALDWIN. THE NATURE OF PROTEIN-FOLDING PATHWAYS - THE CLASSICAL VERSUS THE NEW VIEW. *JOURNAL OF BIOMOLECULAR NMR*, 5(2):103–109, FEB 1995.

[43] Z Ignatova and LM Gierasch. Monitoring protein stability and aggregation in vivo by real-time fluorescent labeling. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 101(2):523–528, JAN 13 2004.

[44] AR Fersht. Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 97(4):1525–1529, FEB 15 2000.

[45] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The Protein Data Bank. *NUCLEIC ACIDS RESEARCH*, 28(1):235–242, JAN 1 2000.

[46] MM Gromiha, J An, H Kono, M Oobatake, H Uedaira, and A Sarai. ProTherm: Thermodynamic database for proteins and mutants. *NUCLEIC ACIDS RESEARCH*, 27(1):286–288, JAN 1 1999.

[47] Vladimir Potapov, Mati Cohen, and Gideon Schreiber. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *PROTEIN ENGINEERING DESIGN & SELECTION*, 22(9):553–560, SEP 2009.

[48] RA Friesner and JR Gunn. Computational studies of protein folding. *ANNUAL REVIEW OF BIOPHYSICS AND BIOMOLECULAR STRUCTURE*, 25:315–342, 1996.

[49] AR Dinner, A Sali, LJ Smith, CM Dobson, and M Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *TRENDS IN BIOCHEMICAL SCIENCES*, 25(7):331–339, JUL 2000.

[50] M Vendruscolo and E Paci. Protein folding: bringing theory and experiment closer together. *CURRENT OPINION IN STRUCTURAL BIOLOGY*, 13(1):82–87, FEB 2003.

[51] LEVINTHAL.C. ARE THERE PATHWAYS FOR PROTEIN FOLDING. *JOURNAL DE CHIMIE PHYSIQUE ET DE PHYSICO-CHIMIE BIOLOGIQUE*, 65(1):44–&, 1968.

[52] M Paterson and T Przytycka. On the complexity of string folding. *DISCRETE APPLIED MATHEMATICS*, 71(1-3):217–230, DEC 5 1996.

[53] Oleg K. Vorov, Dennis R. Livesay, and Donald J. Jacobs. Conformational Entropy of an Ideal Cross-Linking Polymer Chain. *ENTROPY*, 10(3):285–308, SEP 2008.

[54] KA DILL, S BROMBERG, KZ YUE, KM FIEBIG, DP YEE, PD THOMAS, and HS CHAN. PRINCIPLES OF PROTEIN-FOLDING - A PERSPECTIVE FROM SIMPLE EXACT MODELS. *PROTEIN SCIENCE*, 4(4):561–602, APR 1995.

[55] KA DILL and D STIGTER. MODELING PROTEIN STABILITY AS HETEROPOLYMER COLLAPSE. In *ADVANCES IN PROTEIN CHEMISTRY, VOL 46*, volume 46 of *ADVANCES IN PROTEIN CHEMISTRY*, pages 59–104. ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495, 1995.

[56] A Hansen, MH Jensen, K Sneppen, and G Zocchi. Statistical mechanics of warm and cold unfolding in proteins. *EUROPEAN PHYSICAL JOURNAL B*, 6(1):157–161, NOV 1998.

[57] E Shakhnovich. Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *CHEMICAL REVIEWS*, 106(5):1559–1588, MAY 2006.

[58] D. Thirumalai, Edward P. O'Brien, Greg Morrison, and Changbong Hyeon. Theoretical Perspectives on Protein Folding. In Rees, DC and Dill, KA and Williamson, JR, editor, *ANNUAL REVIEW OF BIOPHYSICS, VOL 39*, volume 39 of *Annual Review of Biophysics*, pages 159–183. ANNUAL REVIEWS, 4139 EL CAMINO WAY, PO BOX 10139, PALO ALTO, CA 94303-0897 USA, 2010.

[59] Gregory R. Bowman, Vincent A. Voelz, and Vijay S. Pande. Taming the complexity of protein folding. *CURRENT OPINION IN STRUCTURAL BIOLOGY*, 21(1):4–11, FEB 2011.

[60] KA DILL. THEORY FOR THE FOLDING AND STABILITY OF GLOBULAR-PROTEINS. *BIOCHEMISTRY*, 24(6):1501–1509, 1985.

[61] David E. Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers. Atomic-Level Characterization of the Structural Dynamics of Proteins. *SCIENCE*, 330(6002):341–346, OCT 15 2010.

[62] Kevin E. Riley, Michel Pitonak, Petr Jurecka, and Pavel Hobza. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *CHEMICAL REVIEWS*, 110(9):5023–5063, SEP 2010.

[63] John C. Faver, Mark L. Benson, Xiao He, Benjamin P. Roberts, Bing Wang, Michael S. Marshall, C. David Sherrill, and Kenneth M. Merz, Jr. The Energy Computation Paradox and ab initio Protein Folding. *PLOS ONE*, 6(4), APR 25 2011.

[64] JW Ponder and DA Case. Force fields for protein simulations. In *PROTEIN SIMULATIONS*, volume 66 of *ADVANCES IN PROTEIN CHEMISTRY*, pages 27+. ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495 USA, 2003.

[65] Michal Kolar, Karel Berka, Petr Jurecka, and Pavel Hobza. On the Reliability of the AMBER Force Field and its Empirical Dispersion Contribution for the Description of Noncovalent Complexes. *CHEMPHYSCHEM*, 11(11):2399–2408, AUG 2 2010.

[66] Robert B. Best, Nicolae-Viorel Buchete, and Gerhard Hummer. Are current molecular dynamics force fields too helical? *BIOPHYSICAL JOURNAL*, 95(1):L7–L9, JUL 1 2008.

[67] J Hermans. Hydrogen bonds in molecular mechanics force fields. In *PEPTIDE SOLVATION AND H-BONDS*, volume 72 of *ADVANCES IN PROTEIN CHEMISTRY*, pages 105–119. ELSEVIER ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495 USA, 2006.

[68] Allan Chris M. Ferreon and Ashok A. Deniz. Protein folding at single-molecule resolution. *BIOCHIMICA ET BIOPHYSICA ACTA-PROTEINS AND PROTEOMICS*, 1814(8, SI):1021–1029, AUG 2011.

[69] Y Sugita and A Kitao. Improved protein free energy calculation by more accurate treatment of nonbonded energy: Application to chymotrypsin inhibitor 2, V57A. *PROTEINS-STRUCTURE FUNCTION AND GENETICS*, 30(4):388–400, MAR 1 1998.

[70] B Hess. Convergence of sampling in protein simulations. *PHYSICAL REVIEW E*, 65(3, Part 1), MAR 2002.

[71] Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. de Vries. The MARTINI force field: Coarse grained model for biomolecular simulations. *JOURNAL OF PHYSICAL CHEMISTRY B*, 111(27):7812–7824, JUL 12 2007.

[72] N GO and H TAKETOMI. RESPECTIVE ROLES OF SHORT-RANGE AND LONG-RANGE INTERACTIONS IN PROTEIN FOLDING. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 75(2):559–563, 1978.

[73] DJ Jacobs, S Dallakyan, GG Wood, and A Heckathorne. Network rigidity at finite temperature: Relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *PHYSICAL REVIEW E*, 68(6, Part 1), DEC 2003.

[74] DJ Jacobs and S Dallakyan. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *BIOPHYSICAL JOURNAL*, 88(2):903–915, FEB 2005.

[75] A Kolinski and J Skolnick. Reduced models of proteins and their applications. *POLYMER*, 45(2):511–524, JAN 15 2004.

[76] T Head-Gordon and S Brown. Minimalist models for protein folding and design. *CURRENT OPINION IN STRUCTURAL BIOLOGY*, 13(2):160–167, APR 2003.

[77] P Pokarowski, A Kolinski, and J Skolnick. A minimal physically realistic protein-like lattice model: Designing an energy landscape that ensures all-or-none folding to a unique native state. *BIOPHYSICAL JOURNAL*, 84(3):1518–1526, MAR 2003.

[78] F Ding and NV Dokholyan. Simple but predictive protein models. *TRENDS IN BIOTECHNOLOGY*, 23(9):450–455, SEP 2005.

[79] NV Dokholyan. Studies of folding and misfolding using simplified models. *CURRENT OPINION IN STRUCTURAL BIOLOGY*, 16(1):79–85, FEB 2006.

[80] PG Wolynes. Energy landscapes and solved protein-folding problems. *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY OF*

LONDON SERIES A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES, 363(1827):453–464, FEB 15 2005.

[81] O Collet. Four-states phase diagram of proteins. *EUROPHYSICS LETTERS*, 72(2):301–307, OCT 2005.

[82] Philip Ball. Water as an active constituent in cell biology. *CHEMICAL REVIEWS*, 108(1):74–108, JAN 2008.

[83] PL PRIVALOV and SJ GILL. STABILITY OF PROTEIN-STRUCTURE AND HYDROPHOBIC INTERACTION. *ADVANCES IN PROTEIN CHEMISTRY*, 39:191–234, 1988.

[84] W. Kauzmann. Some Factors in the Interpretation of Protein Denaturation. volume 14 of *Advances in Protein Chemistry*, pages 1 – 63. 1959.

[85] J.H. Hildebrand. Is there a "Hydrophobic Effect"? *Proceedings of the National Academy of Sciences of the United States of America*, 76(1):194–194, January 1979.

[86] T.V. Chalikian. Structural Thermodynamics of Hydration. *The Journal of Physical Chemistry B*, 105(50):12566–12578, December 2001.

[87] Bhimalapuram P. Widom, B. and K. Koga. The hydrophobic effect. *Physical Chemistry Chemical Physics*, 5(15):3085, 2003.

[88] Christopher J. Fennell and Ken A. Dill. Physical Modeling of Aqueous Solvation. *JOURNAL OF STATISTICAL PHYSICS*, 145(2, SI):209–226, OCT 2011.

[89] F Avbelj and RL Baldwin. Limited validity of group additivity for the folding energetics of the peptide group. *PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS*, 63(2):283–289, MAY 1 2006.

[90] Arieh Warshel, Pankaz K. Sharma, Mitsunori Kato, and William W. Parson. Modeling electrostatic effects in proteins. *BIOCHIMICA ET BIOPHYSICA ACTA-PROTEINS AND PROTEOMICS*, 1764(11):1647–1676, NOV 2006.

[91] A WARSHEL and M LEVITT. THEORETICAL STUDIES OF ENZYMIC REACTIONS - DIELECTRIC, ELECTROSTATIC AND STERIC STABILIZATION OF CARBONIUM-ION IN REACTION OF LYSOZYME. *JOURNAL OF MOLECULAR BIOLOGY*, 103(2):227–249, 1976.

[92] M Schaefer and M Karplus. A comprehensive analytical treatment of continuum electrostatics. *JOURNAL OF PHYSICAL CHEMISTRY*, 100(5):1578–1599, FEB 1 1996.

[93] A Papazyan and A Warshel. Continuum and dipole-lattice models of solvation. *JOURNAL OF PHYSICAL CHEMISTRY B*, 101(51):11254–11264, DEC 18 1997.

[94] RL JERNIGAN and PJ FLORY. DISTRIBUTION FUNCTIONS FOR CHAIN MOLECULES. *JOURNAL OF CHEMICAL PHYSICS*, 50(10):4185–&, 1969.

[95] D Shortle. The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB JOURNAL*, 10(1):27–34, JAN 1996.

[96] TR SOSNICK and J TREWHELLA. DENATURED STATES OF RIBONUCLEASE-A HAVE COMPACT DIMENSIONS AND RESIDUAL SECONDARY STRUCTURE. *BIOCHEMISTRY*, 31(35):8329–8335, SEP 8 1992.

[97] M KATAOKA, Y HAGIHARA, K MIHARA, and Y GOTO. MOLTEN GLOBULE OF CYTOCHROME-C STUDIED BY SMALL-ANGLE X-RAY-SCATTERING. *JOURNAL OF MOLECULAR BIOLOGY*, 229(3):591–596, FEB 5 1993.

[98] Y Wang and D Shortle. The equilibrium folding pathway of staphylococcal nuclease: Identification of the most stable chain-chain interactions by NMR and CD spectroscopy. *BIOCHEMISTRY*, 34(49):15895–15905, DEC 12 1995.

[99] EE LATTMAN, KM FIEBIG, and KA DILL. MODELING COMPACT DENATURED STATES OF PROTEINS. *BIOCHEMISTRY*, 33(20):6158–6166, MAY 24 1994.

[100] Patrick Weinkam, Ekaterina V. Pletneva, Harry B. Gray, Jay R. Winkler, and Peter G. Wolynes. Electrostatic effects on funneled landscapes and structural diversity in denatured protein ensembles. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 106(6):1796–1801, FEB 10 2009.

[101] C. Nick Pace. Energetics of protein hydrogen bonds. *NATURE STRUCTURAL & MOLECULAR BIOLOGY*, 16(7):681–682, JUL 2009.

[102] C TANFORD. CONTRIBUTION OF HYDROPHOBIC INTERACTIONS TO STABILITY OF GLOBULAR CONFORMATION OF PROTEINS. *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 84(22):4240–&, 1962.

[103] C. Nick Pace, Hailong Fu, Katrina Lee Fryar, John Landua, Saul R. Trevino, Bret A. Shirley, Marsha McNutt Hendricks, Satoshi Iimura, Ketan Gajiwala, J. Martin Scholtz, and Gerald R. Grimsley. Contribution of Hydrophobic Interactions to Protein Stability. *JOURNAL OF MOLECULAR BIOLOGY*, 408(3):514–528, MAY 6 2011.

[104] GS Ratnaparkhi and R Varadarajan. Thermodynamic and structural studies of cavity formation in proteins suggest that loss of packing interactions rather than the hydrophobic effect dominates the observed energetics. *BIOCHEMISTRY*, 39(40):12365–12374, OCT 10 2000.

[105] JM Chen and WE Stites. Packing is a key selection factor in the evolution of protein hydrophobic cores. *BIOCHEMISTRY*, 40(50):15280–15289, DEC 18 2001.

[106] VV Loladze, DN Ermolenko, and GI Makhatadze. Thermodynamic consequences of burial of polar and non-polar amino acid residues in the protein interior. *JOURNAL OF MOLECULAR BIOLOGY*, 320(2):343–357, JUL 5 2002.

[107] J Vondrasek, L Bendova, V Klusak, and P Hobza. 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*, 127(8):2615–2619, MAR 2 2005.

[108] AE MARK and WF VANGUNSTEREN. DECOMPOSITION OF THE FREE-ENERGY OF A SYSTEM IN TERMS OF SPECIFIC INTERACTIONS - IMPLICATIONS FOR THEORETICAL AND EXPERIMENTAL STUDIES. *JOURNAL OF MOLECULAR BIOLOGY*, 240(2):167–176, JUL 8 1994.

[109] GP BRADY and KA SHARP. DECOMPOSITION OF INTERACTION FREE-ENERGIES IN PROTEINS AND OTHER COMPLEX-SYSTEMS. *JOURNAL OF MOLECULAR BIOLOGY*, 254(1):77–85, NOV 17 1995.

[110] KA Dill. Additivity principles in biochemistry. *JOURNAL OF BIOLOGICAL CHEMISTRY*, 272(2):701–704, JAN 10 1997.

[111] HY Zhou and YQ Zhou. Quantifying the effect of burial of amino acid residues on protein stability. *PROTEINS-STRUCTURE FUNCTION AND GENETICS*, 54(2):315–322, FEB 1 2004.

[112] Kingshuk Ghosh and Ken A. Dill. Computing protein stabilities from their chain lengths. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 106(26):10649–10654, JUN 30 2009.

[113] Andrei Y. Istomin, M. Michael Gromiha, Oleg K. Vorov, Donald J. Jacobs, and Dennis R. Livesay. New insight into long-range nonadditivity within protein double-mutant cycles. *PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS*, 70(3):915–924, FEB 15 2008.

[114] PK PONNUSWAMY and MM GROMIHA. ON THE CONFORMATIONAL STABILITY OF FOLDED PROTEINS. *JOURNAL OF THEORETICAL BIOLOGY*, 166(1):63–74, JAN 7 1994.

[115] D EISENBERG, M WESSON, and M YAMASHITA. INTERPRETATION OF PROTEIN FOLDING AND BINDING WITH ATOMIC SOLVATION PARAMETERS. *CHEMICA SCRIPTA*, 29A:217–221, SEP 1989.

[116] DJ BARLOW and JM THORNTON. ION-PAIRS IN PROTEINS. *JOURNAL OF MOLECULAR BIOLOGY*, 168(4):867–885, 1983.

[117] CN Pace, BA Shirley, M McNutt, and K Gajiwala. Forces contributing to the conformational stability of proteins. *FASEB JOURNAL*, 10(1):75–83, JAN 1996.

[118] T Lazaridis, G Archontis, and M Karplus. Enthalpic contribution to protein stability: Insights from atom-based calculations and statistical mechanics. In *ADVANCES IN PROTEIN CHEMISTRY, VOL 47*, volume 47 of *ADVANCES IN PROTEIN CHEMISTRY*, pages 231–306. ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495, 1995.

[119] CA Rohl, CEM Strauss, KMS Misura, and D Baker. Protein structure prediction using rosetta. In *NUMERICAL COMPUTER METHODS, PT D*, volume 383 of *METHODS IN ENZYMOLOGY*, pages 66+. ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495 USA, 2004.

[120] DJ Wales and HA Scheraga. Review: Chemistry - Global optimization of clusters, crystals, and biomolecules. *SCIENCE*, 285(5432):1368–1372, AUG 27 1999.

[121] Michael C. Prentiss, Corey Hardin, Michael P. Eastwood, Chenghang Zong, and Peter G. Wolynes. Protein structure prediction: The next generation. *JOURNAL OF CHEMICAL THEORY AND COMPUTATION*, 2(3):705–716, MAY 2006.

[122] Zhe Zhang, Lin Wang, Yang Gao, Jie Zhang, Maxim Zhenirovskyy, and Emil Alexov. Predicting folding free energy changes upon single point mutations. *BIOINFORMATICS*, 28(5):664–671, MAR 1 2012.

[123] Shuangye Yin, Feng Ding, and Nikolay V. Dokholyan. Eris: an automated estimator of protein stability. *NATURE METHODS*, 4(6):466–467, JUN 2007.

[124] J Schymkowitz, J Borg, F Stricher, R Nys, F Rousseau, and L Serrano. The FoldX web server: an online force field. *NUCLEIC ACIDS RESEARCH*, 33(2):W382–W388, JUL 1 2005.

[125] JM THORNTON. DISULFIDE BRIDGES IN GLOBULAR-PROTEINS. *JOURNAL OF MOLECULAR BIOLOGY*, 151(2):261–287, 1981.

[126] Milan Oncak, Karel Berka, and Petr Slavicek. Novel Covalent Bond in Proteins: Calculations on Model Systems Question the Bond Stability. *CHEMPHYSCHEM*, 12(17):3449–3457, DEC 9 2011.

[127] PP Ewald. The calculation of optical and electrostatic grid potential. *ANNALEN DER PHYSIK*, 64(3):253–287, FEB 1921.

[128] U ESSMANN, L PERERA, ML BERKOWITZ, T DARDEN, H LEE, and LG PEDERSEN. A SMOOTH PARTICLE MESH EWALD METHOD. *JOURNAL OF CHEMICAL PHYSICS*, 103(19):8577–8593, NOV 15 1995.

[129] G Hummer, LR Pratt, and AE Garcia. Free energy of ionic hydration. *JOURNAL OF PHYSICAL CHEMISTRY*, 100(4):1206–1215, JAN 25 1996.

[130] YY Sham and A Warshel. The surface constraint all atom model provides size independent results in calculations of hydration free energies. *JOURNAL OF CHEMICAL PHYSICS*, 109(18):7940–7944, NOV 8 1998.

[131] MA Kastenholz and PH Hunenberger. Computation of methodology-independent ionic solvation free energies from molecular simulations. I. The electrostatic potential in molecular liquids. *JOURNAL OF CHEMICAL PHYSICS*, 124(12), MAR 28 2006.

[132] S RYBAK, B JEZIORSKI, and K SZALEWICZ. MANY-BODY SYMMETRY-ADAPTED PERTURBATION-THEORY OF INTERMOLECULAR INTERACTIONS - H2O AND HF DIMERS. *JOURNAL OF CHEMICAL PHYSICS*, 95(9):6576–6601, NOV 1 1991.

[133] AJ Misquitta and K Szalewicz. Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. *JOURNAL OF CHEMICAL PHYSICS*, 122(21), JUN 1 2005.

[134] Elangannan Arunan, Gautam R. Desiraju, Roger A. Klein, Joanna Sadlej, Steve Scheiner, Ibon Alkorta, David C. Clary, Robert H. Crabtree, Joseph J. Dannenberg, Pavel Hobza, Henrik G. Kjaergaard, Anthony C. Legon, Benedetta Mennucci, and David J. Nesbitt. Definition of the hydrogen bond (IUPAC Recommendations 2011). *PURE AND APPLIED CHEMISTRY*, 83(8):1637–1641, 2011.

[135] Karel Berka, Roman Laskowski, Kevin E. Riley, Pavel Hobza, and Jiri Vondrasek. 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*, 5(4):982–992, APR 2009.

[136] J Kolafa. Time-reversible always stable predictor-corrector method for molecular dynamics of polarizable molecules. *JOURNAL OF COMPUTATIONAL CHEMISTRY*, 25(3):335–342, FEB 2004.

[137] J Kolafa. Gear formalism of the always stable predictor-corrector method for molecular dynamics of polarizable molecules. *JOURNAL OF CHEMICAL PHYSICS*, 122(16), APR 22 2005.

[138] K. G. Tina, R. Bhadra, and N. Srinivasan. PIC: Protein Interactions Calculator. *NUCLEIC ACIDS RESEARCH*, 35(S):W473–W476, JUL 2007.

[139] Chih-Peng Lin, Shao-Wei Huang, Yan-Long Lai, Shih-Chung Yen, Chien-Hua Shih, Chih-Hao Lu, Cuen-Chao Huang, and Jenn-Kang Hwang. Deriving protein dynamical properties from weighted protein contact number. *PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS*, 72(3):929–935, AUG 15 2008.

[140] M Vendruscolo, NV Dokholyan, E Paci, and M Karplus. Small-world view of the amino acids that play a key role in protein folding. *PHYSICAL REVIEW E*, 65(6, Part 1), JUN 2002.

[141] KV Brinda and S Vishveshwara. A network representation of protein structures: Implications for protein stability. *BIOPHYSICAL JOURNAL*, 89(6):4159–4170, DEC 2005.

[142] KW Plaxco, KT Simons, and D Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *JOURNAL OF MOLECULAR BIOLOGY*, 277(4):985–994, APR 10 1998.

[143] MM Gromiha and S Selvaraj. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *JOURNAL OF MOLECULAR BIOLOGY*, 310(1):27–32, JUN 29 2001.

[144] L HOLM and C SANDER. PROTEIN-STRUCTURE COMPARISON BY ALIGNMENT OF DISTANCE MATRICES. *JOURNAL OF MOLECULAR BIOLOGY*, 233(1):123–138, SEP 5 1993.

[145] Inken Wohlers, Francisco S. Domingues, and Gunnar W. Klau. Towards optimal alignment of protein structure distance matrices. *BIOINFORMATICS*, 26(18):2273–2280, SEP 2010.

[146] AD Michie, CA Orengo, and JM Thornton. Analysis of domain structural class using an automated class assignment protocol. *JOURNAL OF MOLECULAR BIOLOGY*, 262(2):168–185, SEP 20 1996.

[147] D Xu, CJ Tsai, and R Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *PROTEIN ENGINEERING*, 10(9):999–1012, SEP 1997.

[148] Jose M. Duarte, Rajagopal Sathyapriya, Henning Stehr, Ioannis Filippis, and Michael Lappe. Optimal contact definition for reconstruction of Contact Maps. *BMC BIOINFORMATICS*, 11, MAY 27 2010.

[149] S Kundu, JS Melton, DC Sorensen, and GN Phillips. Dynamics of proteins in crystals: Comparison of experiment with simple models. *BIOPHYSICAL JOURNAL*, 83(2):723–732, AUG 2002.

[150] Kussell E. Vendruscolo, M. and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997.

[151] Subramanian B. Kanter I. Domany E. Vendruscolo, M. and J. Lebowitz. Statistical properties of contact maps. *Physical Review E*, 59(1):977–984, January 1999.

[152] M. Vendruscolo, R. Najmanovich, and E. Domany. Protein Folding in Contact Map Space. *Physical Review Letters*, 82(3):656–659, January 1999.

[153] Kanter I. Vendruscolo M. Kabakçioğlu, A. and E. Domany. Statistical properties of contact vectors. *Physical Review E*, 65(4):1–7, March 2002.

[154] J Khatun, SD Khare, and NV Dokholyan. Can contact potentials reliably predict stability of proteins? *JOURNAL OF MOLECULAR BIOLOGY*, 336(5):1223–1238, MAR 5 2004.

[155] M LEVITT, C SANDER, and PS STERN. PROTEIN NORMAL-MODE DYNAMICS - TRYPSIN-INHIBITOR, CRAMBIN, RIBONUCLEASE AND LYSOZYME. *JOURNAL OF MOLECULAR BIOLOGY*, 181(3):423–447, 1985.

[156] MM Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *PHYSICAL REVIEW LETTERS*, 77(9):1905–1908, AUG 26 1996.

[157] Huan-Xiang Zhou and Michael K. Gilson. Theory of Free Energy and Entropy in Noncovalent Binding. *CHEMICAL REVIEWS*, 109(9):4092–4107, SEP 2009.

[158] Ji Guo Su, Xian Jin Xu, Chun Hua Li, Wei Zu Chen, and Cun Xin Wang. Identification of key residues for protein conformational transition using elastic network model. *JOURNAL OF CHEMICAL PHYSICS*, 135(17), NOV 7 2011.

[159] S MIYAZAWA and RL JERNIGAN. ESTIMATION OF EFFECTIVE INTERRESIDUE CONTACT ENERGIES FROM PROTEIN CRYSTAL-STRUCTURES - QUASI-CHEMICAL APPROXIMATION. *MACRO-MOLECULES*, 18(3):534–552, 1985.

[160] S. Miyazawa and R.L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256(3):623–44, March 1996.

[161] Lada Bendova-Biedermannova, Pavel Hobza, and Jiri Vondrasek. Identifying stabilizing key residues in proteins using interresidue interaction energy matrix. *PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS*, 72(1):402–413, JUL 2008.

[162] Karel Berka, Roman A. Laskowski, Pavel Hobza, and Jiri Vondrasek. Energy Matrix of Structurally Important Side-Chain/Side-Chain Interactions in Proteins. *JOURNAL OF CHEMICAL THEORY AND COMPUTATION*, 6(7):2191–2203, JUL 2010.

[163] Dmitri G. Fedorov and Kazuo Kitaura. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *JOURNAL OF PHYSICAL CHEMISTRY A*, 111(30):6904–6914, AUG 2 2007.

[164] Fačkovec, B and Vondrášek,J. *Decomposition of Intramolecular Interactions Between Amino-Acids in Globular Proteins - A Consequence for Structural Classes of Proteins and Methods of Their Classification* . IntechOpen, 2011.

[165] B. Fačkovec. *Potential energy stabilizing a hydrophobic core of protein and its contribution to overall stability.* Bachelor Thesis, 2010.

[166] WL Jorgensen, DS Maxwell, and J TiradoRives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY*, 118(45):11225–11236, NOV 13 1996.

[167] SE Feller and AD MacKerell. An improved empirical potential energy function for molecular simulations of phospholipids. *JOURNAL OF PHYSICAL CHEMISTRY B*, 104(31):7510–7515, AUG 10 2000.

[168] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *JOURNAL OF CHEMICAL PHYSICS*, 126(1), JAN 7 2007.

[169] KA SHARP and SW ENGLANDER. HOW MUCH IS A STABILIZING BOND WORTH. *TRENDS IN BIOCHEMICAL SCIENCES*, 19(12):526–529, DEC 1994.

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
DSC - Differential Scanning Calorimetry
FF - Force Field
HB - Hydrogen Bond
IE - Interaction Energy
IEM - Interaction Energy Matrix
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
SA - Surface Area
SAPT - Symmetry-Adapted Perturbation Theory
SAS - Solvent Accessible Surface
SB - Salt Bridge
SC - Side-Chain
SDoF - Stiff Degree of Freedom
vdW - van der Waals

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
BBBB - backbone-backbone
BBCH - backbone - charged
BBPO - backbone - polar
BBNP - backbone - non-polar
CHCH - charged - charged
CHPO - charged - polar
CHNP - charged - non-polar
POPO - polar - polar
PONP - polar - non-polar
NPNP - non-polar - non-polar