Protein-Protein Interaktionen:

Einfluss des Lösungsmittels und Effekte von Fluorirung

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. Nat.)

vorgelegt

der Fakultät Mathematik und Naturwissenschaften
der Technischen Universität Dresden
von

Master in Physik Sergey Samsonov

geboren am 28.02.1983 in Leningrad, UdSSR

Eingereicht am 15.05.2009

Verteidigt am 16.11.2009

Gutachter: Prof. Daniel Müller (BIOTEC TU Dresden)

Prof. Nikolay Dokholyan (University of NC, USA)

Die Dissertation wurde in der Zeit von August 2006 bis
Mai 2009 im BIOTEC TUD angefertigt.

PROTEIN-PROTEIN INTERACTIONS:

IMPACT OF SOLVENT AND EFFECTS OF FLUORINATION

Sergey A. Samsonov

A dissertation submitted in the fulfillment of
the requirements for the degree of Doctor of Philosophy

Technical University of Dresden

2009

Program Authorized to Offer Degree: Biology

The research in this thesis has been carried out at the Structural Bioinformatics Group headed by Dr. María Teresa Pisabarro at the Biotechnology Center (BIOTEC) of Technical University of Dresden.

The research carried out in this thesis has been supported by the Klaus Tschira Foundation

Technical University of Dresden

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Sergey A. Samsonov

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

_____

Reading Committee:

_____

_____

Date: _____

Selbständigkeitserklärung und Erklärung zur Annerkennung der Promotionsordnung

Hiermit versichere ich, Sergey A. Samsonov, dass ich die vorliegende Arbeit:

Protein-Protein Interaktionen: Einfluss des Lösungsmittels und Effekte von Fluorirung

ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die Dissertation wurde von Dr. M. Teresa Pisabarro, BIOTEC TUD, Structural Bioingormatics betreut und im Zeitraum vom August 2006 bis May 2009 verfasst.

Meine Person betreffend erkläre ich hiermit, dass keine früheren erfolglosen Promotionsverfahren stattgefunden haben.

Ich erkenne die Promotionsordnung der Fakultät für Mathematik und Naturwissenschaften, Technische Universität Dresden an.

Technical University of Dresden

Summary

Protein-protein interactions:

impact of solvent and effects of fluorination

Sergey A. Samsonov

Chair of the Supervisory Committee:
Professor Michael Brand
Faculty of Biology

Proteins have an indispensable role in the cell. They carry out a wide variety of structural, catalytic and signaling functions in all known biological systems. To perform their biological functions, proteins establish interactions with other bioorganic molecules including other proteins. Therefore, protein-protein interactions is one of the central topics in molecular biology. My thesis is devoted to three different topics in the field of protein-protein interactions. The first one focuses on solvent contribution to protein interfaces as it is an important component of protein complexes. The second topic discloses the structural and functional potential of fluorine's unique properties, which are attractive for protein design and engineering not feasible within the scope of canonical amino acids. The last part of this thesis is a study of the impact of charged amino acid residues within the hydrophobic interface of a coiled-coil system, which is one of the well-established model systems for protein-protein interactions studies.

I. The majority of proteins interact *in vivo* in solution, thus studies of solvent impact on protein-protein interactions could be crucial for understanding many processes in the cell. However, though solvent is known to be very important for protein-protein interactions in terms of structure, dynamics and energetics, its effects are often disregarded in computational studies because a detailed solvent description requires complex and computationally demanding approaches. As a consequence, many protein residues, which establish water-mediated interactions, are neither considered in an interface definition. In the previous work carried out in our group the protein interfaces database (SCOWLP) has been developed. This database takes into account interfacial solvent and based on this classifies all interfacial protein residues of the PDB into three classes based on their interacting properties: *dry*

(direct interaction), *dual* (direct and water-mediated interactions), and *wet spots* (residues interacting only through one water molecule). To define an interaction SCOWLP considers a donor–acceptor distance for hydrogen bonds of 3.2 Å, for salt bridges of 4 Å, and for van der Waals contacts the sum of the van der Waals radii of the interacting atoms. In previous studies of the group, statistical analysis of a non-redundant protein structure dataset showed that 40.1% of the interfacial residues participate in water-mediated interactions, and that 14.5% of the total residues in interfaces are wet spots. Moreover, wet spots have been shown to display similar characteristics to residues contacting water molecules in cores or cavities of proteins.

The goals of this part of the thesis were:

1. to characterize the impact of solvent in protein-protein interactions
2. to elucidate possible effects of solvent inclusion into the correlated mutations approach for protein contacts prediction

To study solvent impact on protein interfaces a molecular dynamics (MD) approach has been used. This part of the work is elaborated in section 2.1 of this thesis. We have characterized properties of water-mediated protein interactions at residue and solvent level. For this purpose, an MD analysis of 17 representative complexes from SH3 and immunoglobulin protein families has been performed. We have shown that the interfacial residues interacting through a single water molecule (*wet spots*) are energetically and dynamically very similar to other interfacial residues. At the same time, water molecules mediating protein interactions have been found to be significantly less mobile than surface solvent in terms of residence time. Calculated free energies indicate that these water molecules should significantly affect formation and stability of a protein-protein complex. The results obtained in this part of the work also suggest that water molecules in protein interfaces contribute to the conservation of protein interactions by allowing more sequence variability in the interacting partners, which has important implications for the use of the correlated mutations concept in protein interactions studies. This concept is based on the assumption that interacting protein residues co-evolve, so that a mutation in one of the interacting counterparts is compensated by a mutation in the other. The study presented in section 2.2 has been carried out to prove that an explicit introduction of solvent into the correlated mutations concept indeed yields qualitative improvement of existing approaches. For this, we have used the data on interfacial solvent obtained from the SCOWLP database (the whole PDB) to construct a “wet” similarity matrix. This matrix has been used for prediction of protein contacts together with a well-established “dry” matrix. We have analyzed two datasets containing 50 domains and 10 domain

pairs, and have compared the results obtained by using several combinations of both “dry” and “wet” matrices. We have found that for predictions for both intra- and interdomain contacts the introduction of a combination of a “dry” and a “wet” similarity matrix improves the predictions in comparison to the “dry” one alone. Our analysis opens up the idea that the consideration of water may have an impact on the improvement of the contact predictions obtained by correlated mutations approaches. There are two principally novel aspects in this study in the context of the used correlated mutations methodology :
i) the first introduction of solvent explicitly into the correlated mutations approach; *ii)* the use of the definition of protein-protein interfaces, which is essentially different from many other works in the field because of taking into account physico-chemical properties of amino acids and not being exclusively based on distance cut-offs.

II. The second part of the thesis is focused on properties of fluorinated amino acids in protein environments. In general, non-canonical amino acids with newly designed side-chain functionalities are powerful tools that can be used to improve structural, catalytic, kinetic and thermodynamic properties of peptides and proteins, which otherwise are not feasible within the use of canonical amino acids. In this context fluorinated amino acids have increasingly gained in importance in protein chemistry because of fluorine's unique properties: high electronegativity and a small atomic size. Despite the wide use of fluorine in drug design, properties of fluorine in protein environments have not been yet extensively studied. The aims of this part of the dissertation were:

1. to analyze the basic properties of fluorinated amino acids such as electrostatic and geometric characteristics, hydrogen bonding abilities, hydration properties and conformational preferences (section 3.1)

2. to describe the behavior of fluorinated amino acids in systems emulating protein environments (section 3.2, section 3.3)

First, to characterize fluorinated amino acids side chains we have used fluorinated ethane derivatives as their simplified models and applied a quantum mechanics approach. Properties such as charge distribution, dipole moments, volumes and size of the fluoromethylated groups within the model have been characterized. Hydrogen bonding properties of these groups have been compared with the groups typically presented in natural protein environments. We have shown that hydrogen and fluorine atoms within these fluoromethylated groups are weak hydrogen bond donors and acceptors. Nevertheless they should not be disregarded for applications in protein engineering. Then, we have

implemented four fluorinated L-amino acids for the AMBER force field and characterized their conformational and hydration properties at the MD level. We have found that hydrophobicity of fluorinated side chains grows with the number of fluorine atoms and could be explained in terms of high electronegativity of fluorine atoms and spacial demand of fluorinated side-chains. These data on hydration agrees with the results obtained in the experimental work performed by our collaborators.

We have rationally engineered systems that allow us to study fluorine properties and extract results that could be extrapolated to proteins. For this, we have emulated protein environments by introducing fluorinated amino acids into a parallel coiled-coil and enzyme-ligand chymotrypsin systems. The results on fluorination effect on coiled-coil dimerization and substrate affinities in the chymotrypsin active site obtained by MD, molecular docking and free energy calculations are in strong agreement with experimental data obtained by our collaborators. In particular, we have shown that fluorine content and position of fluorination can considerably change the polarity and steric properties of an amino acid side chain and, thus, can influence the properties that a fluorinated amino acid reveals within a native protein environment.

III. Coiled-coils typically consist of two to five right-handed α-helices that wrap around each other to form a left-handed superhelix. The interface of two α-helices is usually represented by hydrophobic residues. However, the analysis of protein databases revealed that in natural occurring proteins up to 20% of these positions are populated by polar and charged residues. The impact of these residues on stability of coiled-coil system is not clear. MD simulations together with free energy calculations have been utilized to estimate favourable interaction partners for uncommon amino acids within the hydrophobic core of coiled-coils (Chapter 4). Based on these data, the best hits among binding partners for one strand of a coiled-coil bearing a charged amino acid in a central hydrophobic core position have been selected. Computational data have been in agreement with the results obtained by our collaborators, who applied phage display technology and CD spectroscopy. This combination of theoretical and experimental approaches allowed to get a deeper insight into the stability of the coiled-coil system.

To conclude, this thesis widens existing concepts of protein structural biology in three areas of its current importance. We expand on the role of solvent in protein interfaces, which contributes to the knowledge of physico-chemical properties underlying protein-protein interactions. We develop a

deeper insight into the understanding of the fluorine's impact upon its introduction into protein environments, which may assist in exploiting the full potential of fluorine's unique properties for applications in the field of protein engineering and drug design. Finally we investigate the mechanisms underlying coiled-coil system folding. The results presented in the thesis are of definite importance for possible applications (e.g. introduction of solvent explicitly into the scoring function) into protein folding, docking and rational design methods.

The dissertation consists of four chapters:

- Chapter 1 contains an introduction to the topic of protein-protein interactions including basic concepts and an overview of the present state of research in the field.
- Chapter 2 focuses on the studies of the role of solvent in protein interfaces.
- Chapter 3 is devoted to the work on fluorinated amino acids in protein environments.
- Chapter 4 describes the study of coiled-coils folding properties.

The experimental parts presented in Chapters 3 and 4 of this thesis have been performed by our collaborators at FU Berlin.

Sections 2.1, 2.2, 3.1, 3.2 and Chapter 4 have been submitted/published in peer-reviewed international journals. Their organization follows a standard research article structure: Abstract, Introduction, Methodology, Results and discussion, and Conclusions. Section 3.3, though not published yet, is also organized in the same way. The literature references are summed up together at the end of the thesis to avoid redundancy within different chapters.

ACKNOWLEDGEMENTS

First of all I would like to acknowledge my supervisor Dr. Mayte Pisabarro for putting a lot of her time and effort to guide me through the projects as well as for the maintenance of a very friendly and work supporting environment in the group. The independence in research, which she granted together with the constant encouragement, was the very key point for my motivation either in more or less lucky professional and personal moments. Then, I'm grateful to all members of the group in past and present: Andres Palencia, Carsten Baldauf, Jens Laettig, Frank Brand, Jana Sontheimer, Aurelie Tomczak, Ionut David, Anders Thogersen, John Hawkins, Hong Bo Zhu and, especially, Joan Teyra, Gerd Anders and Maciej Paszkowski-Rogacz for their professional and friend's attitude and support I received from them during my PhD. I extend my acknowledgments to Mandy Erlitz and Ralf Gey, without whose help that would be impossible to conduct this work and settle down everything for life in the city of Dresden. I thank also our collaborators in the group of Beate Koksch at FU Berlin: Beate Koksch, Mario Salwiczek, Toni Vagt, Matthias Hakelberg for very interesting common projects that we had and we are going to continue in the future.

I am indebted to my parents, who made a lot for me to be able to start my PhD, helping and supporting me during the whole my life, so this work definitely belongs to them as well. I thank all of my friends in Saint-Petersburg, Germany, Finland, Belarus and Poland for continuous inspiration they contributed to during this time I had been in Dresden doing my PhD.

Last, but not least, I'd like to devote this dissertation to the memory of my grandfather, who passed away several years ago and had an indispensable impact on my life.

"*The scientist does not study nature because it is useful; he studies it because he delights in it.*"

Jules-Henri Poincaré

"*A learned man is an idler who kills time by study.*"

George Bernard Shaw

CHAPTER 1

This chapter introduces the basic concepts of protein-protein interactions and depicts recent achievements and challenges in related studies. When not particularly cited, the chapter is the recompilation of 'Principles of Biochemistry' by Lehninger [1], 'Principles of Biochemistry' by Zubay [2], 'Biochemistry' by Berg et al. [3], 'Principles of Physical Biochemistry' by van Holde et al. [4], 'Computational Chemistry' by Cramer [5], 'Bioinformatics. Sequence and Genome Analysis' by Mount [6], 'Dynamics of proteins and nucleic acids' by McCammon and Harvey [7], AMBER 8.0 manual [8] and Wikipedia, the free multilingual encyclopedia project [9].

1.1 Proteins: a 'sequence-structure-function' dogma

In this section we introduce the basic definitions and concepts of protein chemistry.

1.1.1 Role of proteins in biology of the cell

Proteins are the most versatile macromolecules in living systems and serve crucial functions in essentially all biological processes. Their central place in the cell is reflected in the fact that genetic information is ultimately expressed as protein. Proteins carry out functions, by which they could be arbitrarily classified as:

- Enzymes (possess biocatalytic activities; e.g. lysozyme, ribonuclease)
- Transport proteins (participate in processes, in which specific molecules or ions are transported from one localization to another; e.g. hemoglobin, lipoproteins)
- Nutrient and storage proteins (are used for storage of energy or specific molecules; e.g. ovalbumin, metallothionein)
- Contractile or motile proteins (endow cells and organisms with the ability to contract, to change shape, or to move; e.g. actin, myosin, tubulin)
- Structural proteins (serve to establish filaments, or sheets, to give biological structures strength or protection; e.g. collagen, elastin, keratin)
- Defense proteins (defend cells/organisms against invasion by other species or protect them from injury; e.g. immunoglobulins, toxins).
- Regulatory proteins (regulate cellular or physiological activity; e.g. G-proteins, insulin, MAP-kinases)
- Others (their functions are exotic or not classified; e.g. antifreeze proteins, monellin)

Despite this wide functional variety, all proteins are polymers built up from relatively simple

monomeric subunits: *α-amino-acids*. An α-amino acid consists of a central carbon atom, called the α-carbon, linked to an amino group, a carboxylic acid group, a hydrogen atom, and a distinctive R-group (Figure 1.1.1). The R-group is often referred to as the *side chain*. With four different groups connected to the tetrahedral α-carbon atom, α-amino acids are chiral; the two mirror-image forms are called the L-isomer and the D-isomer.

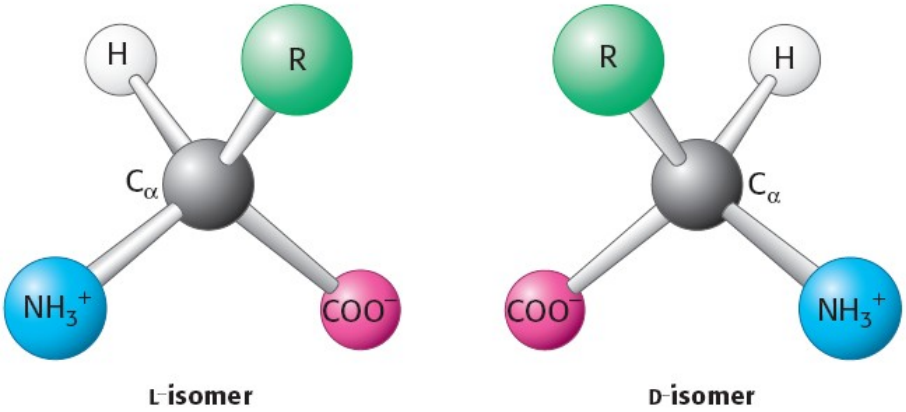

Figure 1.1.1. The L- and D-isomers of amino acids. R refers to the side chain. The L and D isomers are mirror images of each other [2].

Only L-amino acids are constituents of proteins. For almost all amino acids, the L-isomer has S (rather than R) absolute configuration (Figure 1.1.1). Although considerable effort has gone into understanding why amino acids in proteins have this absolute configuration, no satisfactory explanation has been arrived yet. It seems plausible that the selection of L- over D- was arbitrary but, once made, was fixed early in evolutionary history. Amino acids in solution at neutral pH exist predominantly as dipolar ions (also called *zwitterions*). In this dipolar form, the amino group is protonated ($–NH3^+$) and the carboxyl group is deprotonated ($–COO^-$).

Twenty side chains varying in size, shape, charge, hydrogen bonding capacity, hydrophobic character, and chemical reactivity are commonly found in proteins (Table 1.1.1). Indeed, all proteins in all species—bacterial, archaeal, and eukaryotic—are constructed from the same set of 20 amino acids. The remarkable range of functions mediated by proteins results from the diversity and versatility of these 20 building blocks. The side chain functionalities of amino acids include alcohols, thiols, thioethers, carboxylic acids, carboxamides, and a variety of basic groups. Sometimes proteins contain also non-protoigenic groups covalently (e.g. phosphorylation of Ser/Tyr/Thr) and non-covalently (e.g. metal ions coordination).

To quantify the hydrophobicity/hydrophilicity of amino acids, the *hydropathy index* is

corresponded to each of them and represented by the partition coefficient P, measured as the fraction of molecules in the aqueous phase χ_{aq} relative to the fraction in the organic phase χ_{nonaq} at equilibrium:

$$P = -\lg(\chi_{aq}/\chi_{nonaq}) \quad (1.1.1)$$

Table 1.1.1. Standard amino acids and their side chain properties

Amino acid	3-letter code	1-letter code	Side chain polarity	Side chain charge (pH 7)	Hydropathy index (P)
Alanine	Ala	A	nonpolar	neutral	1.8
Arginine	Arg	R	polar	positive	-4.5
Asparagine	Asn	N	polar	neutral	-3.5
Aspartic acid	Asp	D	polar	negative	-3.5
Cysteine	Cys	C	nonpolar	neutral	2.5
Glutamic acid	Glu	E	polar	negative	-3.5
Glutamine	Gln	Q	polar	neutral	-3.5
Glycine	Gly	G	nonpolar	neutral	-0.4
Histidine	His	H	polar	positive	-3.2
Isoleucine	Ile	I	nonpolar	neutral	4.5
Leucine	Leu	L	nonpolar	neutral	3.8
Lysine	Lys	K	polar	positive	-3.9
Methionine	Met	M	nonpolar	neutral	1.9
Phenylalanine	Phe	F	nonpolar	neutral	2.8
Proline	Pro	P	nonpolar	neutral	-1.6
Serine	Ser	S	polar	neutral	-0.8
Threonine	Thr	T	polar	neutral	-0.7
Tryptophan	Trp	W	nonpolar	neutral	-0.9
Tyrosine	Tyr	Y	polar	neutral	-1.3
Valine	Val	V	nonpolar	neutral	4.2

Two amino acid molecules can be covalently joined through a substituted amide linkage, termed a *peptide bond*, to yield a *dipeptide*. Such a linkage is formed by removal of a water moleculed from the α-carboxyl group of one amino acid and the α-amino group of another (Figure 1.1.2). Peptide-bond formation is an example of a condensation reaction. Three amino acids can be joined by two peptide bonds to form a tripeptide etc. When a few amino acids are joined in this fashion, the structure is called an *oligopeptide*. When amino acids are joined, the product is called a *polypeptide*. Proteins may have thousands of amino acid units (*residues*).

Because of the partially double nature of peptide bond it is planar and could represent either the *cis*- or *trans*- isomers. In the unfolded state of proteins, the peptide groups are free to isomerize and adopt both isomers; however, in the folded state, only a single isomer is adopted at each position. The *trans* form is preferred overwhelmingly in most peptide bonds (roughly 1000:1 ratio in *trans*:*cis* populations).

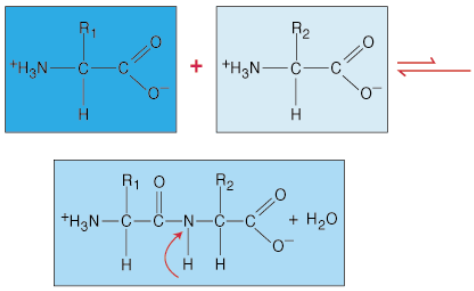

Figure 1.1.2. Dipeptide formation from amino acids with the side chains R_1 and R_2.

1.1.2 Sequence

The *sequence of a protein* is the order of the amino acids that are covalently linked by peptide bonds. Taking into account that there are 20 different standard amino acid monomers, the probability of finding any of these 20 amino acids at each position of the sequence is 1/20. Thus, probability of finding two polypeptides of N amino acids length to be identical by sequence by chance is $1/20^N$. That means that even for a small protein (N=100) there are ~10^{130} different possible amino acid sequences, which could have strikingly different unique structural and functional properties. Millions of protein sequences from different organism sources are nowadays available from the Internet accessible databases (e.g. UniProt Consortium [10], National Center for Biotechnologica Information [11]).

1.1.3 Structure

Despite planarity of the peptide bond the *main chain of protein* (the atoms not belonging to side chain) has two conformational degrees of freedom, which could be corresponded to *dihedral angles* φ and ψ, illustrated at the Figure 1.1.3.

Depending on amino acids side chains physico-chemical properties φ and ψ angles for each residue have different allowed conformational space. The ability of proteins to fold into well-defined structures is remarkable thermodynamically. Consider the equilibrium between an unfolded polymer that exists as a *random coil*—that is, as a mixture of many possible conformations—and the folded form that adopts unique conformation. The favorable entropy associated with the large number of

conformations in the unfolded form opposes folding and must be overcome by interactions favoring the folded form. Thus, highly flexible polymers with a large number of possible conformations do not fold into unique structures. The rigidity of the peptide unit and the restricted set of allowed and φ and ψ angles limits the number of structures accessible to the unfolded form sufficiently to allow protein folding to occur.

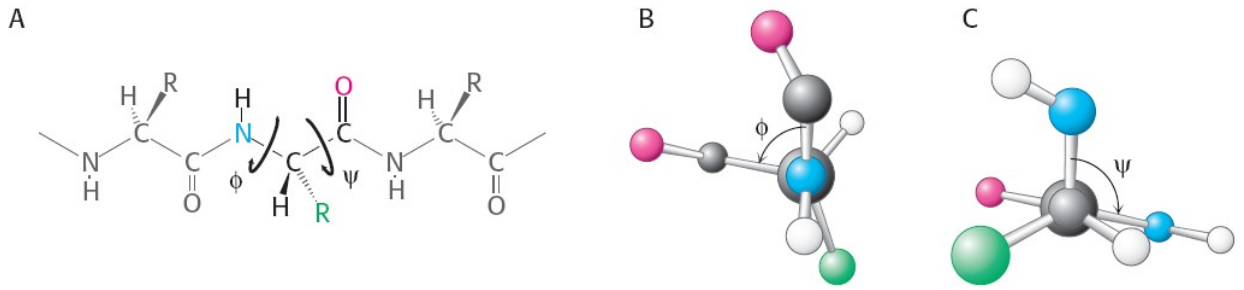

Figure 1.1.3. Rotation about bonds in a polypeptide. The structure of each amino acid in a polypeptide can be adjusted by rotation about two single bonds. A) φ is the angle of rotation about the bond between the nitrogen and the α-carbon atoms, whereas ψ is the angle of rotation about the bond between the α-carbon and the carbonyl carbon atoms. B) A view down the bond between the nitrogen and the α-carbon atoms, showing how φ is measured. C) A view down the bond between the α-carbon and the carbonyl carbon atoms, showing how ψ is measured [2].

Conceptually, protein structure can be considered at four levels (Figure 1.1.4):

- *Primary structure* refers to an amino acids sequence in the protein.
- *Secondary structure* refers to regular, recurring arrangements in space of adjacent amino acid residues in a polypeptide chain. The types of secondary structure relate to a certain area in the (φ, ψ) space. There are few common types of secondary structure, the most prominent being the α-helix and the β-sheet conformation.
- *Tertiary structure* refers to the spatial relationship among all amino acids in a polypeptide; it is the complete three-dimensional (3D) structure of the polypeptide.
- *Quaternary structure* refers to the spatial relationship of polypeptides, or subunits, within the protein (usually not bound covalently).

The relationship between sequence, structure and function of proteins is one of the central topics in molecular biology.

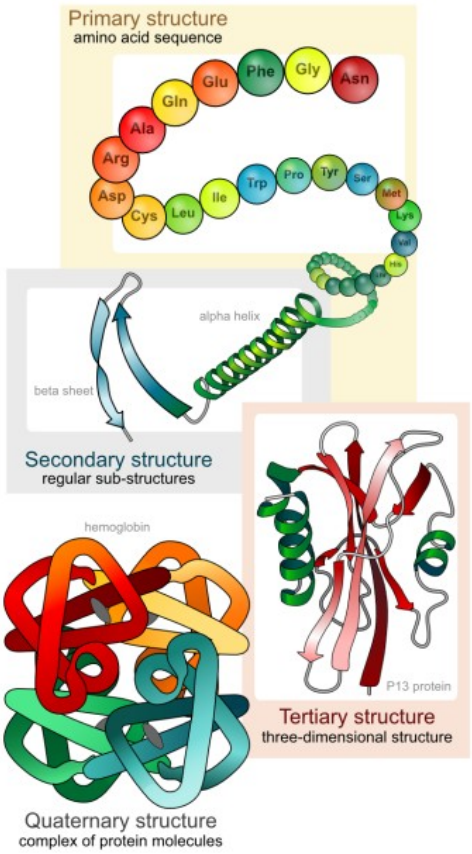

Figure 1.1.4. Main protein structural levels.

1.2 Protein interactions

Protein-protein interactions involve the association of protein molecules. These interactions are crucial for many biological functions. Here we briefly describe the physico-chemical basis of protein interactions, define the concepts of protein complexes and protein interfaces.

1.2.1 Physico-chemical basis of protein interactions

Interactions between protein atoms allow a protein to adopt a 3D-structure as well as define the association between different proteins. All protein-protein interactions could be divided into covalent and non-covalent interactions. A covalent bond is a form of chemical bonding that is characterized by the sharing of pairs of electrons between atoms, or between atoms and other covalent bonds. In short, attraction-to-repulsion stability that forms between atoms when they share electrons is known as covalent bonding. Thus, covalent interactions define primary structure and sometimes tertiary and quaternary structure (e.g. disulfide bonds between side chains of cysteine residues). They could be described approximately by harmonic potential:

$$V_{bond}=V_{bond\,0}+k_{bond}(r-r_0)^2 \qquad (1.2.1),$$

where $V_{bond\,0}$ is an equilibrium potential, $r-r_0$ is the shift of the coordinate from the equilibrium and k_{bond} is a bond "spring" constant.

Non-covalent interactions are critical for maintaining structures of large macromolecules, including proteins and nucleic acids. These interactions could be arbitrarily classified into electrostatic, van der Waals, hydrogen bonding and hydrophobic interactions.

Electrostatic interaction is described by the Coulomb law. For the system of point charges associated with individual atoms, which most of the time represents adequate approximation for proteins, the electrostatic potential is usually represented by charge-charge, dipole-charge, dipole-dipole and charge-induced dipole interactions:

$$V_{el}=\sum_{i,j=1;i>j}^{N}\frac{1}{\varepsilon}\left(\frac{q_i q_j}{r_{ij}}+q_i\vec{\mu}_j\frac{\vec{r}_{ij}}{r_{ij}^{3}}+\frac{\vec{\mu}_i\vec{\mu}_j}{r_{ij}^{3}}-3\frac{(\vec{\mu}_i\vec{r}_{ij})(\vec{\mu}_j\vec{r}_{ij})}{r_{ij}^{5}}+\ldots\right) \qquad (1.2.2),$$

where N is the number of atoms, q_i and q_j, $\vec{\mu}_i$ and $\vec{\mu}_j$ are charges and dipole moments of the i^{th} and j^{th} atoms, respectively, is dielectric constant, $\vec{r}_{ij}$ is the radius-vector between charges and dipole moments.

Typical charge-charge interactions that favor protein folding are those between oppositely charged R-groups such as Lys or Arg and Asp or Glu (*salt bridges*). A substantial component of the energy involved in protein folding is charge-dipole interactions. This refers to the interaction of ionized R-groups of amino acids with the dipole of the water molecule. The slight dipole moment that exist in the polar R-groups of amino acid also influences their interaction with water. It is, therefore, understandable that the majority of the amino acids found on the exterior surfaces of globular proteins contain charged or polar R-groups.

There are both attractive and repulsive *van der Waals (VDW) interactions* that contribute to protein interactions. Attractive VDW forces involve the interactions among induced dipoles that arise from fluctuations in the charge densities that occur between adjacent uncharged non-bonded atoms. Repulsive VDW forces involve the interactions that occur when uncharged non-bonded atoms come very close together but do not induce dipoles. The repulsion is the result of the electron-electron repulsion that occurs as two clouds of electrons begin to overlap (Figure 1.2.1). Although VDW forces

are extremely weak, relative to other forces governing conformation, it is the huge number of such interactions that occur in large protein molecules that make them significant to the folding of proteins.

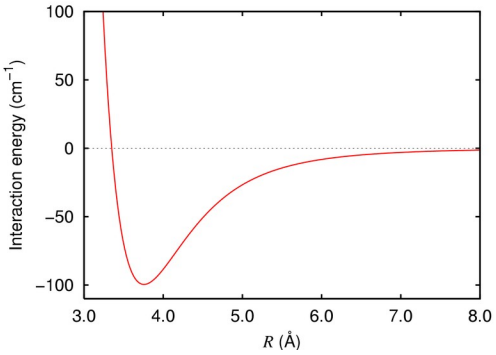

Figure 1.2.1. Van der Waals interaction. Interaction energy of argon dimer.

Van der Waals interaction between two atoms is described by a potential as following:

$$V_{vdw} = \frac{A}{r^m} - \frac{B}{r^6} \qquad (1.2.3),$$

where A and B are positive constants and r is the distance between interacting atoms, m is the power of the repulsive term (usually between 5 and 12). If m=12 the potential is called *Lennard-Jones potential* or *6-12 potential*.

Hydrogen bonding interaction is related to the formation of a *hydrogen bond*, an attractive force between one electronegative atom (*acceptor of hydrogen bond*) and a hydrogen covalently bonded to another electronegative atom (*donor of hydrogen bond)*. It results from a dipole-dipole force with a hydrogen atom bonded to nitrogen, oxygen or fluorine (thus the name "hydrogen bond", which must not be mixed with a covalent bond to hydrogen). *Hydrogen bond interaction* has a potential, which resembles the VDW potential but with involvement of hydrogen atom:

$$V_{vdw} = \frac{C}{r^{12}} - \frac{D}{r^6} \qquad (1.2.4),$$

where C and D are positive constants and r is the distance between interacting atoms. Polypeptides contain numerous proton donors and acceptors both in their backbone and in the R-groups of the amino acids. The environment in which proteins are found also contains ample H-bond donors and acceptors of the water molecule. H-bonding, therefore, occurs not only within and between polypeptide chains

but with the surrounding aqueous medium. Energies of hydrogen bonds strongly depend on a donor-acceptor pair as well as on surrounding environment. However, typically these values for protein environments are in the range of 1-10 kcal/mol.

Hydrophobic interactions are explained in terms of *hydrophobic effect*, which is the property of non-polar molecules to form intermolecular aggregates in an aqueous medium. At the macroscopic level, the hydrophobic effect is apparent when oil and water are mixed together and form separate layers or the beading of water on hydrophobic surfaces such as waxy leafs. At the molecular level, the hydrophobic effect is an important driving force for biological structures and responsible for protein folding, protein-protein interactions, formation of lipid bilayer membranes, nucleic acid structures, and protein-small molecule interactions. In the case of protein folding, the hydrophobic effect is important to understand the structure of proteins that have hydrophobic amino acids, such as Ala, Val, Leu, Ile, Phe, and Met grouped together with the protein. Most folded proteins have a hydrophobic core in which side chain packing stabilizes the folded state, and charged or polar side chains on the solvent-exposed surface where they interact with surrounding water molecules. It is generally accepted that minimizing the number of hydrophobic side chains exposed to water is the principal driving force behind the folding and protein association processes.

All non-covalent interactions could be also classified by distance-dependence (Table 1.2.1) and interaction energies typical values. In comparison to covalent bonds, non-covalent bonds are about one order weaker in terms of energy (~100 kcal/mol vs ~10 kcal/mol) but as stated above, their impact for protein folding and protein-protein association processes is crucial.

Table 1.2.1. Relationship of non-covalent interactions to the distance between interacting atoms (r)

Type of interaction	Distance Relationship
Charge-charge	$1/r$
Charge-dipole	$1/r^2$
Dipole-dipole	$1/r^3$
Charge-induced dipole	$1/r^4$
Dispersion	$1/r^6$
Repulsion	$1/r^{12}$

1.2.2 Protein complexes and interfaces

Protein complex is a group of several proteins, which are associated by non-covalent protein-

protein interactions. Protein complexes could be classified by their stability over time as *transient* or *permanent*. The first ones are characterized by a short and the latter by a long complex lifetime. A *homomultimeric* complex consists of identical subunits and *heteromultimeric* complex is made up of different subunits. Protein complexes formed by protein chains where the process of folding and binding is essentially inseparable are *obligated* (i.e. multi-subunit enzymes) [12,13]. On the other hand, complexed formed by proteins that fold independently and then associate to carry out a particular biological task are *non-obligated*. However, it remains arbitrary to classify protein complexes due to the overlap between types and the limitation of their biological annotation (i.e. localization, coexpression, or binding energies) [14].

Protein interfaces are defined by atoms, which participate in protein-protein interactions. Size of an interface is usually characterized by the difference between *accessible surface area* of unbound interacting counterparts and of a complex. Protein interfaces have been studied at protein chain and domain interface levels [12,15-21]. Many databases containing structural domain-domain interactions have been recently created: 3did [22], PiBase [23], iPfam [19], PSIbase [24], InterPare [25], PRISM [26]. Nevertheless, most current methods do not provide an accurate description of protein interfaces, which is required to be able to establish the bases for understanding the principles that govern molecular recognition and protein function [27]. In addition, all these databases and methods use different definitions for protein interfaces that makes related studies ambiguous to compare with each other. In many studies a simple criteria of distance cut-off between atoms is used to define protein interfaces [28,29]Although this cut-off based definition is easy and fast to implement in computational approaches, there is a low relevance to physico-chemical nature behind protein-protein interactions. In our group, SCOWLP (Structural Characterization of Water, Ligands and Proteins), the database for characterization of protein interfaces, was created based on the interface definition, which contains data obtained from the whole PDB and includes physico-chemical properties of individual atoms in protein complexes and interfacial solvent [27]. To define an interaction SCOWLP considers a donor–acceptor distance for hydrogen bonds 3.2 Å, for salt bridges 4 Å, and for VDW contact the VDW radii distance. Interfacial residues, whose atoms fulfill these criteria, are classified according to the interaction type as *dry* (direct interaction), *dual* (direct and water-mediated interactions), and *wet spots*, which are residues interacting only through one water molecule (Figure 1.2.2 A). The most important feature of this interface description is that it takes into account water explicitly into the protein interface definition. Currently, SCOWLP contains 74907 protein interfaces and 2093976 residue-residue interactions

formed by 60664 structural units (protein domains and peptidic ligands) and their interacting solvent. Statistical analysis of the SCOWLP data showed that 40.1% of the interfacial residues are interacting through water and that wet spots represent a 14.5% of the total, emphasizing the contribution of wet spots to interfacial description (Figure 1.2.2 B).

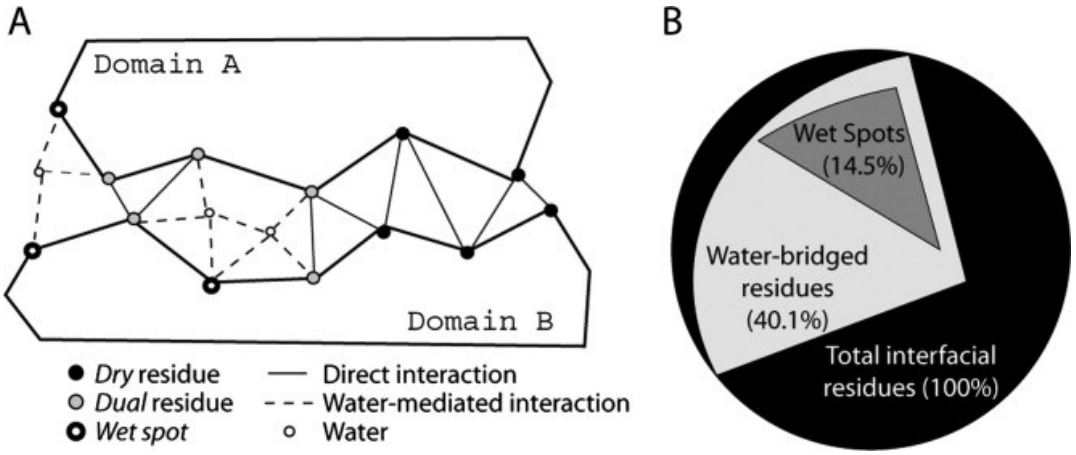

Figure 1.2.2. Residue interaction types. A) Definition of residue interactions: Interface between domains A and B is formed by 13 residues (five dry, five dual, and three wet spots). B) Partition of residue interactions [27].

1.3 Solvent in protein interactions

Despite the fact that protein interactions take place in aqueous solution, solvent is frequently ignored in protein interaction analysis and, for this reason, many protein residues are not taken into account in interface definition. This subsection focuses on the recent findings in the field of protein-solvent relations and gives a short review on how solvent effects are modeled computationally.

1.3.1 Water unique properties.

Physical properties of water, despite its abundance in environment, are unusual compared to other fluids. Although water is similar to them in its VDW attraction and repulsion interaction terms, it has the ability to form hydrogen bonds and a three-dimensional tetrahedral network-like structure. Thus, compared to other liquids with the same molecular size, water is more cohesive, as indicated by its higher boiling and freezing temperatures, surface tension, and vaporization enthalpies. Further, it has a high dielectric constant and exists in numerous crystalline forms. Liquid water's fluidity increases with increasing pressure. The mobility of H^+ and OH^- ions is higher in water than in other liquids. Water also has volumetric anomalies. Whereas most solids are denser than their corresponding liquids, ice floats on water. Also, a typical liquid's density decreases monotonically with increasing temperature. For water, this is true only at high temperatures (above 3.984ºC, the temperature of

maximum density). Other related anomalies include minima in the isobaric heat capacity and isothermal compressibility with temperature in the normal liquid range (at 36ºC and 46ºC, respectively) [30].

1.3.2 Role of solvent in protein interactions.

Though cellular solvent is sometimes erroneously considered as an inert environment for the biomolecular machinery, various studies have indicated particular importance of individual water molecules for structural, catalytic, energetic and dynamical behavior of proteins [31]. At the same time properties of water molecules themselves in close proximity of proteins (surfacial solvent) or in protein cavities differ significantly from the ones of bulk solvent in terms of residence time, mobility, energetic impact and evolutionary conservation [32]. As an example of water-mediated interaction conservation, a water site in coiled-coil interfaces was found to be structurally conserved in many three-stranded coiled-coils, and together with charged residues forming a structural motif that determines three-stranded coiled-coil formation by water bridge formation [33]. In another study, the number of water molecules in a chaperonin cavity observed in molecular dynamics (MD) simulations has been shown to be correlated with experimentally measured folding rates of proteins. Thus water impact has been proved to be important in chaperonin-assisted folding process [34]. These data suggest that water molecules actively participate in the process of folding, one of the key processes in molecular biology. Another case of a crucial impact of hydration of a protein cavity for the function of the protein is demonstrated for β-lactoglobulin pore. Here, apo- and holoforms of the enzyme differ significantly in the rate of ligand association binding affinity, what could be attributed to the essential differences in hydration properties of the catalytic pore [35]. Thermodynamic analysis of high affinity peptides binding to the Abl-SH3 domain revealed that maintenance of a complex hydrogen-bond network mediated by water molecules buried at the binding interface is responsible for the thermodynamic behavior observed in calorimetry experiments [36]. Similarly, a computational study of high affinity epitopes binding to the class I MHC complex has found that the entropy of water molecules participating in binding interfaces is essentially different from the bulk solvent. This work concludes that solvent plays an active role of mediator of interactions in MHC-peptide system [37]. Additionally for MHC-peptide system, 63 conservative clusters of water have been identified in the study of a non-redundant dataset of complexes at high resolution. Water molecules at these positions are supposed to be associated with modulation of peptide recognition [38]. Other analyses of crystal structures of

conserved water molecules with high residence time and low mobility have demonstrated structural and catalytic importance of these water molecules in ribozyme [39] and microbial ribonucleases [40]. These examples underline the importance of detailed solvent studies for broad variety of protein systems.

There is a concept that micro-osmosis explains why water changing properties near protein surfaces are driving forces for protein-protein interactions in the cell. According to this idea, water could be classified into high density and low density water. These two states are metastable and transient, but they appear to last long enough to play significant roles in biochemical and physiological processes and determined by both thermodynamic and kinetic considerations. Low density water is more strongly hydrogen-bonded and more viscous compared to normal water. High density water is more fluid and has highly selective solvent properties compared to low density water. These two different water states exist near surfaces of proteins depending on their electrostatic characteristics. At hydrophobic surfaces, there is a limitation for water molecules to form hydrogen bonds as much as near charged surfaces, that is why water density at hydrophobic surfaces is lower. Aggregation of proteins independently of their hydrophilic nature could be explained in terms of three forces: electrostatic, van der Waals and micro-osmotic force driven by water density [41]. There is a number of biochemical and polymer chemistry evidence that demonstrate different levels of reactivity of solutes in high and low density water. The state of water, in turn, could be determined by chaotropic (disordering) or kosmotropic (ordering) effect of biopolymers irregular surfaces and of ions in the solution [42].

Energetically, water molecules associated with protein interfaces have been shown to contribute essentially to thermodynamic properties of a protein complex. Use of simple models based on the hydrogen bonding propensity of water as a function of temperature gives quantitative estimates of heat capacity that agree well with experimental (calorimetry) observations for both protein folding and ligand binding. Impact of an individual water molecule upon exothermic binding has been estimated with changes of enthalpy, entropy and heat capacity in ranges of −1.5 to −3 kcal/mol, ≈−2.5 kcal/mol K and ≈−20 kcal/mol K, respectively [43]. Free energy perturbation studies also aimed to estimate an energetic impact of individual water molecules to the protein binding site. Application of a thermodynamic integration approach for water molecules in binding pockets of trypsin and HIV protease yielded free energies in ranges from slightly positive values to -4 kcal/mol [44,45]. Monte Carlo simulations allowed to classify water molecules in protein binding sites into two classes: displaced and remaining upon ligand binding. The molecules belonging to the second class, in general, form more hydrogen bonds, locate in more polar environment and are characterized with lower free

energies. The most stable water molecules give an impact of up to -10 kcal/mol to the complex formation [46]. A MM-PBSA study of T-cell receptor with its enterotoxin ligand demonstrated that inclusion of only two structurally conservative interfacial water molecules into free energy calculations significantly lowers the energy of the system [47]. Inclusion of the water term into the Hamiltonian explicitly leads to better results in a protein folding study [48].

Statistically, MD and X-ray analysis suggest that being incorporated into protein environments, water molecules try to retain their tetrahedral hydrogen bond geometry and enable the extension of the 3D chain connection of a hydrogen bond network among hydration water molecules and protein atoms in protein interfaces. These networks of hydrogen bonds are flexible enough to control the conformational changes of proteins as domain motions [49,50]. In general, protein motions and dynamics of interfacial water molecules are tightly interconnected [51]. MD simulations demonstrate that movements of protein structure elements, especially of loops, are strongly correlated with hydration dynamics of protein [52].

There are two hydration shells found near protein surface with the corresponding maximums of water density (radial distribution function for water oxygens) at 2.75 Å and 4.50 Å from protein surface. Within these distances dynamical properties of water molecules could be significantly distinguished from bulk solvent [53]. These differences are often explained in term of chemical heterogeneity at the protein surface or by its surface roughness [54,55]. *Residence time* is one the most well-established characteristics of water molecules on protein surfaces and hydration sites. Usually there are two exponential components describing water residence time distribution in proximity of protein surface: long and short residence time components. The long component dominates with the decrease of the distance to the protein and does not correlate with density of water molecules in space [56]. The long residence time component is related to vacancy times for a single water molecule, corresponding to kinetically bound molecules, which comprise only a small fraction of the total number of occupancy sites and are correlated with local heterogeneities in both surface charge and roughness. Short residence times are physically associated with a high-speed turnover involving multiple water molecules [57]. Fluorescence experiments and MD simulations define long residence time components in ranges of 100 ps and higher, while short residence time component is usually below 10 ps [58-60] . These data add to the evidence that properties of solvent change drastically in proximity of protein surface.

Despite all these above discussed findings suggesting the important role of solvent in protein-

protein interactions, there are still very few computational studies that take water molecules into account. However, an attempt to introduce an explicit solvent term into a Monte Carlo simulation based docking algorithm yielded promising results [61]. To model water networks several algorithms have been developed based either on physico-chemical properties of protein-protein interfaces (WATGEN [62]) or knowledge-based potentials and free energy calculation [63]. Statistical analysis of the data on water including protein interfaces represented in SCOWLP database [27] suggest that wet spots present similar characteristics to residues binding buried water molecules in the core or cavities of proteins [64]. In the same study contact matrices for dry and water mediated interactions have been derived.

The Chapter 2 of this thesis focuses on the role of water in protein interfaces. First, a MD study of water-mediated protein interactions has been carried out on a representative set of protein complexes (section 2.1). Then, the data on water mediated interactions available from the PDB is applied for the protein structure *ab initio* prediction in the section 2.2.

1.3.3 Computational models of solvent.

In general, computational models of solvent could be classified into two groups: explicit and implicit solvent models. *Explicit solvent* models treat solvent as individual molecules, each of which has its microscopic properties. As opposite to explicit solvent, i*mplicit solvent* represents solvent as a continuous dielectric medium instead of individual solvent molecules. Both methods are developed to reproduce macroscopic properties of solvent in simulations. Implicit solvent models (MM-PBSA/MM-GBSA) are described in details in the subsection 1.5.4 of this chapter. Here, we review some explicit solvent models used in computations.

Classification of the explicit solvent models is based on the number of points (Figure 1.3.1) used to define the model (atoms plus dummy sites), bonds flexibility, and inclusion of polarization effects.

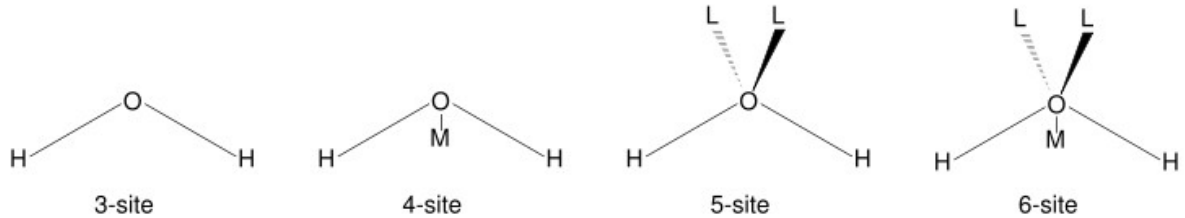

Figure 1.3.1. The 3- to 6-site water models. The OH distance and the HOH angle vary depending on the model. L is a lone pair, M is a dummy atom.

The simplest water models treats the water molecule as rigid and participating only in non-

bonded interactions. Electrostatic interaction is modeled using Coulomb's law and the dispersion and repulsion forces using the Lennard-Jones potential. The potential for SPC or TIP models such as TIPS [65] , TIP3P, TIP4P [66] and TIP5P [67] is represented by:

$$V_{ab} = \sum_{i}^{on\,a} \sum_{i}^{on\,b} \frac{k_c q_i q_j}{r_{ij}} + \frac{A}{r_{OO}^{12}} - \frac{B}{r_{OO}^{\neg}} \qquad (1.3.1),$$

where k_c, the electrostatic constant, has a value of 332.1 Å·kcal/mol; q_i are the partial charges in electron charge unit; r_{ij} is the distance between two atoms or charged sites; and A and B are the Lennard-Jones parameters. The charged sites may be on the atoms or on dummy sites (such as lone pairs). In most water models, the Lennard-Jones term applies only to the interaction between the oxygen atoms.

The models with three interaction sites (TIPS [65], TIP3P [66], SPC [68], SPC/E [69]), corresponding to the three atoms of the water molecule are popular in MD because of their simplicity and computational efficiency. Each atom gets assigned a point charge, and the oxygen atom also gets the Lennard-Jones parameters. Most models use a rigid geometry matching the known geometry of the water molecule. An exception is the SPC [68] model, which assumes an ideal tetrahedral shape (HOH angle of 109.47°) instead of the observed angle of 104.5°.

The 4-site models (BF [70], TIPS2, TIP4P [66], TIP4P-Ew [71], TIP4P/Ice [72], TIP4P/2005 [73]) have a negative charge on a dummy atom (Figure 1.3.1) placed near the oxygen along the bisector of the HOH angle. This improves electrostatic distribution around the water molecule. The first model to use this approach was the Bernal-Fowler model published in 1933, which may also be the earliest water model [70]. However, the BF model does not reproduce well the bulk properties of water, such as density and heat of vaporization. New 4-site models were parameterized by iteratively running Metropolis Monte Carlo or MD simulations and adjusting the parameters until the bulk properties are reproduced well enough.

The TIP4P model published first in 1983 [66] is the most popular 4-site model used for simulation of biomolecular systems. There have been subsequent reparameterizations of the TIP4P model for specific uses: the TIP4P-Ew model, for use with Ewald summation methods [71]; the TIP4P/ Ice, for simulation of solid water ice [72]; and TIP4P/2005, a general parameterization for simulating the entire phase diagram of water [73].

The 5-site models (BNS, ST2 [74], TIP5P [67], TIP5P-E [75]) place the negative charge on dummy atoms (Figure 1.3.1) representing the lone pairs of the oxygen atom, to provide a tetrahedral-like geometry. The BNS and ST2 [74] models do not use Coulomb's law directly for the electrostatic terms, but a modified version that is scaled down at short distances by multiplying it by the switching function. Mainly due to their higher computational cost, these early five-site models were not developed much until recently, when the TIP5P model was published [67]. When compared with earlier models, the TIP5P model results in improvements in the geometry for the water dimer, a more "tetrahedral" water structure that better reproduces the experimental radial distribution functions from neutron diffraction, and the temperature of maximum density of water. The TIP5P-E model is a reparameterization of TIP5P for use with Ewald sums [75] .

A 6-site model that combines all the sites of the 4- and 5-site models [76] is able to reproduce the structure and melting of ice better than other models.

The Mencedez-Benz (MB) model is a model resembling the Mercedes-Benz logo that reproduces some features of water in 2D systems. It is not used for simulations of 3D systems, but it is useful for qualitative studies and for educational purposes [77]. In the MB model, the energy of interaction between two water molecules is the sum of a Lennard-Jones potential and an orientation-dependent hydrogen bond interaction. Neighboring water molecules form an explicit hydrogen bond when an arm of one water molecule aligns with an arm of another water molecule; the corresponding energy is a Gaussian function of both separation and angle. Hydrogen bonding arms are not distinguished as donors or acceptors. The strength of a hydrogen bond is determined only by the degree of alignment of arms on two neighboring waters. In 3D, computational modeling is unable to reversibly freeze water for most of water models because simulations get stuck in deep kinetic traps, whereas reversible freezing and melting are readily studied in the two-dimensional MB model [30].

One- and two-site coarse-grained models of water have also been developed [78]. In coarse grain models, each site can represent several water molecules.

A polarization term could be introduces to equation 1.3.1, as, for example, in SPC/E model [69] by addition of an average polarization correction:

$$V_{pol} = \frac{1}{2} \sum_i \frac{(\mu - \mu_0)^2}{\alpha_i} \qquad (1.3.2),$$

where μ is the dipole of the effectively polarized water molecule (2.35 D for the SPC/E model), μ_0 is the dipole moment of an isolated water molecule (1.85 D from experiment), and α_i is an isotropic polarizability constant, with a value of 1.608 × 10−40 F m. Since the charges in the model are constant, this correction just results in adding 1.25 kcal/mol to the total energy. The SPC/E model results in a better density and diffusion constant than the SPC model.

Choice of a water model could be critical for the results of the study if some specific properties of solvent are investigated. In all simulations we utilized widely the used TIP3P water model.

1.4 Protein engineering and non-canonical amino acids

Protein engineering is a directed construction of proteins with new properties such as increased thermostability, altered binding specificity, improved binding affinity or enhanced enzymatic activity. Protein engineering contributes to the understanding of protein folding and protein recognition for protein design principles. Two general strategies for protein engineering are used. The first one is rational design, in which the detailed knowledge of the structure and function of a protein are used to make desired changes in its properties. The second strategy for protein engineering is a directed evolution. In this case random mutagenesis is applied to a protein, and a selection criteria is used to find variants with the desired qualities.

Protein design (rational design technique used in protein engineering) is the design of new protein molecules by making calculated variations on a known structure. Computational protein design algorithms require an understanding of the molecular interactions that stabilize proteins in specific folded configurations. However, protein design does not require an understanding of the dynamical process by which proteins fold. In a sense it is the reverse of structure prediction: a tertiary structure is specified, and an amino acid sequence is identified which will fold to it. Protein design investigations pursue both scientific and engineering goals, using design to test and advance our understanding of underlying biophysical interactions. There are some challenges in the field of computational protein design, which one faces while applying related approaches.

1. *Development of energy functions*. Protein design strongly relies on energy functions used to evaluate obtained results. Development of energy functions includes understanding and validating their applicability to design studies, developing new potentials and target functions where appropriate, and improving efficiency through both better algorithms and approximations. A common difficulty reported in computational design efforts is the accurate evaluation of electrostatic solvation and interaction

terms. Sometimes the electrostatic term of the binding energy was even found to be a better predictor than the total energy for affinity improvements [79].

2. *Treatment of solvent*. Because proteins are surrounded by water, which is highly polarizable, any accurate description of protein requires to treat electrostatics as a function of the solvent environment, making the electrostatic energy a manybody term. Thus it is necessary to reconcile the limitations of the pairwise approximation with the need for an accurate description of electrostatics. Furthermore, modeling of water explicitly is currently intractable for the number of conformational energies that must be calculated for protein design. Therefore, continuum or empirical models have been used in most protein design force fields to address electrostatic interactions as well as polar desolvation [80]. However, the placement of individual water molecules, particularly bridging protein complexes, could be crucially important in natural and designed proteins. Baker and colleagues have introduced a new energetic description of water-mediated hydrogen bonds and combined it with a 'solvated rotamer' approach to place interfacial water molecules using conventional rotamer search techniques [63]. Solvent treatment has always a trade-off between accuracy and effectiveness of an used approach. More details on this topic can be found in the subsection 1.3.2 and the section 1.5 of this thesis.

3. *Dealing with conformational ensembles*. Such useful characteristics of a studied protein system as kinetic energy, full free energy or entropy could be properly estimated only when taking into account statistical ensembles for analysis. The role of entropy calculations, which are usually the most computationally demanding part in protein design, is still widely discussed [81]. Kuhlman and colleagues used a protein design procedure based on Monte Carlo search to include side-chain conformational entropy in the design of 110 native protein backbones. They found very little difference in the resulting sequence designs whether entropy was included or not, with the largest differences involving long, flexible side chains [82]. Even if conformational entropy contributions are not dominant in protein design calculations, the use of ensembles is likely to have other benefits in protein design engineering [79].

4. *Taking into account non-contacting residues near binding interaface*. An approach for introduction of non-contacting residues near a binding interface allowed to enhance affinity by virtue of paying very little desolvation penalty yet making larger 'action-at-a-distance' intermolecular interactions [83].

5. *Enzymes characterization*. It is not well understood how many of natural enzymes function, and thus the optimization objectives for computational enzyme design are unclear. Factors that may be important include binding to transition state of substrates, accommodation of substrate, release of product, protein

flexibility and dynamics, and active-site catalytic residues. Though computational design has been used to stabilize proteins, enzyme stabilization is complicated by the need to maintain catalytic activity [79].

6. *Search and optimization algorithms*. To speed up computationally demanding tasks in protein design numerous new algorithms are being developed. Energy minimization uses wide variety of methods such as discrete approaches (dead-end elimination, genetic algorithms, and integer programming etc.), as well as faster, non-guaranteed methods (Monte-Carlo, self-consistent mean field theory etc.) [79,84] .

7. *Human intervention*. Unfortunately for high-throughput approaches, most protein design methods are not free of human intervention. The use of hand curation is common for selecting or refining designs, as opposed to a fully automated methodology.

Utilization of non-standard for protein environments components as, for example, non-biological amino acids and cofactors opens up new broad perspectives for protein engineering, yielding protein systems poised to present new functions that have not previously been accessible [84]. This approach requires both basic theoretical studies of non-standard components and introduction of them into simplified model systems emulating protein environments. Chapter 3 of this thesis describes this challenge in respect to fluorinated amino acids. In particular, we provide a theoretical study of fluorinated amino acid properties and a study behavior of fluorinated amino acids within two distinct protein model systems: coiled-coil and chymotrypsin catalytic site.

1.5 Computational approaches to study protein interactions

This section does not represent an exhaustive overview of existing computational methodologies in the field of protein-protein interactions but rather gives a brief general introduction to methodologies used in this thesis and the theoretical concepts behind them.

1.5.1 Sequence-based approaches

A *sequence alignment* is a way of arranging sequences of DNA, RNA, or proteins to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships. Aligned sequences of nucleotides or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. Very short or very similar sequences can be aligned by hand. However, often an alignment of lengthy, highly variable or extremely numerous sequences is required, which cannot be

carried out solely by human effort. Therefore, algorithms are developed to produce high-quality sequence alignments, and the final results are curated to reflect patterns that are difficult to represent algorithmically. Computational approaches to sequence alignment generally fall into two categories: global alignments and local alignments. Calculating a *global alignment* is a form of global optimization that "forces" the alignment to span the entire length of all query sequences. By contrast, *local alignments* identify regions of similarity within long sequences that are often widely divergent overall. Local alignments are often preferable, but can be more difficult to calculate because of the additional challenge of identifying the regions of similarity. A variety of computational algorithms have been applied to the sequence alignment problem, including optimizing methods like dynamic programming, heuristic algorithms and probabilistic approaches designed for a large-scale database search.

Sequence alignments could be divided into two other groups: pairwise alignments and multiple sequence alignments. *Pairwise sequence alignment* methods are used to find the best-matching local or global alignments of two query sequences. *Multiple sequence alignment* is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all sequences in a given query set. Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related. Such conserved sequence motifs can be used in conjunction with structural and mechanistic information to locate the catalytic active sites of enzymes, for example. Alignments are also used to aid in establishing evolutionary relationships by constructing *phylogenetic trees*. Multiple sequence alignments are computationally difficult to produce and most formulations of the problem lead to combinatorial optimization problems [85]. Nevertheless, the utility of these alignments in bioinformatics has led to the development of a variety of methods suitable for aligning three or more sequences [86].

In the thesis we utilized *CLUSTAL* [85,87-91], one of the most popular global multiple sequence alignments programs. It works as follows:

1. Performs pairwise alignments of all sequences.
2. Uses alignment scores to produce a phylogenetic tree joining the closest neighbours.
3. Aligns the sequences sequentially guided by the phylogenetic relationship indicated by the tree.

Thus, the most closely related sequences are aligned first, and then additional sequences and groups of sequences are added, guided by the initial alignments to produce multiple sequence alignment showing in each column the sequence variations among the sequences. For producing of phylogenetic tree, genetic distances (the number of mismatched positions in an alignment divided by

the total number of matched positions with ignoring the positions with opposite gaps) between the sequences are required. Sequence contributions to multiple sequence alignment are weighted according to their relationships on the predicted evolutionary tree based on the distance of each sequence from the root. The alignment scores between two positions in the multiple sequence alignment are calculated using the resulting weights as multiplication factors (Figure 1.5.1).

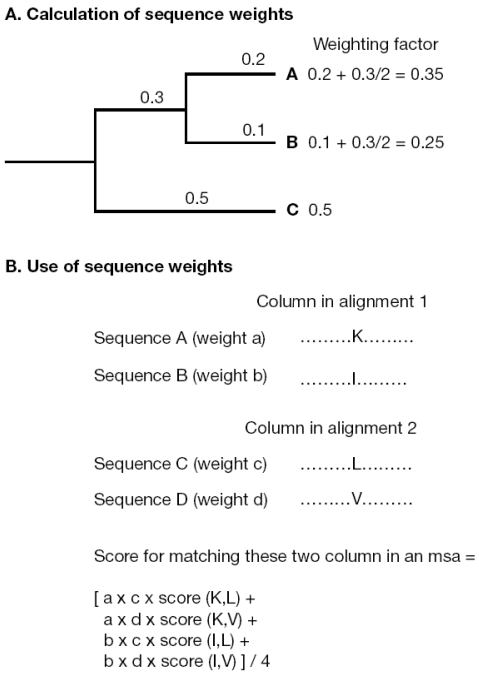

Figure 1.5.1. Weighting scheme used by CLUSTAL [89]. A. Sequences that arise from a unique branch deep in the tree receive a weighting factor equal to the distance from the root. Other sequences that arise from the branches shared with other sequences receive a weighting factor that is less than the sum of the branch lengths from the root. For example, the length of a branch common to two sequences will only contribute one-half of that length to each sequence. Once the specific weighting factors for each sequence have been calculated, they are normalized so that the largest weight is one. As CLUSTAL aligns sequences or group of sequences, these fractional weights are used as multiplication factors in the calculations of alignment scores. B. Illustration of using sequence weights for aligning two columns in two separate alignments.

The scoring of gaps in multiple sequence alignment is performed in a different manner from scoring gaps in pairwise alignment. As more sequences are added to a profile of an existing multiple sequence alignment, gaps accumulate and influence the alignment of further sequences [87]. Like other alignment programs, CLUSTAL uses a penalty for opening gap in a sequence alignment and an additional penalty for extending the gap by one residue. These penalties are user-defined. Gaps found in initial alignments remain fixed. New gaps introduced as more sequences are added also receive this same gap penalty, even when they occur within an existing gap, but the gap penalties for an alignment are then modified according to the average match value in the substitution matrix, the percent identity

between the sequences [89].

CLUSTAL also has options for adding one or more additional sequences with weights or an alignment to an existing alignment. Once an alignment has been made, a phylogenetic tree may be made by the neighbor-joining method, with corrections for possible multiple changes at each counted position in the alignment.

Multiple sequence alignments contain information, which reflects evolutionary and functional conservation of protein sequences. Thus, results obtained by sequence alignments are utilized for protein structure and function predictions. *Correlated mutations* approach is a representative of sequence alignment based structure prediction methods. The idea of the concept behind is that interacting protein residues co-evolve, so that a mutation in one of the interacting counterparts is compensated by a mutation in the other (Figure 1.5.2) [92]. This co-evolution is revealed for the residues within the same protein domain and for the pairs of different interacting protein domains.

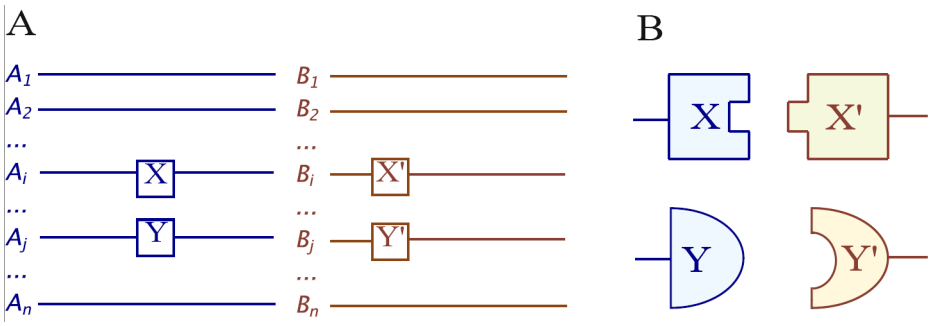

Figure 1.5.2. Concept of correlated mutations. $\{A\}_i$ and $\{B\}_i$ are families of sequences. A) Sequence alignment of $\{A\}_i$ and $\{B\}_i$ sequences. For the i^{th} and j^{th} members of both families there are mutations X to Y and X' to Y', corresponding to one of positions in each alignment, in which the residues are supposed to interact. B) Mutation of X to Y in A family is compensated by the mutation X' to Y' in B family to keep the existing interaction between the positions in the counterparts.

There are many different methods and algorithms used for the implementation of this concept for practical applications. The differences between the methods correspond to the following steps:

1. Alignment. Sequence alignments or structure-based sequence alignments could be used at this stage. The following parameters for input sequences should be specified before the final alignment is done: highest and lowest sequence similarity, minimum and maximum length, and minimum number of input sequences.

2. Scoring positions in an alignment. First of all, the measure of variance for each of positions in alignments should be defined. *Similarity matrices* for amino acids are used for this purpose. The source of data used for similarity matrices could be based on physico-chemical properties of amino acids or probabilistic information obtained from different data. Then, after assigning scores for positions, only

the positions within the specified allowed range of variance are taken into account for the further analysis.

3. Scoring pairs of positions. To predict if residues in the analyzed positions could interact, the positions scores should be used for application of a function, which deals with pairs of positions and ranked. There is a vast variety of scores used for pairs based on: covariance, informational entropy measures, probabilistic models of co-occurence etc. [93]

4. Ranking and making predictions. After obtaining scores for pairs of sequential positions, they are ranked. Best N contacts are compared with experimental data. N is usually dependent on the length L of sequences in an alignment, e.g. N=L, L/2, L/10. Distribution of distances between predicted to be in contact residues is analyzed. One of restrictions that could be introduced at this step is a *sequence separation*: pairs are considered for analysis only if they are separated in sequence by more than *n* positions.

5. Contact definition. It is important to notice that different definitions of residue contacts could lead to different results of predictions. Some contact definitions are based only on distant cut-offs between residue atoms (e.g. in [94]) and others take into account physico-chemical properties of the interacting atoms (see section 2.2).

6. Predictions assessment. Usually standard parameters to estimate predictions such as accuracy, sensitivity and specificity are used. However, depending on a length of a sequence it could be useful to analyze how big is an improvement of prediction compared to random chance to predict a pair of residues to form a contact.

7. Neural nets approach could be implemented into the predictive pipeline (e.g. [95]).

However, despite significant progress in development of the approaches, relatively high levels of false-positive predictions typically render such methods of little use in the *ab initio* prediction of protein structure, when they are used alone [96]. Typically, accuracy of the approach is in the range of 0.1-0.4 [97].

In section 2.2 the study on topic of correlated mutations is presented and more details of methodology can be found.

1.5.2 Structure-based sequence alignment

Structure-based sequence alignment is a form of sequence alignment that is based on structural comparison. These alignments attempt to establish equivalences between sequences of two or more proteins based on their structural similarity. Structure-based sequence alignment is a valuable tool for

comparison of the protein sequences with low sequence similarity, where evolutionary relationships between proteins cannot be easily detected by standard sequence alignment techniques. Structural alignment can, therefore, be used to imply evolutionary relationships between proteins that share very little common sequence. However, caution should be taken in using the results as evidence for shared evolutionary ancestry because of the possible confounding effects of convergent evolution by which multiple unrelated amino acid sequences converge to a common tertiary structure. Similarly to sequence-based sequence alignments structure-based alignments can deal with pairwise or multiple sequences. Since these alignments rely on structural information related to all the query sequences, the method can be only applied if all the structures being compared are known. Structure-based sequence alignments can be used as comparison points to evaluate alignments produced by purely sequence-based bioinformatics methods [98].

The output of a structure-based alignment is a superposition of the atomic coordinate sets corresponding to the minimal root mean square distance (*RMSD*) between the structures. The RMSD of two aligned structures indicates their divergence from one another. Multiple structural alignment can be complicated by the existence of multiple protein domains within one or more of the input structures. In this case an attempt to align whole proteins can artificially inflate the RMSD even if certain domains in these proteins are very similar.

MAMMOTH (MAtching Molecular Models Obtained from Theory) is a structure-based sequence alignment program used in this thesis. Benchmarks on targets of blind structure prediction (the CASP experiment) and automated GO annotation have shown it is tightly rank correlated with human curated annotation [99,100]. A highly complete database of MAMMOTH-based structural annotation for the predicted structures of unknown proteins covering 150 genomes (http://homepages.nyu.edu/~rb133/wcg/thread_8036.html) facilitates genomic scale normalization.

MAMMOTH-based structure alignment methods decompose the protein structure into short peptides (heptapeptides) which are compared with the heptapeptides of another protein. Similarity score between two heptapeptides is calculated using a unit-vector RMS (URMS) method, which uses comparison of C_α coordinates [101]. These scores are stored in a similarity matrix, and the optimal residue alignment is calculated with a hybrid (local-global) dynamic programming. Protein similarity scores calculated with MAMMOTH is derived from the likelihood of obtaining a given structural alignment by chance [99].

1.5.3 Quantum mechanics calculations

To obtain microscopic basic properties of relatively small molecules methods of quantum chemistry are used. These methods deal with the atoms as a complex system consisting of electrons and nuclei explicitly. Here, we describe the very basics of QM methods, which have been utilized in the section 3.1 of the thesis.

Schrödinger equation. For a system of atoms the Schrödinger equation describes its quantum state in time:

$$\mathbf{H}\psi = E\psi \qquad (1.5.1),$$

where **H** is the Hamiltonian operator, ψ is a wave function of the system and E is an eigenvalue of the operator **H**. The typical form of the Hamiltonian operator for QM calculations (non-relativistic, without external electric field etc.) takes into account five contributions to the total energy of a system (molecule): the kinetic energies of the electrons and nuclei, the attraction of the electrons to the nuclei, the interelectronic and internuclear repulsions:

$$H = -\sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_k \frac{\hbar^2}{2m_k} \nabla_k^2 - \sum_i \sum_k e^2 \frac{Z_k}{r_{ik}} + \sum_{i<j} \frac{e^2}{r_{ij}} + \sum_{k<l} \frac{e^2 Z_k Z_l}{r_{kl}} \qquad (1.5.2),$$

where i and j run over electrons, k and l run over nuclei, ħ is Planck's constant divided by 2π, m_e is the mass of the electron, m_k is the mass of nucleus k, ∇^2 is the Laplacian operator, e is the charge on the electron, Z is an atomic number, and r_{ab} is a distance between the particles a and b. In general, the equation 1.5.1 has many acceptable eigenfunctions ψ_i (perhaps infinite) for a given molecule, each characterized by a different associated eigenvalue E_i. Without loss of generality, the functions are chosen to be orthonormal:

$$\int\int\int \psi_i \psi_j \, dx\, dy\, dz = \delta_{ij} \qquad (1.5.3),$$

where the integration is taken over the whole phase space. And:

$$\int\int\int \psi_j H \psi_i \, dx\, dy\, dz = \langle \psi_i | H | \psi_j \rangle = E_i \delta_{ij} \qquad (1.5.4).$$

Born-Oppenheimer approximation, often accepted for the calculations, decouples the Hamiltonian into “slow” nucleus movement and electronic term, because the nuclei move much slower than electron. Even then, there is no analytic solution for the equation 1.5.1 even for such simple systems as the molecular hydrogen. That is why the most precise quantum chemistry methods dealing with wave functions give only approximate solutions for molecules (in this case solutions ψ_i are called *molecular orbitals*). Molecular orbitals are searched as *a linear combination of atomic orbitals (LCAO)*, which is a superposition of atomic orbitals. Since electron configurations of atoms are described as wave functions, these wave functions are the *basis set* of functions, which describes the electrons of a given molecule (system):

$$\phi = \sum_{i=1}^{N} a_i \varphi_i \qquad (1.5.5),$$

where the set of N functions φ_i is the basis set and each of basis functions is weighted by a coefficient a_i. So, in principal, one of problems in application of QM methods is the choice of basis set.

Hartree-Fock methods (HF) use assumption that electronic Hamiltonian could be decomposed into the one-electron Hamiltonians:

$$H = \sum_{i=1}^{N} h_i \qquad (1.5.6),$$

where h_i are related to one-electron Hamiltonians:

$$h_i = -\frac{1}{2} \nabla_i^2 - \sum_{k=1}^{M} \frac{Z_k}{r_{ik}} + V_i\{j\} \qquad (1.5.7),$$

where M is the number of nuclei and $V_i\{j\}$ is the part describing interelectronic interactions:

$$V_i\{j\} = \sum_{i \neq j} \int \frac{\rho_j}{r_{ij}} dr \qquad (1.5.8),$$

where ρ_j is the charge density associated with electron j. The repulsive third term in 1.5.7 is analogous

to the attractive second term, except that nuclei are treated as point charges, while electrons, being treated as wave functions, have their charge spread out, so an integration over all space is necessary. The solution for molecular orbital is searched in form of Slater determinant:

$$\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \cdots & \chi_N(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \cdots & \chi_N(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_1(\mathbf{x}_N) & \chi_2(\mathbf{x}_N) & \cdots & \chi_N(\mathbf{x}_N) \end{vmatrix} \quad (1.5.9),$$

where N is the total number of electrons and χ_i is a spin-orbital (spin space and cartesian space dependences are decomposed into two independent functions, the interaction between spin moment and electric charge is supposed to be zero). The solution could be also expressed by a linear combination of Slater determinants (physically they correspond to excited states of the molecule).

Thus, the methods are classified by:

1. Number of the electrons, taking into account in the calculations.
2. Number of Slater determinants (one Slater determinant is used only if the full spin of the system is 0).
3. Number of electronic states of each atom taken into account (excitation by spin variable change, for example).
4. If the coefficients a_i from 1.5.5 are known (direct variational methods) or not.

The functional $<\psi H \psi>$ is then minimized iteratively.

The *Hartree-Fock (HF) method* assumes that the exact, N-body wave function of the system can be approximated by a single Slater determinant (in the case where the particles are fermions) of N spin-orbitals. By invoking the variational principle, one can derive a set of N-coupled equations for the N spin-orbitals. Solution of these equations yields the HF wavefunction and energy of the system, which are approximations of the exact ones. In process of solving an equation resulted from a functional minimization, 4-bodies integrals appear, that means that complexity of calculations increases as m^4, where m is the number of taken into account electrons. Basis sets in HF-methods normally consist of Slater functions expressed by Gaussian functions, which are simpler to use for integration. The most critical limitation of HF-methods is that they do not take into account electron correlation (electrons search the space independently if other ones already occupy this part of the space). Normally this error is of 1% order of magnitude of the calculated energies. Roothaan method is used to improve HF-

method by use of perturbation. In this method the Hamiltonian is precisely expressed as a sum HF-Hamiltonian and the difference between precise Hamiltonian of the system and HF-Hamiltonian:

$$H=F+(H-F) \qquad (1.5.10),$$

where the last component is considered as perturbation. Such method is called Hartree-Fock-Roothaan method (HFR). The second of order of perturbation theory (MP2) is widely used to deal with electron correlation [102].

Semiempirical methods of quantum chemistry are based on the HF formalism, but make many approximations and obtain some parameters from empirical data. They are very important in computational chemistry for treating large molecules, where the full Hartree-Fock method without the approximations is too expensive. The use of empirical parameters appears to allow some inclusion of electron correlation effects into the methods. Though being computationally less expensive (calculations scale as m^3 instead of m^4), these methods do not necessarily sacrifice accuracy compared to HF methods.

One of the simplest semiempirical methods Complete Neglect of Differential Overlap (CNDO) could be obtained from HFR method by several assumptions and approximations:

1.Only the electrons at the atomic open shells are taken into account. The electrons at the closed shells are localized on the point nuclei.

2. Basis set corresponds to the occupied states of the atomic orbitals.

3. All integrals containing $\varphi_i^*$and φ_j are considered to be zero if $i \neq j$.

Some of the one-electron integrals are not calculated but correspond to empirical value obtained from an experiment. Other semiempirical methods could have different number of electrons and empirical parameters. Different semiempirical methods are constructed to reproduce distinct properties of molecules (e.g., ZINDO/1 is a method for calculating ground state properties such as bond lengths and bond angles and covers a wide range of the periodic table, including the rare earth elements).

Density Functional Theory (DFT) in contrast to *ab initio* HF or semiempirical methods does not deal with wave functions (which are mathematically formal and only their absolute square is physically relevant) but with physically observable charge densities. Within this theory, the properties of a many-electron system can be determined by using functionals, which in this case is the spatially dependent on electron density. DFT was put on a firm theoretical basis by the two Hohenberg-Kohn theorems (H-K)

[103]. The original H-K theorems held only for non-degenerate ground states in the absence of a magnetic field, although they have since been generalized to encompass these.

The first H-K theorem demonstrates that the ground state properties of a many-electron system are uniquely determined by an electron density that depends on only 3 spatial coordinates. It lays the groundwork for reducing the many-body problem of N electrons with 3N spatial coordinates to only 3 spatial coordinates, through the use of functionals of the electron density. This theorem can be extended to the time-dependent domain to develop time-dependent density functional theory (TDDFT), which can be used to describe excited states.

The second H-K theorem defines an energy functional for the system and proves that the correct ground state electron density minimizes this energy functional.

Within the framework of Kohn-Sham DFT, the intractable many-body problem of interacting electrons in a static external potential is reduced to a tractable problem of non-interacting electrons moving in an effective potential. The effective potential includes the external potential and the effects of the Coulomb interactions between the electrons, e.g., the exchange and correlation interactions. Modeling the latter two interactions becomes the difficulty within KS DFT. The simplest approximation is the local-density approximation (LDA), which is based on exact exchange energy for a uniform electron gas, which can be obtained from the Thomas-Fermi model, and from fits to the correlation energy for a uniform electron gas. Non-interacting systems are relatively easy to solve as the wave function, which can be represented as a Slater determinant of orbitals. Further, the kinetic energy functional of such a system is known exactly. The exchange-correlation part of the total-energy functional remains unknown and must be approximated. Distinct DFT methods differ in this part of equation, and different exchange-correlation parameters could be tuned by theoretical precalculations as well as by empirically obtained values. Computationally these methods are scale as m^3 expensive, which is less than for HF-methods but scaling also depends on the exchange-correlation part. However, they have the limitations related to treatment of the systems which cannot be described by use of single Slater determinant (excited states).

BSSE-correction. The *basis set superposition error (BSSE)* is a consequence of basis set truncation, i.e. of the unavoidable fact that finite basis sets have to be chosen. Upon a complex formation, a complex has more basis functions employed in the calculations than in either of monomers. That means that each monomer is able to utilize, at least in part, the basis functions of its interaction partners. This is not the case when a monomer is treated alone, e.g. when a binding or

interaction energy is calculated. Therefore, the energy of the whole system is computed lower in comparison to the separated subsystems which do not benefit from the basis functions of their interaction partners.

In the counterpoise correction the basis set for subsystems contain also the basis functions of the whole molecule. In the uncorrected calculation of a dimer *AB*, the dimer basis set is the union of the two monomer basis sets. The uncorrected interaction energy is

$$V_{AB}(\mathrm{G}) = E_{AB}(\mathrm{G,AB}) - E_A(\mathrm{A}) - E_B(\mathrm{B}) \qquad (1.5.11),$$

where G denotes the coordinates that specify the geometry of the dimer and E_{AB}(G,AB) the total energy of the dimer *AB* calculated with the full basis set AB of the dimer at that geometry. Similarly, E_A(A) and E_B(B) denote the total energies of the monomers *A* and *B*, each calculated with the appropriate monomer basis sets A and B, respectively. This is the procedure for calculating an interaction energy without BSSE correction.

The *counterpoise corrected interaction energy* [104] is:

$$V_{AB}^{\mathrm{cc}}(\mathbf{G}) = E_{AB}(\mathrm{G,AB}) - E_A(\mathrm{G,AB}) - E_B(\mathrm{G,AB}) \qquad (1.5.12)$$

where E_A(G,AB) and E_B(G,AB) denote the total energies of monomers *A* and *B*, respectively, computed with the dimer basis set AB at geometry G, i.e. in the calculation of monomer *A* the basis set of the other monomer *B* is present at the same location as in dimer *AB*, but the nuclei and electrons of *B* are not. In this way, a basis set for each monomer is extended by the functions of the other monomer. Important to note that the counterpoise correction provides only an estimate of the BSSE since the monomer basis set is enhanced not only by empty orbitals, but also by orbitals occupied by electrons of the other monomer molecule.

1.5.4 Molecular dynamics and related methods

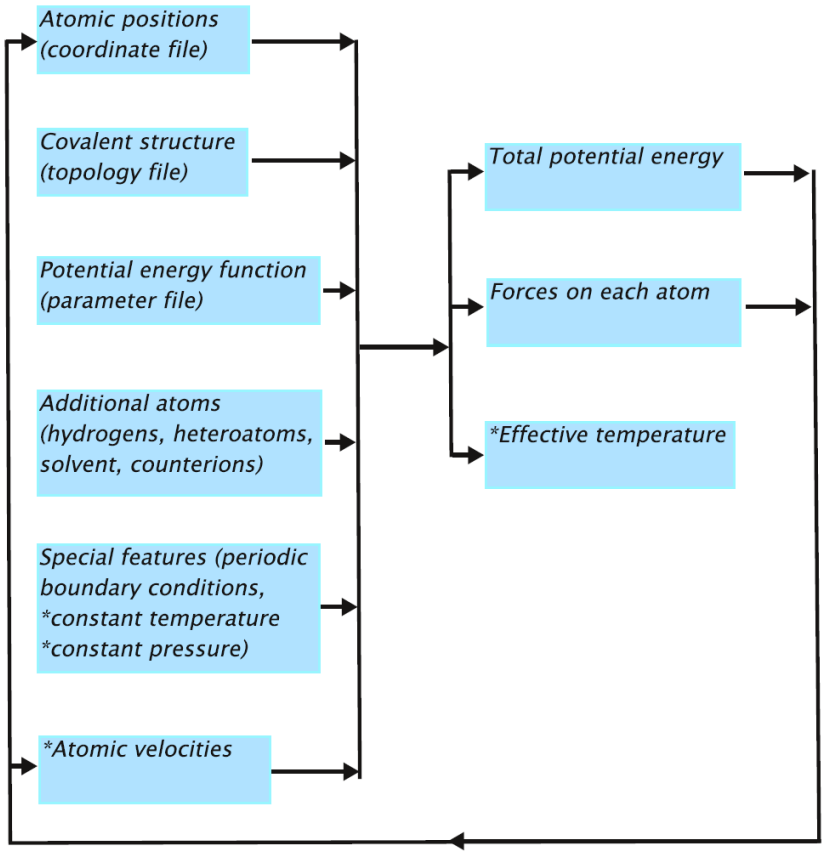

Figure 1.5.3. Schematic flow chart of algorithms for energy minimization and MD. Features which apply only to molecular dynamics are indicated with asterix. Each cycle of energy minimization represents a step in conformation space, while each cycle of molecular dynamics represents a step in time [7].

- *Molecular dynamics (MD)* is a form of computer simulation that solves Newton's equations of motion for a system of N interacting atoms:

$$m_i \frac{\delta^2 \vec{r}_i}{\delta t^2} = \vec{F}_i \qquad (1.5.12),$$

where forces are:

$$\vec{F}_i = \frac{-\delta V}{\delta \vec{r}_i} \qquad (1.5.13),$$

where V is the potential.

The equations are solved simultaneously in small time steps, so that the temperature and

pressure remain at the required values, and the coordinates are written to an output file at regular intervals. The coordinates as a function of time represent a *trajectory of the system*. After initial changes, the system usually reaches an equilibrium state. By averaging over an equilibrium trajectory many macroscopic properties can be extracted from the output file. The pipeline typical for energy minimization and MD in common is represented in the Figure 1.5.3.

Further, we give some definitions and details on the stages and aspects important for MD runs.

- *Forcefield* refers to the functional form and parameter sets used to describe the potential energy of a system of particles (typically but not necessarily atoms). Force field functions and parameter sets are derived from both experimental work and high-level QM calculations. *All-atom forcefields* provide parameters for every atom in a system, including hydrogens, while *united-atom forcefields* could treat the hydrogen and carbon atoms in methyl and methylene groups, for example, as a single interaction center. *Coarse-grained forcefields*, which are frequently used in long-time simulations of proteins, provide even more abstracted representations for increased computational efficiency. More detailed description of the physical nature behind potential included in the force fields is given in the subsection 1.2.1. In comparison to analytical expression for the potentials mentioned before, bonded interactions are explicitly decomposed into bond, torsion angle and dihedral angle potentials:

$$V_{bonded} = V_{bond} + V_{angle} + V_{dihedral} \tag{1.5.14}$$

Electrostatic interaction usually contains explicitly only the first component of the equation 1.2.2, which describes point-charge interactions, unless a force-field is polarizable and, therefore, has a different expression for the electrostatic potential. So, in general, the potential has the following form:

$$V(\vec{r}) = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}(1 + \cos[n\phi - \gamma]) + \sum_{i<j}^{atoms} \left(\frac{A_{ij}}{r_{ij}^2} - \frac{B_{ij}}{r_{ij}^2}\right) + \sum_{i<j}^{atoms} \frac{q_i q_j}{\epsilon R_{ij}} \tag{1.5.15}$$

and polarization component is usually expressed as:

$$E_{pol} = -\frac{1}{2}\sum_{i}^{atom} \vec{\mu}_i \vec{E}_i^{(0)} \qquad (1.5.16)$$

where μ_i is an induced atomic dipole and $E_i^{(0)}$ is an initial electric field causing this polarization. In addition, charges that are not centered on atoms, but are off-center (as for lone-pairs) that can be included in the forcefield.

- *Energy minimization* methods are common techniques to compute an equilibrium configuration of molecules, which should be a stable state and correspond to a local minimum of their potential energy. This kind of calculations generally start from an arbitrary state of molecules, then the mathematical procedure of optimization allows to move atoms (to vary coordinates) in a way to reduce the net forces (the gradients of potential energy) to nearly zero (or defined cut-off value). From a computational viewpoint, the problem of minimizing the energy of a model macromolecular system falls into the general area of nonlinear optimization problems. The functional V($\vec{r}$) should be minimized in the multidimensional space $\vec{r}$. In case of a model with N atoms, there are 3N-dimensional space (since each atom has 3 cartesian coordinates), and V is the potential energy of the system. There are two most popular algorithms used for the minimization: *steepest descent* and *conjugate gradient.*

- *Steepest descent* is an first-order optimization algorithm. To find a local minimum of a function using steepest descent, one takes steps proportional to the negative of the gradient (or the approximate gradient) of the function at the current point. For each next iteration:

$$\vec{r}_k = \vec{r}_{k-1} + \lambda_k \frac{\vec{F}}{F} \qquad (1.5.17),$$

where λ_k is a positive coefficient. Since the vector $\frac{\vec{F}}{F}$ is parallel to the negative gradient of the energy, it points straight downhill. The weaknesses of steepest descent are:

1. The algorithm can take many iterations to converge towards a local minimum, if the curvature in different directions is very different.
2. Finding the optimal λ per step can be a time-consuming task. Conversely, using a fixed λ can yield poor results.

- Because of these limitations, steepest descent is usually followed by *conjugate gradient* in the

minimization procedure. This technique combines information on the current gradient with that based on the gradient at previous steps. Iteratively:

$$\vec{r}_k = \vec{r_{k-1}} + \lambda_k (\vec{F}_k + \frac{F_k^2}{F_{k-1}^2} \frac{\vec{F_{k-1}}}{F_{k-1}}) \tag{1.5.18}$$

It can be proven that for a quadratic surface, the search direction specified by the last equation passes through minimum on the N^{th} step for an N-dimensional surface, as long as the minimum along each successive search direction is found. Even if the step size is not optimal, the conjugate gradient method still yields a search direction that is superior to that of steepest descent.

- *Verlet integration* [105] is a numerical method frequently used to integrate Newton's equations of motion. The Verlet integrator offers greater stability than the much simpler Euler method, as well as other properties that are important in physical systems such as time-reversibility. Stability of the technique depends fairly heavily upon either a uniform update rate, or the ability to accurately identify positions at a small time intervals into the past.

The idea for Verlet algorithm comes from the third-order Taylor expansions for the positions $\vec{r}(t)$ of atom:

$$\vec{r}(t+\delta t) = \vec{r}(t) + \vec{v}(t)\delta t + \vec{a}(t)\frac{\delta t^2}{2} + \vec{b}(t)\frac{\delta t^3}{6} + O(\delta t^4)$$

$$\vec{r}(t-\delta t) = \vec{r}(t) - \vec{v}(t)\delta t + \vec{a}(t)\frac{\delta t^2}{2} - \vec{b}(t)\frac{\delta t^3}{6} + O(\delta t^4) \tag{1.5.19}.$$

Therefore:

$$\vec{r}(t+\delta t) = 2\vec{r}(t) - \vec{r}(t-\delta t) + \vec{a}(t)\delta t^2 + O(\delta t^4) \tag{1.5.20}.$$

Considering that acceleration is expressed as:

$$\vec{a}(t) = \frac{1}{m}\vec{F}(\vec{r}(t)) \tag{1.5.21},$$

an atomic position is:

$$\vec{r}(t+\delta t)=2\vec{r}(t)-\vec{r}(t-\delta t)+\frac{1}{m}\vec{F}(\vec{r}(t))\delta t^{2}+O(\delta t^{4}) \qquad (1.5.22)$$

So, the truncation error of the algorithm, when evolving the system by δt, is of the order of δt^4, even if third derivatives do not appear explicitly. This algorithm is simple to implement, accurate and stable, explaining its large popularity among molecular dynamics simulators. For velocities:

$$\vec{v}(t)=\frac{\vec{r}(t+\delta t)-r(t+\delta t)}{2\delta t} \qquad (1.5.23).$$

However, the error associated with this expression is of order of δt^2.

If in Taylor expansions δt is rewritten as 1/2δt+1/2 δt, the final equations for integration could be obtained for propagating the position and velocities in a coupled fashion:

$$\vec{r}(t+\delta t)=\vec{r}(t)+\vec{v}(t+\frac{1}{2}\delta t)\delta t$$

$$\vec{v}(t+\frac{1}{2}\delta t)=\vec{v}(t-\frac{1}{2}\delta t)+\vec{a}(t)\delta t \qquad (1.5.24).$$

In this algorithm, position depends on velocities as computed one-half time step out of phase, thus, scaling of the velocities can be accomplished to control the temperature. Forces, however, are computed at integral time steps, half-time-step-forward velocities are computed therefrom, and used to update particle positions. The possible source of algorithm's instability is ignoring the third derivatives in Taylor expansions.

- *Periodic boundary conditions (PBC)* are a set of boundary conditions that are often used to simulate a large system by modeling a small part that is far from its edge. A unit cell of a certain geometry is defined, and when an object passes through one face of the unit cell, it reappears on the opposite face with the same velocity. The simulation is of an infinite perfect tiling of the system. The

tiled copies of the unit cell are called images, of which there are infinitely many. During the simulation, only the properties of the unit cell need to be recorded and propagated. The minimum-image convention is a common form of PBC particle bookkeeping in which each individual particle in the simulation interacts with the closest image of the remaining particles in the system.

In MD PBC are usually applied to simulate bulk gases, liquids, crystals or mixtures. A common application is to use PBC to simulate solvated macromolecules in a bath of explicit solvent. Since MD contains electrostatic interactions, the net electrostatic charge of the system must be zero to avoid summing to an infinite charge when PBC is applied. However, there are still some artifacts originated from the correlations between unit cells and artificial interactions between “heads” and “tails” of different unit cells.

PBC requires the unit cell to be a shape that tiles perfectly into a three-dimensional crystal. Thus, a spherical or elliptical droplet cannot be used. A cube or rectangular prism is the most intuitive and common choice, but can be computationally expensive due to unnecessary amounts of solvent molecules in the corners, distant from the central macromolecules. A common alternative that requires less volume is the truncated octahedron.

- *Particle Mesh Ewald (PME)* method is utilized for electrostatic calculations in PBC. PME uses Ewald summation, which is a special case of the Poisson summation formula, replacing the summation of interaction energies in real space with an equivalent summation in Fourier space. The advantage of this approach is the rapid convergence of the Fourier-space summation compared to its real-space equivalent when the real-space interactions are long-ranged. Because electrostatic energies consist of both short- and long-range interactions, it is maximally efficient to decompose the interaction potential into a short-range component summed in real space and a long-range component summed in Fourier space [106]. In practice, the *cut-off* for PME is defined by a researcher, who uses MD, and depends on the size of a unit in PBC. The cut-off should be less or equal to the half of unit's dimension, to assure the convergence of electrostatic summation in the direct space.

- *Temperature coupling* is a technique used for maintenance of constant temperature in PBC in NTP or NTV microcanonical ensembles. There are several ways to carry out the temperature couplings. In the *weak-coupling* algorithm a single scaling factor is used for all atoms [107]. This algorithm ensures that the total kinetic energy is appropriate for the desired temperature but does not control that the temperature is even over all parts of the molecule. Atomic collisions tend to ensure an even temperature distribution, but in reality this is not guaranteed, and there are many subtle problems that

can arise with weak temperature coupling [108]. *Andersen temperature coupling* scheme [109] implies imaginary collisions, which randomize the velocities to a distribution corresponding to the fixed constant temperature. The dynamics process is Newtonian. Hence, time correlation functions can be computed and the results averaged over an initial canonical distribution. Too high collision rate slows down the speed at which the molecules explore configuration space, whereas too low rate means that the canonical distribution of energies is sampled slowly [110]. Use of *Langevin dynamics* is an another approach for temperature coupling. There is a collision frequency parameter in this algorithm to be defined and a simple leapfrog integrator (a variant of Verlet integration) is used to propagate the dynamics, with the kinetic energy adjusted to be correct for the harmonic oscillator case [111,112]. A collision frequency parameter is not necessary equal to the physical collision frequency. In fact, it is often advantageous, in terms of sampling or stability of integration, to use much smaller values for this parameters than the physically relevant ones.

- *Pressure coupling* adjusts the volume of the unit cell (gradually on each step) to make the computed pressure approach the target pressure. Equilibration with NTP microcanonical ensemble is generally necessary to adjust the density of the system to appropriate values. Pressure coupling algorithms are often analogous to weak temperature coupling [107].

- *Constraints* in MD are used to restrict movements of a group of atoms. There are several ways of the most popular constraints: *freezing* and putting *harmonic restraints*. In the first case positions of the constrained atoms are absolute rigidly fixed, they do not move in a MD simulation and all equations of motion are rewritten with consideration that masses, for example, of these atoms are infinite. When harmonic restrains are applied, all bonds of the constrained atoms receive additional rigidity by increasing spring constants for these bonds. In this case the equations of motion look the same as in unconstrained dynamics but with changed coefficients. Important to remember, that application of these constraints does not speed up calculations since atoms are still described by the same equations of motion. Another type of constraints include *positional and rotational constraints*, when an an additional potential is acting upon a group of atoms to bias them to be at certain distance of to keep a certain angle with a reference point. Mathematical form of this kind of potential could be very different.

- *SHAKE algorithm* is used to perform bond length constraints [113]. It is normally utilized for hydrogens in MD simulations. The size of the MD time step is determined by the fastest motions in the system. SHAKE removes the bond stretching freedom, which is the fastest motion, and consequently

allows a larger time step to be used, resulting in speeding up the calculations. For water models, a special "three-point" algorithm is used [114]. Since SHAKE is an algorithm based on dynamics, the minimizer is not aware of what SHAKE is doing; for this reason, minimizations generally should be carried out without SHAKE.

- *Counterions* are used in MD with PBC to make a charge of a unit neutral to avoid problems with electrostatics. For counterions Na+, K+ and Cl- are usually used. A study on counterions impact to MD results in AMBER shows that the simulations of solvated proteins are moderately sensitive to the presence of counterions. However, this sensitivity is highly dependent on the starting model and different procedures of equilibration used. The neutralized systems tend to evince smaller root mean square deviations regardless of the system investigated and the simulation procedure used. The results of parameterized fitting of the simulated structures to the crystallographic data, giving quantitative measure of the total charge influence on the stability of various elements of the secondary structure, revealed a clear scatter of different reactions of various systems' secondary structures to counterions addition: some systems apparently were stabilized when neutralized, while the others were not. Thus, one cannot unequivocally state, despite consideration of specific simulation conditions, whether protein secondary structures are more stable when they have neutralized charges. This suggests that caution should be taken when claiming the stabilizing effect of counterions in simulations involving small, unstable polypeptides or highly charged proteins [115].

- *MM-PBSA/MM-GBSA* (Molecular Mechanics-Poisson-Bolzmann Surface Area/Molecular Mechanics-Generalized Born Surface Area) approach represents the postprocessing method to evaluate free energies of binding or to calculate absolute free energies of molecules in solution. The MM_PBSA/GBSA method combines molecular mechanical energies with the continuum solvent approaches. Often, the key quantity that needs to be computed is the total free energy of the molecule in the presence of solvent, which could be written as:

$$E_{tot} = E_{vac} + \delta G_{solv} \qquad (1.5.25),$$

where E_{vac} represents a molecule's energy in vacuum (gas-phase), and δG_{solv} is the free energy of transferring the molecule from vacuum into solvent, i.e. solvation free energy. Usually it is assumed

that E_{vac} is given by a classical potential function, or force-field, that breaks the interaction down into various physical components, such as bond and angle stretching, torsional twist, and VDW and Coulomb interactions between its atoms (see above in this subsection).

To estimate the total solvation free energy of a molecule δG_{solv} one typically assumes that it can be decomposed into the electrostatic and non-electrostatic components:

$$\delta G_{solv} = \delta G_{el} + \delta G_{nonel} \quad (1.5.26),$$

where δG_{nonel} is the free energy of solvating a molecule from which all charges have been removed (i.e. partial charges of every atom are set to zero), and δG_{el} is the free energy of first removing all charges in the vacuum, and then adding them back in the presence of a continuum solvent environment. The above decomposition, which is yet another approximation, is the basis of the widely used MM-PBSA scheme [116]. Generally speaking, δG_{nonel} comes from the combined effect of two types of interaction: the favorable van der Waals attraction between the solute and solvent molecules, and the unfavorable cost of breaking the structure of the solvent around the solute. δG_{nonel} is described in terms of solvent *accessible surface area* (ASA), the surface area of a biomolecule that is accessible to a solvent. The ASA is usually measured in $Å^2$. ASA is typically calculated using the 'rolling ball' algorithm [117], which uses a sphere (of solvent) of a particular radius to probe the surface of the molecule. The choice of the probe radius does have an effect on the observed surface area, as using a smaller probe radius detects more surface details and, therefore, reports a larger surface. A typical value is 1.4 Å (also used as a default value in a popular NACCESS program [118]), which approximates the radius of a water molecule. Another factor that affects the results is the definition of the VDW radii of the atoms in the studied molecule. For example, the molecule may often lack hydrogen atoms which are implicit in the structure. The hydrogen atoms may be implicitly included in the atomic radii of the heavy atoms, with a measure called the group radii.

The ASA is closely related to the concept of the solvent-excluded surface (also known as the molecular surface or Connolly surface), which is imagined as a cavity in bulk solvent (effectively the inverse of the solvent-accessible surface). In practice, it is also calculated via a rolling-ball algorithm [119] and independently implemented three-dimensionally in two studies works [120,121]. Connolly spent several more years perfecting the method [122] and it is thus sometimes called the Connolly surface.

The ASA can be used in protein interfaces characterization (difference of the ASA of complex and unbound components gives the size of an interface of the complex) and for empirical estimation of hydration energy, which is considered to be proportional to ASA. Within the PBSA, δG_{nonel} is supposed to be proportional to the total solvent ASA of the molecule, with a constant derived from experimental solvation energies of small non-polar molecules:

$$\delta G_{nonel} \sim ASA \qquad (1.5.27),$$

which is an approximation, but arguably not the most critical one in the hierarchy of assumptions that form the foundation of the implicit solvent methodology [123].

In a model of continuous solvent the remained component δG_{el} can be easily calculated if the potential distribution $\varphi(r)$ in space is known. This distribution is described by Poisson-Bolzmann equation for the charge density $\rho(r)$ distribution in a dealectric with constant $\varepsilon(r)$:

$$\nabla\varepsilon(r)\nabla\varphi(r) = -4\pi\rho(r) + \kappa^2\varepsilon(r)\varphi(r) \qquad (1.5.28),$$

where κ is Debye-Huckel parameter. However, in molecular dynamics applications, the associated computational costs are often very high, as the *Poisson-Boltzmann* equation needs to be solved every time the conformation of the molecule changes. The *Generalized Born* (GB) model is an approximation to the exact (linearized) Poisson-Boltzmann equation. The GB approach is computationally more effective than PB approach and it is an approximation to the exact Poisson-Boltzmann equation. It is based on modeling a protein as a volume whose internal dielectric constant differs from the external solvent. The model has the following functional form:

$$G_{el,GB} = \frac{1}{8\pi}\left(\frac{1}{\epsilon_0} - \frac{1}{\epsilon}\right)\sum_{i<j}^{N}\frac{q_i q_j}{f_{GB}} \qquad (1.5.29),$$

where:

$$f_{GB} = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}} \qquad (1.5.30)$$

and:

$$D=(\frac{r_{ij}}{2a_{ij}})^2, a_{ij}=\sqrt{a_i a_j} \quad (1.5.31)$$

where, ε_0 is the dielectric constant *in vacuo*, q_i is the electrostatic charge on particle i, r_{ij} is the distance between particles i and j, and a_i is a quantity (with the dimension of length) known as the effective Born radius [124]. The effective Born radius of an atom characterizes its degree of burial inside the solute; qualitatively it can be thought of as the distance from the atom to the molecular surface. Accurate estimation of the effective Born radii is critical for the GB model [125].

Often, the challenge in these calculations is to extract frames from trajectory to essentially sample the conformational space. Depending on a system, there is a different number of randomly chosen frames to be enough for this purpose. Averaged structure of a system is usually not used since even for two equally probable rotamer states of a side chain its average would not correspond to the physically relevant state of the system. The most important impact of conformational changes is on electrostatic energy component, that is why at the first step, statistical sampling could be carried out only for less computationally expensive electrostatic component (for the details see Methodology in section 2.1) [126].

For calculations of entropies in MM-PBSA method *normal mode analysis* is used. Nevertheless, entropy components of free energies are still the least accurate in MM-PBSA energy calculations [81].

- *Free energy perturbation (FEP)* is a method based on statistical mechanics that is used for computing free energy differences from MD or Metropolis Monte Carlo simulations [127]. According to free-energy perturbation theory, the free energy difference for going from state A to state B is obtained from the following equation, known as the Zwanzig equation:

$$\delta G = G_A - G_B = -k_b T \ln[\exp(\frac{-E_B - E_A}{k_b T})]_A \quad (1.5.32),$$

where T is the temperature, k_B is Boltzmann's constant, and the quadratic brackets denote an average over a simulation run for state A. In practice, one runs a normal simulation for state A, but each time a new configuration is accepted, the energy for state B is also computed. The difference between states A and B may be in the atom types involved, in which case the δG obtained is for "mutating" one molecule into another, or it may be a difference of geometry, in which case one obtains a free energy map along one or more reaction coordinates. These reaction coordinates could be related, for example,

to the charge of perturbed atoms or to their van der Waals radii. Mathematically, free energy could be expressed by thermodynamical integration approach as:

$$\delta G = G(\lambda = 1) - G(\lambda = 0) = \int_0^1 \left(\frac{\delta V}{\delta \lambda}\right)_\lambda d\lambda \qquad (1.5.33),$$

where V is a potential function and λ is a reaction coordinate.

For numerical integration Gaussian integration procedure could be used:

$$\delta G = \sum_{i=1}^{n} w_i \left(\frac{\delta V}{\delta \lambda}\right)_{\lambda_i} \qquad (1.5.34),$$

where w_i are weights corresponding to the number of quadrature points (n).

The main limitation of this method is the difficulty in convergence for δG for large perturbed groups of atoms. Normally, the method is applied for perturbation of single atoms of geometrically similar chemical groups or for "eliminating" atoms.

In this thesis FEP has been utilized to calculate free energy impact of water molecules in protein interfaces (section 2.1) and for the study of fluoromethylated groups, which were compared to their non-fluorinated analogues (section 3.1). One can find more computational details on the application of FEP in these studies in the corresponding methodology related subsections.

- *Charge derivation (ESP/RESP procedure)* for a new residue/molecule, not yet parametrized for a given forcefield, is an important step in MD. To derive such atom-centered charges three steps need to be followed:

1. The molecule studied is optimized to determine a stable minimum (using a QM approach).
2. Then, this minimized structure is used to calculate a Molecular Electrostatic Potential (MEP) on a 3D grid (also using QM).
3. This grid is exported into the RESP program which is used to fit atom-centered charges to the MEP [128].

1.5.5 Molecular docking

In the field of molecular modeling, docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex [129]. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules. Hence docking plays an important role in the rational design of drugs [130]. Given the biological and pharmaceutical significance of molecular docking, considerable efforts have been directed towards improving the methods used to predict docking. Molecular docking can be thought of as a problem of "*lock-and-key"*, where one is interested in finding the correct relative orientation of the "*key"* which will open up the "*lock"* .

Two principal approaches are particularly popular within the molecular docking community. The first approach uses a matching technique that describes a protein and a ligand as complementary surfaces [131,132]. Complementarity descriptors are usually related to molecular surface area of receptor and ligand characterized by its hydrophobic and hydrophilic properties. Shape complementarity based approaches are typically fast and robust, but they cannot usually model the movements or dynamic changes in ligand/protein conformations accurately, though recent developments allow these methods to investigate ligand flexibility. Shape complementarity methods can quickly scan through several thousand ligands in a matter of seconds and actually figure out whether they can bind at the protein's active site, and are usually scalable to even protein-protein interactions. The second approach simulates the actual docking process in which the ligand-protein pairwise interaction energies are calculated [133]. In this approach, a protein and a ligand are separated by a certain physical distance, and the ligand finds its position into the protein's active site after a certain number of 'moves' in its conformational space. This motions incorporate rigid body transformations such as translations and rotations, as well as internal changes in ligand's structure including torsion angle rotations. Each of these moves in the conformation space of the ligand induces a total energetic change of the system, and hence after every move the total energy of the system is calculated. The obvious advantage of the method is that it is more amenable to incorporate ligand flexibility into its modeling whereas shape complementarity techniques have to use some ingenious methods to incorporate flexibility in ligands. Another advantage is that the process is physically closer to what happens in reality, when a protein and a ligand approach each other after molecular recognition. An obvious disadvantage of this technique is that it takes longer time to evaluate the optimal pose of binding since they have to explore a rather large energy landscape. However grid-based techniques as

well as fast optimization methods have significantly addressed these problems.

To accomplish a docking search in conformational space many different search algorithms (MD, linear combinations of multiple structures to emulate flexibility of receptor/ligand, genetic algorithms) and scoring functions (molecular mechanics force fields, Generalized Born, Poisson-Bolzmann etc.) could be applied.

Autodock, the docking program used in this study (section 3.3), has a complementary surface based approach implemented with Lamarckian genetic algorithm, in which environmental adaptations of an individual's phenotype are reverse transcribed into its genotype and become heritable traits. The implemented algorithm is characterized by *i)* efficiency of search in terms of lowest energy reached within a given number of energy evaluations; *ii)* reliability in terms of reproducibility of finding the lowest energy structure in independent docking simulations, as measured by the number of conformations in the top ranked cluster; *iii)* success in terms of reproducing the known crystal structure [132].

Despite many promising results achieved by docking, there are still many limitations (especially for protein-protein docking) for this technique because of the problem of trade-off between conformational space exhaustive search and accuracy of scoring function.

CHAPTER 2

2.1 A molecular dynamics approach to study the importance of solvent in protein interactions

by Sergey Samsonov, Joan Teyra, and M. Teresa Pisabarro

Proteins: Structure, Function, and Bioinformatics. 2008 Nov 1;73(2):515-525.

2.1.1 Abstract

Water constitutes the cellular environment for biomolecules to interact. Solvent is important for protein folding and stability, and it is also known to actively participate in many catalytic processes in the cell. However, solvent is often ignored in molecular recognition and not taken into account in protein-protein interaction studies and rational design. Previously we developed SCOWLP, a database and its web application (http://www.scowlp.org), to perform studies on the contribution of solvent to protein interface definition in all protein complexes of the PDB. We introduced the concept of *wet spots*, interfacial residues interacting only through one water molecule, which were shown to considerably enrich protein interface descriptions. Analysis of interfacial solvent in a non-redundant dataset of protein complexes suggested the importance of including interfacial water molecules in protein interaction studies. In this work we use a molecular dynamics approach to gain deeper insights into solvent contribution to protein interfaces. We characterize the dynamic and energetic properties of water-mediated protein interactions by comparing different interfacial interaction types (direct, dual and wet spot) at residue and solvent level. For this purpose, we perform an analysis of 17 representative complexes from 2 protein families of different interface nature. Energetically wet spots are quantitatively comparable to other residues in interfaces, and their mobility is shown to be lower than protein surface residues. The residence time of water molecules in wet spots sites is higher than of those on the surface of the protein. In terms of free energy, though wet-spots-forming water molecules are very heterogeneous, their contribution to the free energy of complex formation is considerable. We find that water molecules can play an important role in interaction conservation in protein interfaces by allowing sequence variability in the corresponding binding partner, and we discuss the important implications of our observations related to the use of the correlated mutations concept in protein interactions studies. The results obtained in this work help to deepen our understanding of the physico-chemical nature underlying protein-protein interactions and strengthen the idea of using the wet spots concept to qualitatively improve the accuracy of folding, docking and rational design algorithms.

2.1.2 Introduction

Water plays an extremely important role in all biological processes because of its unique physical and chemical properties. It represents not only an environment for interacting components of biochemical reactions but it is also an active participant. Without a critical level of hydration proteins are not functional, and water molecules presence is crucial in catalytic sites of many enzymes [32]. It has been shown that water molecules can be structurally conserved in protein complexes, and that their residence time and diffusion characteristics are distinct from bulk and surface solvent [39,37,134]. Thermodynamically, water molecules can contribute favorably to protein complex formation [135,136]. Computationally, the inclusion of water in the Hamiltonian of protein systems has improved folding predictions compared to *in vacuo* folding models [48]. It is widely accepted that exclusion of water molecules from the proteins contact area in the process of complex formation is associated with a free energy decrease. The entropic component increase due to the transfer of water molecules into bulk solvent is considered to have the decisive energetic impact. This entropy gain is usually thought to exceed the corresponding enthalpy loss [137]. Despite all, solvent is often ignored in the analysis of protein-protein interactions.

The protein interface concept is very important in the description of protein-protein interactions. Protein interfaces could be defined differently depending on the criteria introduced for the cut-off for interacting atoms of the complex counterparts [138]. In our previous work we have developed SCOWLP, which, taking into account interfacial solvent, classifies all interfacial protein residues of the PDB into three classes based on their interacting properties: *dry* (direct interaction), *dual* (direct and water-mediated interactions), and *wet spots* (residues interacting only through one water molecule) [27]. Also in our preceding studies, statistical analysis of a non-redundant protein structure dataset showed that 40.1% of the interfacial residues participate in water-mediated interactions, and that 14.5% of the total residues in interfaces are wet spots. Moreover, wet spots have been shown to display similar characteristics to residues contacting water molecules in cores or cavities of proteins [64]. This suggests that water-mediated interactions should not be disregarded in a detailed definition of protein interfaces. Our and other studies have revealed that water-mediated interactions are highly heterogeneous [64,15,139], making protein-protein interactions studies challenging. In fact, certain aspects of water-mediated interactions still remain unclear.

This study aims to gain insights into solvent contribution to protein interactions. Our focus is the contribution of wet spots to protein interfaces in comparison to other interfacial residues. For this

purpose, we use a molecular dynamics (MD) approach to characterize dynamic and energetic properties of wet spots and the water molecules forming them. We pay special attention to the role of water molecules in interaction conservation in protein interfaces.

For our studies we use representatives of two protein families of different physico-chemical interface nature and interface size in complex with other proteins and peptides: Src-homology 3 domains (SH3) and immunoglobulin domains (Ig). SH3 are small recognition domains comprising 5 antiparallel β-strands and widely presented in signaling pathways [140]. Immunoglobulin domains are bigger and comprise 7 to 9 antiparallel β-strands [141]. Ig interfaces are more hydrophilic than SH3 domain interfaces.

The results of this study show that water-mediated interactions are similar in both protein families, and that the dynamic and energetic characteristics of wet spots and dual residues are comparable to dry interfacial residues. Interfacial water molecules display properties such as residence time and mobility more similar to water molecules in cores and cavities of proteins than to bulk or protein surface solvent. Our findings emphasize the important role of water contributing to interaction conservation of protein interfaces and strongly imply the significance of including interfacial solvent in detailed protein interface descriptions.

2.1.3 Methodology

Protein complexes dataset (Table 2.1.1). The following criteria for protein complex selection was applied: resolution ≤ 2.5 Å and existence of wet spots (as they are annotated in SCOWLP [27]). The complex 1avz (FYN SH3 domain with HIV1-Nef) was also taken in our dataset because of its biological relevance. Structural alignments of the proteins were done with the MAMMOTH algorithm [142].

Molecular dynamics simulations. MD simulations were carried out with the AMBER 8.0 package. All hydrogen atoms were added using the Xleap tool. Standard ff03 force field parameters were used. Parameters for phosphorylated residues (complexes 1opk, 2src, 1qcf) were taken from the phosphorylated amino acids library set [143], and parameters for the N-substituted glycine residue (complex 1b07) were derived in the Antechamber module of AMBER 8.0. Each complex was solvated in a truncated octahedron periodic box filled with TIP3P water molecules and neutralized by counterions. MD simulations were preceded by two energy-minimization steps: 500 cycles of steepest descent and 1000 cycles of conjugate gradient with harmonic force restraints on protein atoms, then

1000 cycles of steepest descent and 1500 cycles of conjugate gradient without constraints. This was followed by heating of the system from 0 to 300K for 10 ps, and a 30 ps MD equilibration run at 300K and 10^6 Pa in isothermal isobaric ensemble (NPT). Following the equilibration procedure, 10 ns of productive MD runs were carried out in periodic boundary conditions in NPT ensemble with Langevin temperature coupling with collision frequency parameter γ=1 ps^{-1} and Berendsen pressure coupling with a time constant of 1.0 ps. The SHAKE algorithm was used to constrain all bonds that contain hydrogen atoms. A 2 fs time integration step was used. An 8 Å cutoff was applied to treat non-bonded interactions, and the Particle Mesh Ewald (PME) method was introduced for long-range electrostatic interactions treatment. MD trajectories were recorded each 2 ps. For the analysis of the trajectories PTRAJ module was used.

Table 2.1.1. Complexes dataset

PDB ID	SCOP family	Resolution Å	PP/Pp	Short description
1ujo*	SH3	1.70	Pp	Signal transducing adaptor molecule-2(STAM-2) SH3 domain with peptide derived from deubiquitinating enzyme (UBPY)
1bbz	SH3	1.65	Pp	Abl Tyrosine Kinase SH3 domain with p40 synthetic peptide
1fyn	SH3	2.30	Pp	Fyn Tyrosine Kinase SH3 domain with 3BP-2 synthetic peptide
1oeb	SH3	1.76	Pp	Grb2-like adaptor protein Mona/Gads SH3 domain with the peptide derived from T-cell receptor signal transducer SLP-76
1b07	SH3	2.50	Pp	Proto-oncogene Crk2 (Serine-Threonine kinase) SH3 domain with peptoid inhibitor
1uti	SH3	1.50	Pp	Grb2-like adaptor protein Mona/Gads with the peptide derived from Hematopoietic Progenitor Kinase 1 (Hpk1)
1abo	SH3	2.00	Pp	Abl Tyrosine Kinase SH3 domain with 3BP-1 synthetic peptide
1opk	SH3	1.80	PP	Mouse Abl Tyrosine Kinase (SH3-PI)
2src	SH3	1.50	PP	Human Src Tyrosine Kinase (SH3-PI)
1avz*	SH3	3.00	PP	Fyn Tyrosine Kinase SH3 domain with HIV1-Nef protein
1qcf	SH3	2.00	PP	Human Hck Tyrosine Kinase (SH3-PI)
1sm3	Ig	1.95	Pp	Tumor specific Fab fragment with its peptide epitope
1qkz*	Ig	1.95	Pp	Fab fragment with bacterial antigen
1ejo	Ig	2.30	Pp	Monoclonal 4C4 Fab fragment with G-H loop from virus FMDV
1g7i	Ig	1.80	PP	Monoclonal Fab fragment with Hen Egg White Lysozyme
1jps	Ig	1.85	PP	Fab D3h44 fragment with tissue factor
1lk3	Ig	1.91	PP	Fab 9D7 fragment with engineered IL-10

PP, protein-protein complex; pp, protein-peptide complex. *Soft harmonic force restraints (2 kcal/mol) were applied on C_α of the ligands to keep them in their binding sites during the simulation.

Trajectory processing. We defined interfacial interactions based on physico-chemical and distance

criteria between atoms. For hydrogen bonds, we considered a donor-acceptor distance of 3.2 Å, for salt bridges 4 Å, and for van der Waals interactions the van der Waals radii distance. Three classes of residues were introduced: dry (direct interaction), dual (direct and water-mediated interactions), and wet spots (residues interacting only through one water molecule) [27]. Each frame of the trajectory was processed so that the relative time fractions (TFs) of total, dry, dual and wet spot interactions (TF_T, TF_D, TF_d, TF_{ws}) during the simulation were corresponded to each residue. The total interaction was defined as a sum of all three defined interaction types. A residue was considered interacting if the total time of interaction was at least 5% of the simulation time. A residue was considered to be a wet spot if it interacted only through a single water molecule more than 10% of the simulation time. Such cut-offs were chosen arbitrarily in order to consider the wide range of the interactions for analysis and to restrict the definition of wet spots in molecular dynamics to a certain intuitively significant value.

Effective interface area calculations. The area of interface is usually defined as the difference between solvent accessible areas for the unbound molecule and for the same molecule in complex. We introduced "effective" interface areas related to water-less (ΔASA_{wl}) and water-mediated (ΔASA_w) interactions in order to estimate the impact of water-mediated interactions on the interface definition. The introduction of "effective" interface areas allows to consider that during the simulation the same interfacial residue could belong to dry (D), dual (d) and wet spots (ws) residue class for certain respective times:

$$\Delta ASA_{wl}=\sum_i \Delta ASA_i\,(TF_{D,i} + \tfrac{1}{2}\, TF_{d,i}) \qquad (2.1.1)$$

$$\Delta ASA_w=\sum_i \Delta ASA_i\,(TF_{ws,i} + \tfrac{1}{2}\, TF_{d,i}) \qquad (2.1.2),$$

where $TF_{D/d/ws,\,i}$ are the relative time fractions of residue i; ΔASA_i is the accessible surface area of the i^{th} residue calculated in the NACCESS program with a standard water probe radius of 1.4 Å [118].

Fluctuation analysis. The average fluctuation (F) for each interfacial residue was obtained with the PTRAJ module of AMBER 8.0 as a mass-weighted sum of fluctuations for atoms belonging to this residue. To implicitly decompose the impact of each type of interaction (total, dry, dual, wet spots) on the fluctuation as an analytically unknown function $F(TF_T, TF_D, TF_d, TF_{ws})$, the following method was used. The function values were averaged regarding to all other TFs except for the one of interest in order to obtain dependence on this certain TF: $\langle F(TF_{ij}, TF_{kj} \geq a)\rangle_i$, where i contains all interaction types (T= total, D= dry, d= dual, ws= wet spot) except k ($i \neq k$), k is the TF of interest, j is the summing index for interfacial residues and $a \in [0,100]$ expressed in %. Dependences on these 4 TFs were compared

qualitatively.

MM-GBSA free energy decomposition per residue. Energetic post-processing of the trajectories was done in a continuous solvent model as implemented in the AMBER 8.0 MM-GBSA (Molecular Mechanics- Generalized Born Surface Area) module. MM-GBSA is a method for free energy calculation utilizing implicit solvent model and is based on a Generalized Born approximation to the exact (linearized) Poisson-Boltzmann equation for electrostatics. The snapshots for the calculations were chosen as described by Lafont and coworkers [126]. To achieve better conformational space sampling, first, all frames of the trajectory were sorted by *in vacuo* calculated electrostatic energy values and the range of these energies was divided into 10 equal intervals. For each interval the number of corresponding conformations was calculated and served as a weight function for the interval. Then, for a conformation, most closely corresponding to the interval mean value of electrostatic energy, full MM-GBSA energy calculations were carried out. The final result was calculated as a weighted sum of values for each interval. The energy components per residue were compared by TFs (i.e. TF_T, TF_D, TF_d, TF_{ws}) in a similar way as it is described in the "Fluctuation analysis" section .

Residence time analysis of water molecules. The distance from the interacting heavy atoms of each wet spot counterpart to water molecules in each frame was calculated using the PTRAJ module of AMBER 8.0. If the distance did not exceed 3.6 Å, the wet spot site was considered to be occupied. A surface water site was defined by the volume that was closer than 3.6 Å to one of the protein polar groups and was not located in an interface. Solvent was considered as bulk at a distance ≥ 5 Å from the protein surface. Surface and bulk sites are defined so that their total occupancy is 100%. This makes them *a priori* more occupied in comparison to interfacial sites, where total occupancy is 64% on average. Therefore, when differences between surface, bulk and interfacial sites exist, conclusions can be made even stronger. The frequency of consecutively occupied frames for the site was presented as residence time distribution density. Maximum number of consecutively occupied frames for the site was corresponded to maximum residence time (T_{max}) of a water molecule in the site, and total occupancy was defined as the sum of all time intervals when the site was occupied.

Free energy perturbation calculations. For free energy calculations of water molecules in wet spot sites, the double decoupling method of free energy perturbation as described in the work of Hamelberg D., McCammon J.A. was used [44]. This method is based on two steps of perturbation. First, electrostatic interactions are gradually turned off; then van der Waals radii of chargeless atoms are decreased to 0. The free energy difference between two states was calculated using the thermodynamic

integration approach at discrete points of the coupling parameter λ, which was variated from 0 to 1 and then back from 1 to 0 with a 0.01 step along the path. Simulation for each λ value was equilibrated for 10 ps followed by a productive MD sampling for 10 ps. In the case of two water molecules in the same spot, the less mobile one was first removed. If the removal energy of the first water was negative (favorable) and the two waters were establishing hydrogen bond interactions in the site, both waters were removed from the spot at a time. AMBER prevents other water molecules from occupying the site of the perturbed one by considering the volume corresponding to the site of the perturbed water molecule as occupied by default.

Statistical analysis. Statistical analysis of data was carried out with the R-package [144].

2.1.4 Results and discussion

We have performed MD studies to deepen our understanding of the properties of protein interfacial residues (direct, dual, wet spots) and interfacial solvent. In this work we have studied protein interactions in terms of mobility, free energy and interaction conservation. For this purpose we have analyzed water-mediated interactions in a representative set formed by 11 complexes of SH3 domains (7 with peptides and 4 with proteins) and 6 complexes of Ig domains forming variable antibody fragments (3 with peptides and 3 with proteins) (Table 2.1.1). A total of 292 interfacial residues including 110 wet spots were analyzed.

Interaction patterns in MD simulations. MD runs of 10 ns were performed for each complex. To define interfacial residues the trajectories were processed and the corresponding TFs were calculated. A summary of properties averaged over all structures of a family for the interface description of the 2 representative protein families used in this study is presented in Table 2.1.2.

Table 2.1.2. Summary of interface properties

Domains	SH3	Ig
Number of interacting residues per domain	14±2	24±5
Interface area, $Å^2$	733±195	1291±471
Interacting residues/1000$Å^2$	19	19
Total observed wet spots in MD (SCOWLP)	49 (15)	61(19)
Wet spots in MD (SCOWLP)/complex	4.5 (1.4)	10 (3.2)
Wet spots in MD (SCOWLP)/1000$Å^2$	6.1 (1.9)	7.7 (2.5)
Wet spots in MD (SCOWLP)/interface residue	0.32(0.1)	0.42(0.13)
Effective areas ratio $\Delta ASA_w/\Delta ASA_d$, %	28±7	39±13

Immunoglobulin interfaces are almost twice bigger than SH3 domain interfaces, both in number of residues and in area size. However, the density of interactions (number of interfacial residues per area unit) is the same for both families. In both families the number of wet spots observed in MD simulations is about three times higher than in the corresponding PDB initial structures. This could be explained, first of all, because of the different nature of the source of information (obtained from PDB files, static; defined in MD, dynamic), and partly by resolution and quality of data contained in PDB files. Classification of interfacial residues based on X-ray data falls into 3 classes (direct, dual and wet spots), while in MD this discrete division is not possible due to the fact that the same residue may present different interacting modes during the simulation. A continuous model of residue interaction pattern requires more complicated, often implicit, not direct mathematical approaches for analysis. Table 2.1.2 illustrates roughly same ratios for the parameters for the two representative domain families though Ig interfaces have higher relative number of wet spots than SH3 domains interfaces. The last displayed parameter in Table Table 2.1.2 corresponds to the ratio of effective areas of dry and water-mediated interfaces (see Methods section). These ratios reflect how larger the area of the interfaces would be if we would include wet spots in the interface definition. We obtained roughly 30% and 40% of interface size increase for SH3 and Ig, respectively. Considering the importance of the interfacial area as empirical parameter in algorithms implemented for energy calculations, these numbers suggest that exclusion of the water molecules from protein interface analysis may lead to significantly biased or incomplete results.

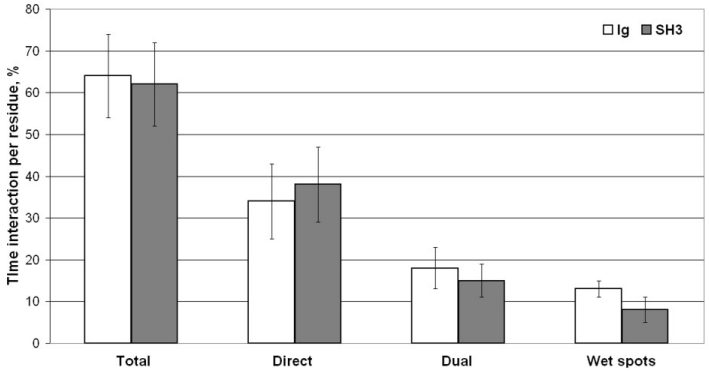

Figure 2.1.1. Average time fractions of interaction in the Ig and SH3 complexes.

In terms of total interactions per residue, despite the differences in size and chemical nature, both SH3 and Ig interfaces have comparable averaged contribution of interfacial residues to each class (Figure 2.1.1). The t-Test shows only a significant difference for the wet spots impact (at the level of p-value=0.05). Although the total occurrence of wet spots interactions is about 3 times lower than of

direct interactions, overall water-mediated interactions correspond to almost the same TF as direct interactions. The percentage of interactions in our complexes agrees with the 14% presence of wet spots in total interfacial residues obtained for a non-redundant dataset of protein complexes [64]. That means that, in terms of wet spots contribution, both SH3 and Ig families are close to average protein families. Direct and water-mediated interactions reveal differences in distributions of TFs (Figure 2.1.2). Direct interactions are almost uniformly distributed on all time fraction intervals, while dual and wet spots interactions are distributed mostly on intervals up to 30% of relative interaction time. At the same time, the distribution of total interactions shows that most of the residues are interacting during more than half of the simulation. Comparison of the distributions suggests that there are few interfacial residues forming wet spots interactions for a long time during simulations. However, the contribution of wet spots to total interaction is substantial (Figure 2.1.1). The analysis of the distributions shows that it is not correct to consider that an interfacial residue unambiguously belongs to only one class of interfacial residues.

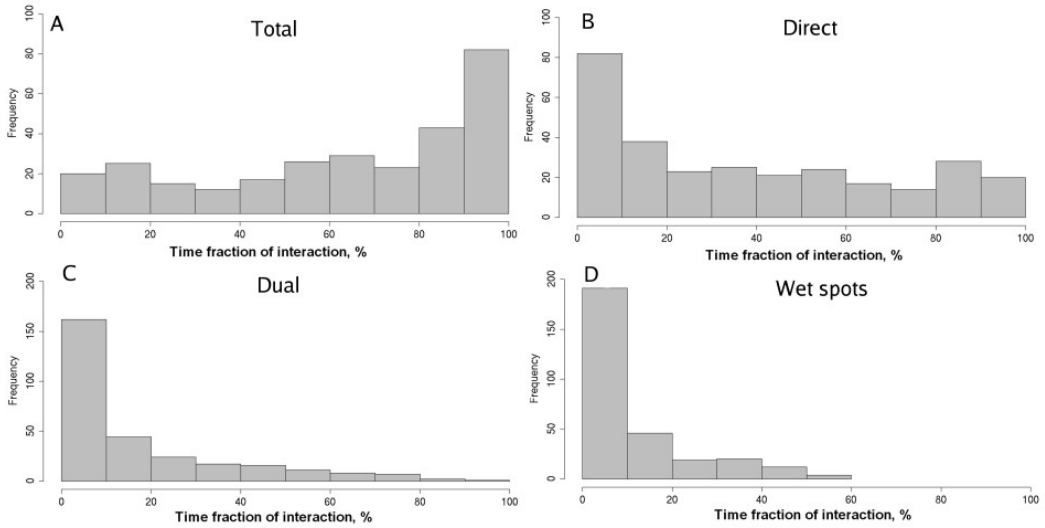

Figure 2.1.2. Distribution of time fractions of interaction for all simulated complexes.

We monitored wet spots in the MD simulations of the SH3 and Ig domains, and classified them by interaction type (main-chain/side-chain) as well as by amino acid composition (Figure 2.1.3). For both families side-chain interactions slightly prevailed (55% of all wet spots), which agrees with previous results obtained for transient complexes based on crystallographic data [64]. Wet spots show in general a strong preference for polar and negatively charged residues, mostly interacting through their side-chains. At the same time, water mediation increases the participation of hydrophobic amino acid residues through their main-chains in the formation of interfaces with a certainly less hydrophobic

character.

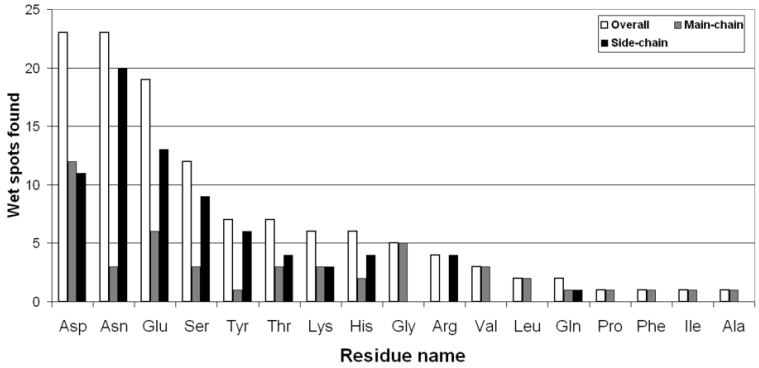

Figure 2.1.3. Participation of different residues in wet spots.

Interaction conservation through water. Water plays a role of "molecular glue" in protein interfaces [145]. Water may mediate conserved interactions in protein families and thus participate in specificity. Certain water-mediated interactions can be present in most of the interfaces of a protein family (e.g. Asn52 in SH3 domains; Figure 2.1.4). In addition, water may also keep the interactions conserved despite the introduction of semi- or non-conservative mutations in a specific position in a protein family (e.g. sites I, IV and V; Figure 2.1.4).

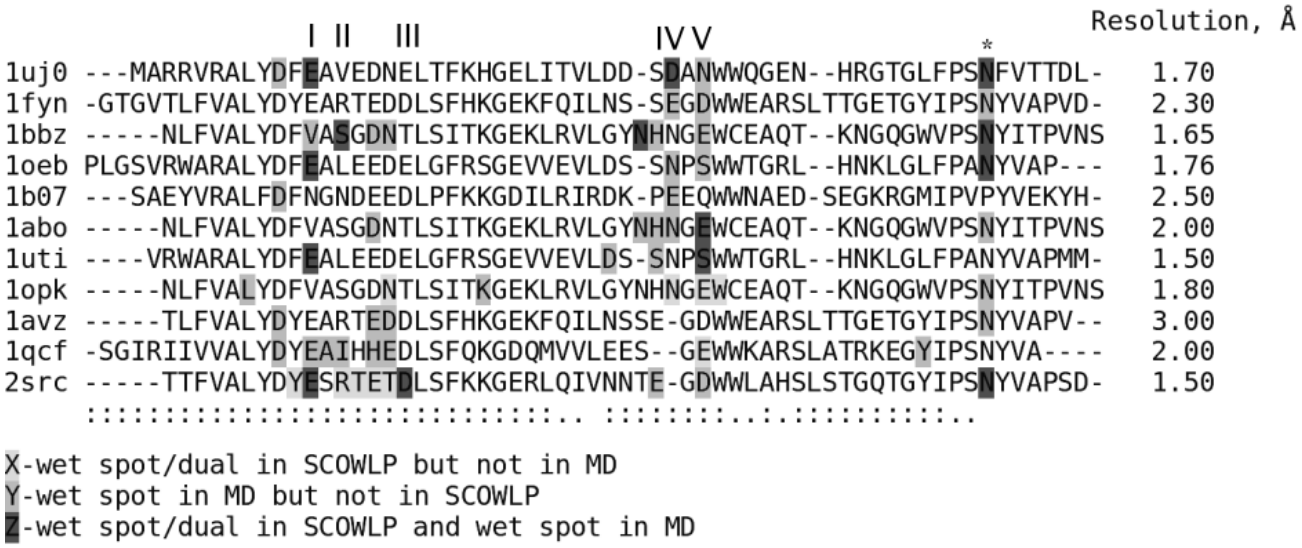

Figure 2.1.4. Structure-based sequence alignment of SH3 domains. Residues are colored by their participation in wet spots. The position of Asn52 (numbering by 1uj0) is labeled with an asterisk. Interaction sites from Table 2.1.3 are labeled with roman numbers at the top of the alignment.

Correlated mutations in binding partners are expected to appear as a result of co-evolution and aim to keep a specific interaction conservative [92,146]. However, due to the participation of water in protein interfaces conservation of an interfacial interaction may occur despite non-correlated mutations (Table 2.1.3).

Table 2.1.3. Examples of interaction conservation in SH3 domain interfaces

Site	PDB ID	SH3	Ligand	Site	PDB ID	SH3	Ligand
I	1uj0	E12m	R64s	III	1bbz*	T16s	Y63s
	1bbz	V10m	Y63s		1uti*	E17s	R66s
	1oeb	E15m	R65s	IV	1uj0	D34s	N66m
	1uti	E11m	R66s		1oeb	N37s	T67m
II	1bbz	S12s	Y63s		1uti*	N33s	E68m
	1qcf	I16m	E175s	Va	1uj0	N36s	N66s
	1fyn*	R16s	Y66s		1oeb	S39s	T67m
	1uj0*	V14s	R64s		1uti	S35s	E68s
	1oeb*	L17s	R65s	Vb	1fyn	N38s	Y66m
	1b07*	N14s	R68s		1abo	E35s	M62m
	1uti*	I13s	R66s	Vc	1bbz	E35s	P65m
III	2src	D16s	I172m		1avz	D237s	L106m
	1uj0*	E18s	R64s		1qcf	R35s	W174m
	1fyn*	D20s	Y66s		1b07*	R36s	P67m

s=side-chain, m=main-chain interactions; *=direct interaction; 5a, 5b, 5c correspond to different interactions in the same site.

Examples of interaction conservation through water found in SH3 domains are graphically shown in Figure 2.1.5. Figure 2.1.5 A illustrates site I with no correlation between the mutations in protein and ligand. The conserved interaction Glu(SH3)-Arg(ligand) is replaced in one of the complexes (1bbz) by the interaction Val(SH3)-Tyr(ligand). The interaction formed by the side-chain of different ligand residues with the protein is conserved due to the establishment of water-mediated main-chain interactions. Figure 2.1.5 B illustrates site III with direct interacting residues being replaced by wet spots. The conserved interaction between side-chains of ligand and side-chains of the protein is maintained in one of the complexes with a non-conservative mutation (1fyn) thanks to the establishment of a water-mediated main-chain interaction.

The concept of correlated mutations in protein-protein interaction studies was introduced in the 90s [92,146] and has been used since then to optimize protein design, predictions of protein interfaces and docking algorithms. Several matrices of residue-pairwise interacting probabilities have been built using different mathematical approaches and empirical parameters [147,97,93]. However, none of them considers solvent as mediator of interactions. Our results indicate that disregarding interfacial solvent may cause inaccuracies in the application of correlated mutations based approaches in the complete

analysis of protein interfaces and the prediction of protein interactions.

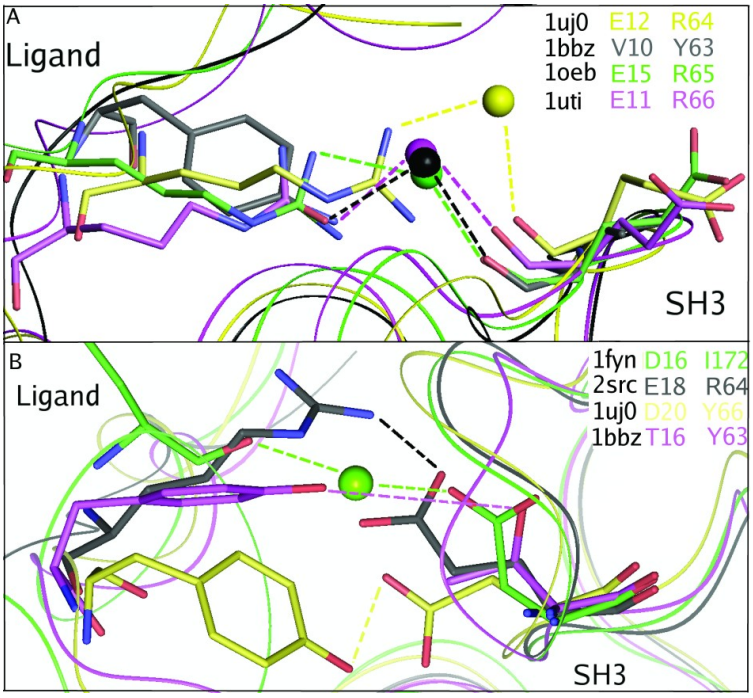

Figure 2.1.5. Examples of interfacial interaction conservation through water in SH3 interfaces. A) Site I of Table 2.1.3. with no correlation between the mutations in protein and ligand. B) Site III of Table 2.1.3. with direct interacting residues being replaced by wet spots. Proteins and ligands are represented by ribbons and labeled. Interacting residues are shown in sticks, and water molecules as spheres. The color code of the sequences in the upper right panels corresponds to the colors of the residues in the proteins and ligands as well as in the water molecules. Hydrogen bonds are represented with dash lines.

Fluctuation analysis. In previous work we showed that thermal B-factors of wet spots are comparable to those of other interfacial residues [64]. Prior to checking if the mobility properties of different interfacial residue classes could be distinguished, we compared the mobility of surface residues and interfacial residues in terms of average fluctuations. Our results show that fluctuations of interfacial residues (side-chain and also full residue) are significantly lower than those of surface residues (at the level of t-Test p-value=0.05), while there is no significant difference for backbone fluctuations. Implicit decomposition of the average fluctuation function calculated for all interfacial residues in the studied complexes shows that, in general, residue fluctuations decrease with the increase of residue interaction time (Figure 2.1.6). Dual residues fluctuate more than dry and less than wet spots. At the same time the residues in the protein interior were shown to be significantly (at the level of t-Test p-value=0.05) less mobile than those in the interface or surface. For example, for the 1UJ0 complex, the average fluctuations for protein interior, interfacial and surface residues were 0.50±0.06 Å, 0.74±0.30 Å and 1.00±0.48 Å, respectively. The fluctuation analysis of 111 surface and 151 interfacial residues also revealed significant difference (at the level of t-Test p-value=0.05) between these two groups of protein

residues (1.20±0.61, 0.83±0.59, respectively).

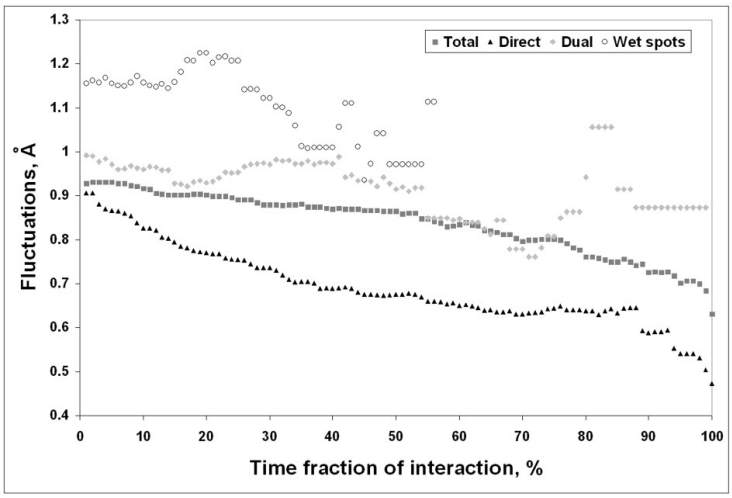

Figure 2.1.6. Fluctuations of interfacial residues decomposed by interaction type.

Our data agree with the thermal factor analysis performed on a large dataset of protein complexes, which found that the closer the residues are to the core of interfaces, the higher their stability [148]. Since the fluctuations of wet spots and surface residues at temperatures higher than 180K could be roughly explained in terms of surrounding water molecules mobility [149], it suggests that water flow around wet spots is slower in general than water molecules motion in the surface hydration shell. A similar trend was obtained for the interfacial residues participating in water-mediated interactions in another study where the speed of surrounding water flow and the mobility of the residues were analyzed in MD simulations [138]. Therefore, dynamical properties of interfacial residues and solvent mediating interaction are tightly interconnected, and they could be mechanically described as a coupling of harmonic oscillator to solvent modes via small springs (hydrogen bonds) [51]. This means that dynamic analysis of the interfacial residues might be biased without the consideration of surrounding water molecules.

Free energy decomposition per residue in interfaces. The MM-GBSA method applied for free energy decomposition calculations per residue allows to obtain the following independent components describing the energetics of a protein complex in implicit solvent: electrostatic component *in vacuo*, van der Waals interactions component *in vacuo*, Generalized Born reaction field energy [150] and hydrophobic component of solvation. The differences in energy values for interfacial residue classes were compared with the characteristic thermal motion energy value at 300K (RT~0.6 kcal/mol). The

differences in hydrophobic component of solvation were lower than this value, so we considered that this component does not differ significantly among the interfacial residue classes. Generalized Born reaction field energy roughly compensates for the electrostatic component in vacuo, so we discuss only the results obtained for van der Waals and electrostatic components *in vacuo*. The general trend for both components and all types of interactions is that the energy values decrease with the increase of residue interaction time, suggesting that both energy components stabilize complexes independently of the residue class.

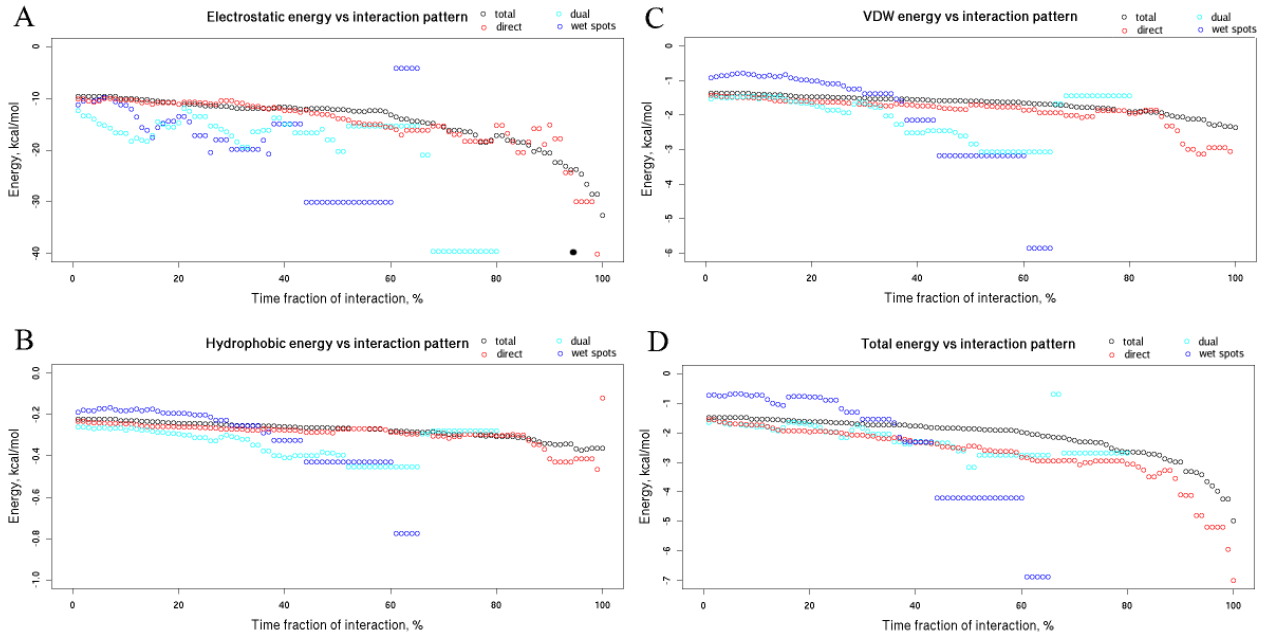

Figure 2.1.7. Free energy decomposition for interfacial residues decomposed by interaction type. A) Electrostatic energy. B) Van der Waals energy. C) Hydrophobic component of solvation energy. D) Total MM-GBSA energy.

Energy decomposition (Figure 2.1.7) illustrates that the electrostatic impact of water-mediated interactions is at least of the same order as of direct interactions, considering that the dielectric constant of water in the protein interface within the analyzed distance scale is approximately one order lower than the dielectric constant in bulk (and several times higher than in dry interfaces) [151]. The Van der Waals energy component is the lowest for dual residues and the highest for wet spots. Such a benefit for dual residues is explained by more tight contacts of the atoms additionally summed up with water atoms contacts. In wet spots there are only contacts with water atoms, which are not so tightly packed.

Despite the quantitative differences observed for the SH3 and Ig interfaces, the important finding out of the energy decomposition is that all three interfacial residue classes are energetically comparable even in implicit solvent, meaning that wet spots interactions are energetically of the same order as direct interactions. This conclusion could be generalized for all protein interfaces since the

analyzed interface families substantially differ in physico-chemical properties. The obtained small differences between the families in the free energy components are due to intrinsic properties of the dataset, and the larger size and more hydrophilic nature of Ig interfaces. Non-interfacial residues were analyzed in the same way but their free energy contribution values were at least one order lower than of the interfacial ones. This suggests that inclusion of water-mediated interactions in protein interface definition is energetically well grounded.

Residence time of water molecules in wet spot sites. The analysis of the residence time distribution density of wet spot sites obtained from MD simulations suggests that the best theoretical model to describe the distribution density function should be defined as $\rho(t)=Ct^{-k}$, where C is a constant that can be obtained by normalization for each site individually, and the constant $k>0$ is the only distribution parameter. This k constant and the maximum residence time (T_{max}) in sites were taken as the parameters to compare different water sites. For wet spot sites the linear regression adjusted correlation coefficient r is equal to 0.97±0.04 with p-value ranging from $8*10^{-13}$ to $3*10^{-2}$. r and p-values for most of the surface and bulk solvent sites were not defined because the observed residence times were in the most cases less than 20ps, meaning that there were just 2 points in the distribution (each point was obtained summing up the number of events on a 10ps interval to avoid big fluctuations in the density function).

Table 2.1.4. Residence time parameters of different water sites

Site type	Sample size	Tmax, ps	k
Wet spot	110	137±12	3.0±1.0
Surface	30	18±6	7.1±1.1
Bulk	10	14±2	8.7±0.9

Sample size is the number of analyzed sites of each water site type. T_{max} is maximal residence time.
k is residence time distribution parameter.

As it is shown in Table 2.1.4, both k and T_{max} significantly differ (at the level of t-Test p-value=0.05) for different sites, indicating that water molecules in wet spot sites are less mobile than in bulk solvent or in surface hydration sites. At the same time, in each wet spot site many occupancy events occur that have as short residence as in bulk or surface sites. That agrees with the model proposed by Makarov et al., where the correlation function for residence time in hydration sites is decomposed into the sum of fast and slow diffusion exponent components. These components characterize bulk water motions and specific for hydration site events, respectively [56]. Other theoretical and experimental studies obtained similar residence time values for different water sites, which vary from 1-10 ps for bulk solvent to 10^1-10^3 ps for protein hydration sites, cavities and cores

[32]. T_{max} and k are well correlated (adjusted correlation coefficient r=0.81 for $\ln(T_{max})$~k linear regression, Figure 2.1.8 A), meaning that maximum residence time of water molecules does not correspond to an opportunistic event of site occupation but is expected from the residence time distribution. The correlation between total residence time/maximum residence time and water mediated interaction amount is weak but the dependence is clearly observed (Figure 2.1.8 B, C). There was no correlation between total occupancy of the sites and T_{max} ($r<0.3$, Figure 2.1.8 D) because these parameters are independent and describe different kinetic characteristics of the site. While T_{max} is defined only by the energy barrier required for the molecule to leave the site, total occupancy is also dependent on the energy barrier of water transfer from bulk solvent to the site. The residence time analysis suggests that the potential barriers for wet spots sites are significantly higher than those for surface sites. However, it does not mean that the potential energy level of water molecules in wet spot sites are necessarily lower than in bulk solvent.

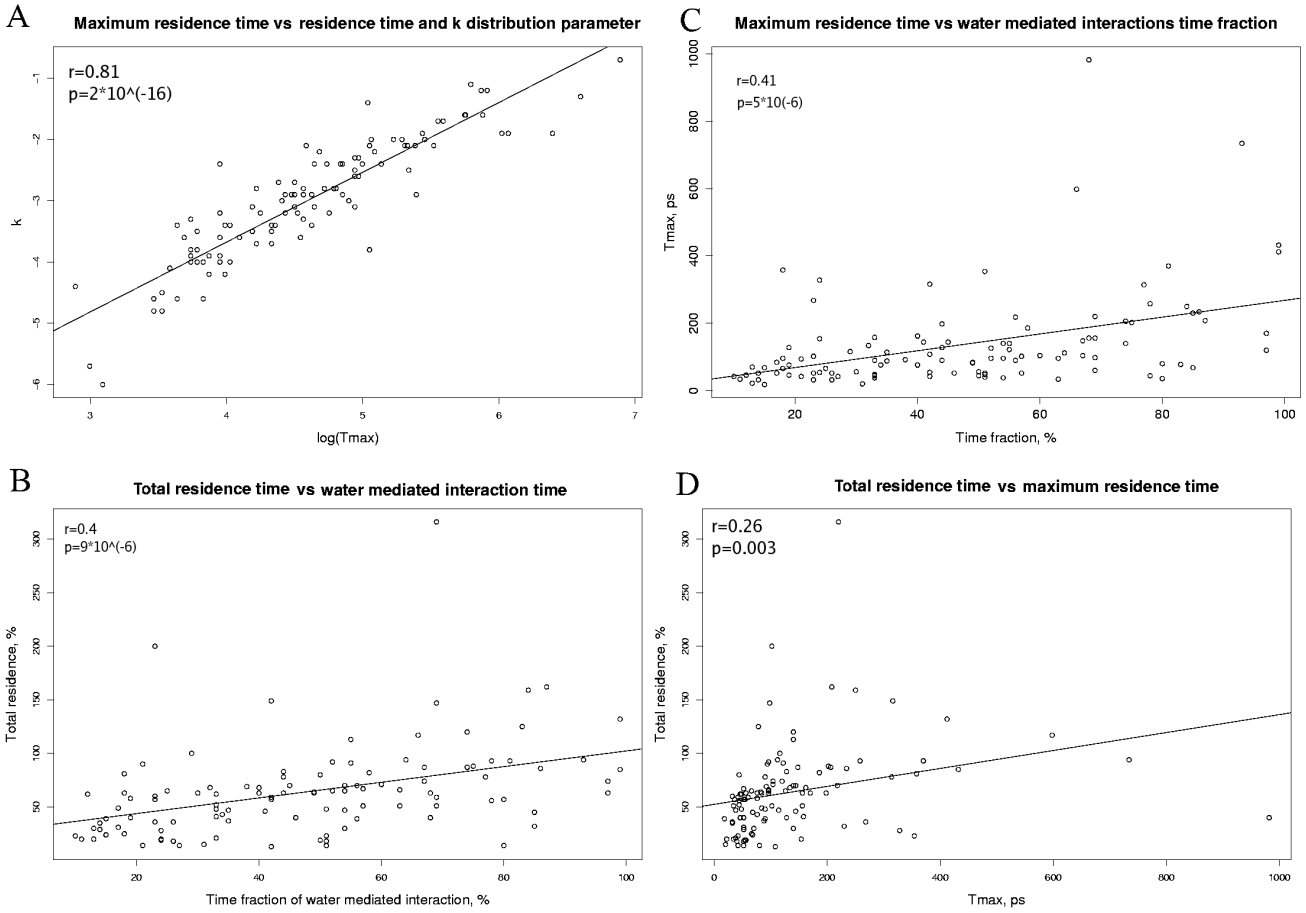

Figure 2.1.8. Interdependence of different time related wet spot sites parameters. A) T_{max} *vs* k. B) T_{max} *vs* water mediated interaction time. C) Total residence time *vs* water mediated interaction time. D) Total residence time *vs* T_{max}.

Free energy of water molecules in wet spot sites. To determine if water molecules contribute

energetically favorably to complex formation we calculated their free energy using the free energy perturbation double decoupling method [44]. As a first step, free energy of removing a water molecule from bulk solvent was calculated. Electrostatic and van der Waals components were equal to 8.2 and -2.2 kcal/mol, respectively, which agrees well with the results obtained from similar calculations correlated with experimental data [44,45] (Table 2.1.5). The second step consisted of the transfer of a water molecule from the wet spot site to vacuum. The difference of these two energy components makes up the total energy of a water molecule transfer from bulk solvent to the wet spot site.

Table 2.1.5. Free energy perturbation of water molecules in 1uj0 complex using the double-decoupling method

E (kcal/mol) Site	Site type	E_{Elect}	E_{VDW}	$-RT^* \ln(S_w S_p / S_{w^*p})$	$RT^* \ln(C_0 V_1)$	E_{rot}	E_{total}	ΔG^0
E12-R64	Wet spot	12.9	-1.5	0.4	-4.4	0	7.4	-1.4
D34-N66	Wet spot	8.3	0.1	0.4	-4.1	0	4.7	1.3
D34-N66, 2 water molecules	Wet spot	22.9	-3.7	0.8	-8.2	0	12.6	-6.6
N52-M61,N63	Wet spot	8.9	0.1	0.4	-4.2	0	5.2	0.8
N52-M61,N63 2 water molecules	Wet spot	18.1	0.1	0.8	-7.2	0	11.2	0.1
L58-R6	Surface	9.8	0.2	0.4	-3.8	0	6.4	-0.4
D31-S33	Surface	7.6	-0.6	0.4	-3.7	0	3.7	2.3
Lysozyme	Cavity	13.5	0.0	0.4	-3.9	0	10.0	-4.0
Bulk-vacuo transfer	Bulk	8.2	-2.2	-	-	-	6.0	-
Bulk-vacuo transfer [20]	Bulk	8.2	-2.2	-	-	-	6.0	-
Bulk-vacuo transfer [34]	Bulk	8.3	-2.4	-	-	-	5.9	-

E_{Elect}, electrostatic energy; E_{VDW}, van der Waals energy; $-RT^* \ln(S_a S_b / S_{a^*b})$, the free energy component related to the symmetry of water molecule (S_a), protein (S_b) and the complex of water molecule with protein (S_{a^*b}); $RT^* \ln(C_0 V_1)$, the free energy component associated with translational constraints in the V_1 volume; E_{rot}, the free energy component associated with rotational constraints; E_{total}, total free energy of a water molecule transfer from the site to vacuo; ΔG^0, free energy of transfer of a water molecule from bulk to the site [20].

The obtained results for several water sites of the SH3 domain complex 1uj0 show that the sites are very heterogeneous. In particular, the free energy of water molecule transfer from bulk to the site formed by the carboxyl oxygen of Glu12 in the SH3 domain and the side-chain of Arg64 in the ligand is -1.4 kcal/mol, meaning a favorable impact of a water molecule on the complex formation. The calculations carried out for another site formed by the side-chain of Asp34 in the SH3 domain and the side-chain of Asn66 in the ligand revealed a positive change of free energy (1.3 kcal/mol). However, as it was observed in the trajectory, another water molecule was present in this site and establishing a

hydrogen bond with the first water molecule forming the wet spot. Consideration of both water molecules in the free energy calculations revealed an energy gain of -6.6 kcal/mol (Table 2.1.5). In this case removal of two water molecules from a wet spot site leads to an increase of the free energy value, while a removal of each water molecule independently leads to a free energy decrease. Another example of such an effect was found in the site formed by the side-chain of Asn52 in the SH3 domain and the main-chain of Met61 in the ligand. Here, although the energy became more favorable by taking into account two water molecules transfer, water contribution was still not favorable. For the comparison of the free energies of wet spot sites we took several surface sites and a site in the cavity of lysozyme, which is not exposed to bulk solvent (1HEL, 1.7 Å). The X-ray structure of lysozyme presents a very stable water molecule [152], which is present in the site during the whole simulation with a residence time of several nanoseconds. The energetic impact of the cavity water to the stability of the lysozyme was quite significant (-4 kcal/mol). In surface sites no big negative values for free energy were found. In similar calculations performed with the double decoupling method for free energy calculation with AMBER, the obtained values for the free energy of water in hydration sites changed from slightly positive up to -5 kcal/mol [44,45]. The examples of favorable energetic impact of water molecules on complex formation was also found in a study of various protein complexes by Monte Carlo calculations using different force fields [46].

The most important conclusion that can be driven from this free energy analysis is that water molecules in wet spot sites can not be characterized uniformly in energetic terms since in some cases they manifest properties similar to cavity waters and in other do not even contribute favorably to the complex free energy (just occupying an empty space between the residues). Nevertheless, it is realistic to claim that the introduction of water into protein interface description would crucially change the energy function of the system. Interestingly, a recent attempt of solvated protein docking has shown promising results [61].

2.1.5 Conclusions

We present a detailed molecular dynamics study on 17 protein complexes representatives of two families of different interface nature. Our aim has been to gain insights into the contribution of interfacial solvent in protein-protein interactions. We show that water molecules in protein interfaces contribute to the conservation of protein interactions by allowing more sequence variability in the interacting partners, which has important implications for the use of the correlated mutations concept in

protein interactions studies.

Interfacial residues interacting through water are more mobile than those interacting directly but less than protein surface residues. Despite their broad heterogeneity, all interfacial residues are quantitatively comparable in terms of their contribution to the energy of complex formation, independently of their type of interaction. In the case of interfacial solvent, water molecules forming wet spots have significantly longer residence time than those on the protein surface, meaning that in terms of mobility interfacial protein residues and interfacial solvent are alike. Although interfacial water molecules are very diverse energetically, their contribution to the free energy of complex formation should be not be ignored.

Our data confirm that water plays an important active role in protein interfaces, suggesting that consideration of solvent in the development of energetic functions describing protein interactions is essential. Moreover, the introduction of water-mediated interactions into protein interface definitions should substantially increase the accuracy of protein interaction predictions based on protein contacts. We believe that the results obtained in this work could be useful for deeper understanding of the physico-chemical properties underlying protein-protein interactions in order to improve the accuracy of protein folding, docking and rational design methods.

2.2 Analysis of the impact of solvent on contacts prediction in proteins

by Sergey Samsonov, Joan Teyra, Gerd Anders and M. Teresa Pisabarro

BMC Structural Biology 2009 Apr 15;9(1):22.

2.2.1 Abstract

Background. The correlated mutations concept is based on the assumption that interacting protein residues coevolve, so that a mutation in one of the interacting counterparts is compensated by a mutation in the other. Approaches based on this concept have been widely used for protein contacts prediction since the 90s. Previously, we have shown that water-mediated interactions play an important role in protein interfaces. We have observed that current "dry" correlated mutations approaches might not properly predict certain interactions in protein interfaces due to the fact that they are water-mediated.

Results. The goal of this study has been to analyze the impact of including solvent into the concept of correlated mutations. For this purpose we use linear combinations of the predictions obtained by the application of two different similarity matrices: a standard "dry" similarity matrix (DRY) and a "wet" similarity matrix (WET) derived from all water-mediated protein interfacial interactions in the PDB. We analyze two datasets containing 50 domains and 10 domain pairs from PFAM and compare the results obtained by using a combination of both matrices. We find that for both intra- and interdomain contacts predictions the introduction of a combination of a "wet" and a "dry" similarity matrix improves the predictions in comparison to the "dry" one alone.

Conclusions. Our analysis, despite the complexity of its possible general applicability, opens up that the consideration of water may have an impact on the improvement of the contact predictions obtained by correlated mutations approaches.

2.2.2 Introduction

The correlated mutations concept was introduced in the 90s [92,153-155] and has been widely used for protein contacts prediction [93]. The method is based on the assumption that interacting protein residues co-evolve, so that a mutation in one of the interacting counterparts is compensated by a mutation in the other. Therefore, it is possible to introduce an exchange matrix or other measures of similarity for each sequence position in a multiple sequence alignment and to use covariance (correlation coefficient) between two positions to predict if the residues at these positions may establish physical contact in 3D space, and develop contact maps. Several different similarity measures and algorithms have been implemented in the concept of correlated mutations [93,156,96]. Most exchange

matrices are based either on physico-chemical properties of amino acids or on statistical data on the substitutions obtained from multiple sequence alignments [157]. Statistically it is clear that the distribution of distances between the residues at highly correlated positions is shifted towards lower values compared to the distance distribution of all residues. This has been demonstrated in the study of correlated mutations for residues within one protein domain (intradomain), for residues from different domains in multidomain proteins (interdomain intraprotein) [158,159] and in transmembrane proteins [160]. At the same time, attempts to use the concept of correlated mutations to predict thermodynamically coupled residues have suggested that the method is successful only for residues in evolutionary constrained positions [161].

The concept of correlated mutations has been intensively developed recently. The implementation of neural nets into algorithms of contact predictions has allowed to substantially improve the accuracy of the methods in a number of studies [147,162-164]. Also the application of filtering procedures such as the similarity of sequences in a dataset and the number of sequences in multiple sequence alignments, introduction of weights for physico-chemical properties of the residue pairs and creation of sub-multiple sequence alignments were successfully used to increase a true positive ratio of contact predictions [97]. Nowadays, different correlated mutations based approaches yield predictions accuracies in the range of 0.1-0.4 [97] but they are still of little use in the *ab initio* prediction of protein structure [96] .

Previously, we have shown that water-mediated interactions play an important role in protein interfaces [64,165]. In particular, we observed that the interfacial residues interacting only through one water molecule (wet spots) are more similar in terms of dynamic and energetic properties to residues in the core of proteins than to residues on the protein surface. Moreover, in our studies interfacial water molecules show significantly longer residence times than water molecules on the protein surface or in bulk solvent, and have been shown to give an indispensable energetic impact on complex formation [165]. In other studies it has been demonstrated that inclusion of solvent term into the Hamiltonian of protein systems has improved folding predictions compared to *in vacuo* folding models [48]. Also consideration of solvent explicitly in protein docking approaches has recently shown promising results [61]. In addition, we have observed that water molecules in protein interfaces may contribute to the conservation of interactions by allowing more sequence variability in the interacting partners. In particular, we have observed water-mediated interactions in protein complex interfaces that are not predicted by “dry” correlated mutations approaches [165]. Interestingly, in one of the recent studies on

correlated mutations, protein contacts prediction has been shown to be more accurate for protein cores than for the whole protein [94]. This could be partly explained by a higher conservation of residue contacts in protein cores, especially the hydrophobic ones [166] and probably also by the fact that the participation of solvent in protein contacts is being ignored.

The goal of this study has been to analyze the impact of including solvent into the concept of correlated mutations. For this purpose, we use a linear combination of predictions obtained by the use of two similarity matrices: a standard and widely used "dry" similarity matrix (DRY) [167] and a "wet" similarity matrix (WET) derived from data on all water-mediated protein-protein interfacial interactions in the PDB [168]. We compare the predictive results obtained with different combinations of these two similarity matrices in terms of number of correctly predicted contacts, accuracy, improvement ratio over random prediction for intradomain contacts and distributions of distances between residues in interdomain pairs.

Our results show that, despite a partial interdependence of both WET and DRY matrices, there is a clear trend pointing that a combination of these two matrices yields improved predictions over the single use of the DRY matrix for both intra- and interdomain contacts. The results obtained in this work underline the importance of water-mediated interactions in the description of protein-protein interactions, and that implementing combinations of "dry" and "wet" matrices could possibly improve the results obtained by correlated mutations-based approaches.

2.2.3 Methodology

Dataset and multiple sequence alignments. We based the generation of our dataset on previous similar studies [92,158,94]. Our dataset includes 50 domains and 10 domain pairs extracted from the PFAM database [169]. Consecutive increase of the size of our dataset for intradomain contacts did not significantly change our results.

For most of the families, only seed sequences were used, except for the cases when the number of seed sequences was less than 20. Datasets with a smaller number of sequences are not supposed to be useful in correlated mutations analysis [94]. The *reference sequence* (corresponding to the structure used for predictions evaluation) was added to the set of sequences, if this did not already contain it, following the same procedure that Eyal and co-workers used for obtaining a substitution matrix for protein structure prediction purposes [94]. Multiple sequence alignments were obtained with CLUSTALW [87]. Sequences with more than 95% of identity were not taken into account.

For the interdomain dataset the sequences from the two domain families were aligned

independently. Except for the case of immunoglobulins, where light and heavy chains were used as two interacting domains, all interdomain entries in the dataset contained pairs of two different PFAM domains. Reference structures had resolution ≤ 2.0 Å except for five of them (1BU1 and 1A19 taken from the Eyal et al dataset and 2HB2, 1WMG, 1ZWW taken into account to enrich the dataset with bigger domains and highly represented families).

Source and analysis of atomic data on protein structures. An in-house relational database of protein structures (XMLRPDB) and the SCOWLP database [27,168] were used to obtain interaction information including solvent from X-ray structures in the PDB.

Contact definition. Residue contacts in a reference structure were defined by following the physico-chemical criteria from SCOWLP [27]. We considered a 3.2 Å donor-acceptor distance for hydrogen bonds, 4 Å for salt bridges, and van der Waals radii for van der Waals interactions.

Similarity matrices. We used the McLachlan similarity matrix (based on structural and genetic similarities of amino acids) as a 'dry' matrix (DRY) [167]. To build a 'wet' matrix (WET) we extracted information on protein interfacial residues and solvent from all available X-ray PDB structures using the SCOWLP database [27,168]. In this database, three classes of interacting residues are defined based on their interactions: dry (direct interaction), dual (direct and water-mediated interactions), and wet spots (residues interacting only through one water molecule). For each type of amino acid residue the probability of participation in water-mediated interactions (by establishing hydrogen bond by main chain or side chain) in protein interfaces was calculated as:

$p_i=N_{i,w}/N_{i,total}$ (Table 2.2.1), where i corresponds to any of the 20 amino acids; $N_{i,w}$ is the number of the residues of this type forming wet spots or dual interactions; and $N_{i,total}$ is the total number of residues of this type participating in interfaces in all PDB structures. Each element of the WET similarity matrix was then defined as:

$WET_{ij}=1-|p_i-p_j|$, where i and j correspond to any of the 20 amino acids.

The fact that for the creation of the wet matrix we take low resolution structures containing either none or few water molecules into account when considering the whole PDB does not bias the WET matrix because it affects each probability proportionally.

Correlation coefficient calculations. For both DRY and WET similarity matrices the corresponding covariance matrices were calculated as previously described (Göbel et al 1994) using formula:

$$r_{ij} = \frac{1}{N^2}\sum_{k,l}^{N}\frac{W_{kl}(S_{ikl}-<S>_i)(S_{jkl}-<S>_j)}{\sigma_i\sigma_j} \qquad (2.2.1),$$

where N is the number of sequences; i and j are sequence position numbers; S_{ikl} is a value from the similarity matrix (DRY or WET); S_i is the mean of S_{ikl}; σ_i is the standard deviation of S_{ikl}; and W_{kl} is a weight matrix defined as:

$$W_{kl} = 1 - \frac{1}{L}\sum_{i=1}^{L}\delta(R_{ik}, R_{il}) \quad (2.2.2),$$

where L is the sequence length; R_{ik} and R_{il} are the residue types at position i in the sequences k and l, respectively; and δ is Kronecker delta [170].

For the interdomain dataset the weight matrix W_{kl} was calculated as an average for the domains and weighted by sequence length. The positions with more than 10% of gaps as well as completely conserved positions were not included in thSCOWLP criteriae calculations (zero was assigned to the corresponding correlation coefficient). After calculating covariance matrices based on the DRY and WET similarity matrices, we built their linear combinations:

$$r_{ij}=r_{ij\,DRY} + \alpha \cdot r_{ij\,WET} \quad (2.2.3),$$

where α takes values from {0, 0.1, 0.2, 0.5, 1, 2, 4, 10, 20}, so that the weight ratio between the impact of DRY and WET represents the range from completely dry ($\alpha=0$) to extremely WET-biased covariance ($\alpha=20$).

Evaluation of intradomain predictions. For evaluation of intradomain contacts predictions we used previously described methodology [92]. Sequence separation of 0, 6, 12 and 24 was used. Prediction *accuracy* was defined as the ratio between the number of correctly predicted contacts (C_{corr}) and total number of predicted contacts (C_{tot}). *Random accuracy* corresponds to the probability of correct prediction of the contact by chance and is equal to the ratio between experimentally observed contacts (C_{obs}) and maximum number of possible contacts. The ratio between accuracy and random accuracy was introduced as *improvement ratio over random prediction*. *Wet prediction ratio* is equal to accuracy normalized by the accuracy obtained by using only the DRY matrix (α=0). For the reference structures C_{corr} was taken as the number of contacts defined by SCOWLP criteria (see the Contact definition section in Methods).

Distance calculation and harmonic average (X_d). In the analysis of interdomain contacts the accuracy calculated in the same way as for intradomain contacts (typical value C_{obs}~10^2) is expected to be at least one order of magnitude lower (typical value C_{obs}~10^1). That is why comparison of accuracy, improvement ratio over random prediction and C_{corr} as functions of α is not appropriate in this case. It has been shown that the distribution of distances between the correlated pairs is shifted to lower values

compared to the distribution of distances for all residue pairs in two domains [158]. In our study we use a harmonic weighted difference statistic X_d described before [158]:

$$X_d = \sum_{i=1}^{n} \frac{P_{ic} - P_{ia}}{d_i n} \quad (2.2.4),$$

where n is the number of distance bins; d_i is the upper limit for each bin normalized to the maximum value of the distributed distances; P_{ic} is the percentage of the analyzed correlated pairs at the distances between d_i and d_{i-1}; and P_{ia} is the same percentage for all pairs of residues. The width of bin was 4 Å. The higher the X_d value, the more successful a prediction is.

Different definitions for the distance between residues resulted in all cases in the same trends and quantitatively only slightly affected X_d values. For interdomain pairs we used distances between the centers of mass of residues in order not to be biased to either main-chain or side-chain contacts.

For X_d calculations we took the best L/2 contacts for intradomain and $(L_1+L_2)/2$ contacts for interdomain contact predictions, where L_1 and L_2 are the reference sequences of the two interacting domains.

Although both the wet prediction ratio and X_d characterize the predictive power of the method, it is irrelevant to compare the results obtained for these parameters with each other. The same applies to α values corresponding to best predictions.

Statistical analysis. Statistical analysis of data was carried out with the R-package [144].

2.2.4 Results and discussion

Residue-solvent relations in proteins. Independently of residue types, we calculated the average ratios between the number of residues found to be in contact with water and all residues in X-ray PDB structures. A negligible difference was found between these ratios for interfaces and the whole protein (0.33 and 0.35, respectively). The ratios by residue type (Figure 2.2.1 and Table 2.2.1) correlate with an adjusted squared correlation coefficient R^2=0.90 (p-value~10^{-10}) and there is also a clear trend of residue ratios distribution in interfaces, which relates to their hydrophilic properties. This agrees with observations obtained from other datasets not including the whole PDB [171]. The better correlation between the ratios and the hydrophilicity index for interfaces compared to the whole protein (R^2=0.62 p-value~10^{-5} and R^2=0.44 p-value~10^{-3}, respectively) could be explained by the fact that the whole protein includes many residues in the core that are not accessible to water. This further supports the evidence that residue-solvent relations in protein interfaces are different from the ones in the proteins as

a whole [64,165].

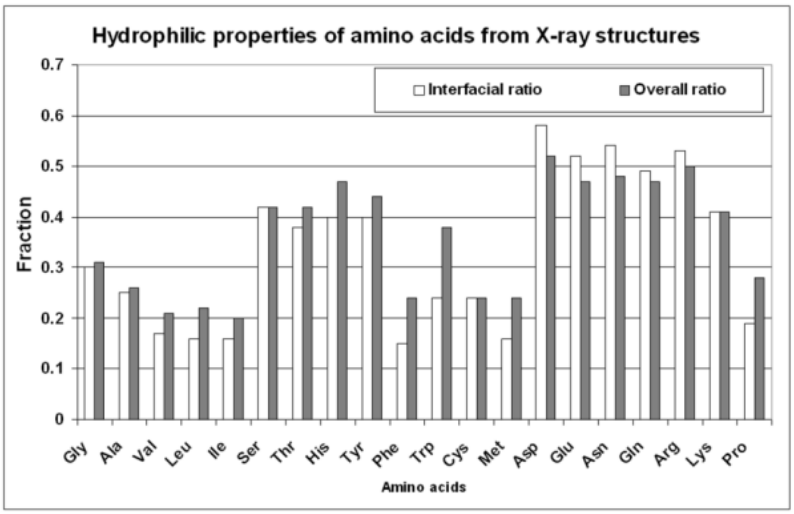

Figure 2.2.1.Water contacts of residues in PDB. Fractions of residues found to be in contact with water in protein interfaces (white) and in whole proteins (grey) in the PDB.

Table 2.2.1. Probabilities for residues to be in contact with water in protein interfaces

Residue	Total in interfaces	In contact with water	Probability	Residue	Total in interfaces	In contact with water	Probability
Gly	131875	40188	0.30	Pro	104724	19398	0.19
Ala	133562	33008	0.25	His	71046	28339	0.40
Val	128573	21609	0.17	Met	56221	8871	0.16
Leu	188008	29506	0.16	Cys	20393	4913	0.24
Ile	111915	18277	0.16	Asp	134113	78111	0.58
Ser	119168	50556	0.42	Asn	102592	55597	0.54
Thr	123482	47469	0.38	Glu	147932	77461	0.52
Tyr	114596	45580	0.40	Gln	94758	46319	0.49
Phe	106920	15936	0.15	Arg	163652	86656	0.53
Trp	42958	10448	0.24	Lys	116322	47565	0.41

The probabilities are derived from SCOWLP data for protein interfaces.

Relations between the DRY and WET similarity matrices. Both DRY and WET similarity matrices are created in a way that each column or row is a vector, which coordinates correspond to the similarity between certain amino acid residue type and other residue types. It is possible to define whether these vectors are interdependent for both matrices by application of linear regression analysis. The data obtained and averaged for all types of residues are presented in Table 2.2.2.

Table 2.2.2. Correlation between vectors per residue type in the DRY and WET matrices

Residue	p-value	Adjusted R^2	Residue	p-value	Adjusted R^2
Ala	0.90	-0.05	Leu	$6 \cdot 10^{-3}$	0.31
Arg	$4 \cdot 10^{-3}$	0.35	Lys	$8 \cdot 10^{-3}$	0.29
Asn	$4 \cdot 10^{-5}$	0.65	Met	$6 \cdot 10^{-3}$	0.31
Asp	$6 \cdot 10^{-4}$	0.46	Phe	0.02	0.24
Cys	0.14	0.07	Pro	0.62	-0.04
Gln	$5 \cdot 10^{-4}$	0.47	Ser	$2 \cdot 10^{-3}$	0.39
Glu	$4 \cdot 10^{-4}$	0.49	Thr	0.07	0.12
Gly	0.53	-0.03	Trp	0.18	0.05
His	0.02	0.22	Tyr	0.71	-0.05
Ile	$8 \cdot 10^{-4}$	0.44	Val	$4 \cdot 10^{-3}$	0.33

High degree of correlation is observed for some vectors, which correspond to hydrophilic residues (excluding Thr and Tyr) and for Ile, Leu, Met, Val, suggesting that these vectors in the matrices are close to be collinear in 20-dimensional space. This can be explained by the properties of these residues. In particular, hydrophilic residues interact by electrostatic forces through their polar atoms, and water mediation in this case can only change the electrostatic forces by introducing water dipoles oriented in a way to weaken the initial electric field. For hydrophilic residues there is a correlation between hydrophilicity indexes and co-linearity of the corresponding vectors in the DRY and WET matrices, which explains also relatively low co-linearity for Tyr and Thr residues in comparison to other hydrophilic residues (Figure 2.2.2).

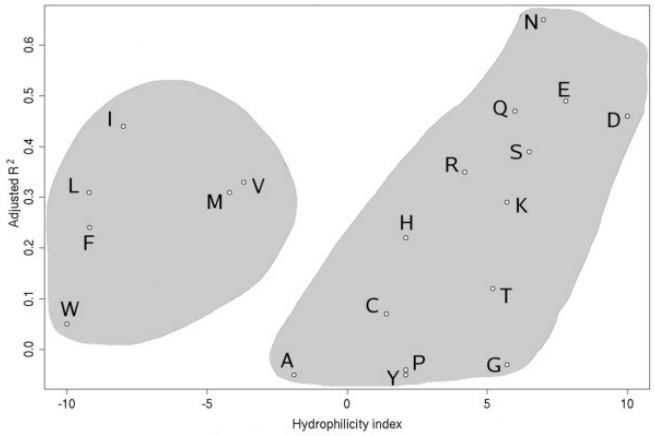

Figure 2.2.2. Hydrophilicity index *vs* correlation for the DRY and WET matrices per residue type. The grey shading highlights two areas resulting from the different trends.

Direct and water-mediated interactions formed by main chains of Ile, Leu, Met and Val in interfaces have been previously shown to be especially important, whereas other residues that present no correlation have been shown to predominantly participate in side-chain interactions in interfaces [64]. We conclude that the DRY and WET similarity matrices contain partially interdependent information for some of amino acid residues, and the found similarities can be explained by the physico-chemical properties of these residues.

Intradomain contacts prediction. Our dataset for intradomain contacts prediction consisted of domains of 50 PFAM protein families (Table 2.2.3). The lengths of the reference sequences varied from 30 to 195 residues. Initially we analyzed L, L/2, L/3, L/5 and L/10 best correlated contacts for each family (L is the length of the reference sequence). The number of sequences considered for the multiple sequence alignments was in the range of 20 to 295 sequences. Previous studies have shown that accuracy (ratio between the number of correctly predicted contacts and the number of total predicted contacts) and improvement ratio over random prediction (ratio between accuracy and the probability of predicting a contact by chance) decrease with the increase of the number of analyzed contacts [92,93,156]. Table 2.2.4 shows accuracy and improvement ratio over random prediction for α =0.5 (weight for WET matrix prediction when for DRY is 1), which corresponds to the average best accuracy obtained for different numbers of analyzed predicted contacts. The results obtained for other α values followed the same trend (data not shown). Independent of the number of analyzed contacts the best predictions in average did not correspond to $\alpha = 0$. The obtained values for accuracy and improvement ratio over random prediction are within the ranges obtained by other correlated mutations approaches [97,94]. However, direct quantitative comparison of these methods is not appropriate because of their substantial differences in their residue contacts definitions. In particular, some of these approaches utilize for contact definition (see contact definition in Methods section) a chosen distance cut-off of 6-8 Å between atoms [92,97,163], whereas we use physico-chemical properties of protein residues, which results in a ≤ 4 Å cut-off [27].

Table 2.2.3. Dataset used for intradomain contact predictions.

PFAM ID	PDB ID[a]	Res, Å	N[b]	% iden[c]	L[d]	Ran acc[e]	Acc[f]	R[g]	Optimal α [h]	$X_{d\ dry}$ [i]	Opt $_{Xd}$ α [j]	$X_{d\ wet\|opt\ \alpha}$ [k]
PF00014	6pti	1.70	151	33	52	0.096	0.346	3.61	1	9.37	1	11.16
PF03705	1af7	2.00	85	20	57	0.081	0.241	2.65	0.5, 4, 10	6.14	2	7.63
PF00062	5lyz	2.00	22	46	127	0.043	0.078	1.91	0, 0.5	2.68	0	2.68
PF00018	1bu1	2.60	61	28	56	0.088	0.357	4.06	0.5	12.99	0	12.99
PF03900	1pda	1.76	21	25	74	0.062	0.237	3.82	2	9.18	0.2	9.99
PF00034	1ctj	1.10	35	17	89	0.061	0.250	4.10	1	9.13	0.1	10.34
PF01568	1dmr	1.82	88	18	113	0.044	0.050	1.14	0.2, 0.5	10.62	2	12.53
PF00127	8paz	1.60	31	29	89	0.055	0.102	1.85	2	0.50	1	4.82
PF01814	2mhr	1.30	295	12	49	0.098	0.400	4.08	0.5, 2	8.39	2	13.14
PF00017	1bmb	1.80	59	28	93	0.058	0.212	3.66	0-0.5	5.98	1	8.37
PF01320	1ayi	2.00	45	47	86	0.056	0.233	4.15	0.2	16.04	0	16.04
PF08666	1ame	1.65	171	14	66	0.074	0.273	3.69	0	10.25	0	10.25
PF01337	1a19	2.76	30	25	89	0.065	0.178	2.87	0, 0.1	4.55	0.1	4.72
PF00595	2hb2	2.30	56	19	85	0.062	0.233	3.75	0.5-2	10.16	1	11.67
PF00531	1wmg	2.10	92	14	82	0.066	0.250	3.79	0-0.5	7.67	0.2	7.95
PF00397	1eg3	2.00	73	32	30	0.143	0.467	3.26	2-20	6.59	2	8.81
PF01335	2f1s	1.40	40	21	76	0.072	0.237	3.88	0.1, 0.2	5.66	0.2	5.96
PF00619	1cy5	1.30	61	16	85	0.066	0.209	3.43	0.2-2	5.09	2	9.42
PF02213	1syx	2.35	112	28	58	0.083	0.241	2.91	0.5-2	7.37	0.5	7.77
PF05743	1uzx	1.85	28	27	118	0.035	0.068	1.98	0.1	7.22	0	7.22
PF00536	1b4f	1.95	69	28	74	0.076	0.395	5.19	0.2-2	15.53	2	16.36
PF03114	1zww	2.30	29	19	195	0.021	0.074	3.53	0.2	2.41	20	3.99
PF00169	1nty	1.70	139	10	112	0.050	0.071	1.43	0, 0.2, 0.5	5.46	2	7.53
PF08416	1wvh	1.50	49	28	132	0.040	0.106	2.65	2, 4	0.53	0.1	1.24
PF01981	1wn2	1.20	69	43	116	0.049	0.172	3.52	0.1-0.5	7.63	20	12.38
PF03992	1xbw	1.90	116	15	65	0.068	0.125	1.84	0.5	3.34	0	3.34
PF00907	1h6f	1.70	23	49	183	0.032	0.033	1.03	All the same	3.30	2	6.03
PF02237	1wpy	1.60	47	21	48	0.094	0.167	1.77	0.5-2	-2.83	0.5	0.22
PF08031	2axr	1.98	64	34	34	0.135	0.235	1.74	0.1, 0.2	-0.05	2	3.37
PF02861	1k6k	1.80	165	21	51	0.098	0.440	4.49	1, 4, 10, 20	9.55	20	13.21
PF02834	1vgj	1.94	106	14	85	0.048	0.119	2.48	4-20	-0.51	4, 10	3.21
PF01423	1kq1	1.55	128	23	60	0.079	0.167	2.11	0.2, 0.5	5.78	0.1, 0.2	7.14
PF01472	1as0	1.80	106	24	78	0.058	0.128	2.21	1-20	3.57	2, 4	11.45
PF01909	1no5	1.80	119	14	91	0.059	0.133	2.26	0.1-1	4.97	0.2	6.01
PF09261	1r34	1.95	79	31	78	0.069	0.205	2.97	0.1, 0.2	4.87	0.1	6.64
PF01315	1vlb	1.28	28	19	117	0.041	0.207	5.05	1, 2	7.70	2	10.28
PF04545	1ku3	1.80	128	31	54	0.096	0.370	3.86	0, 0.1, 1, 10, 20	12.37	10, 20	12.76
PF00984	1mv8	1.55	24	17	98	0.048	0.184	3.83	0.5-20	8.27	0.2	9.78
PF01658	1u1i	1.90	20	31	105	0.049	0.096	1.96	0.1-20	1.93	0.5	6.28
PF00745	1gpj	1.95	34	23	99	0.048	0.100	2.08	0.1-0.5	3.17	0.1	4.17
PF03099	1wnl	1.60	65	14	117	0.043	0.121	2.81	0	13.7	0.2	14.20
PF01985	1jo0	1.37	50	23	84	0.064	0.167	2.60	0-0.2	6.96	0	6.96
PF08436	1q0q	1.90	77	57	94	0.049	0.213	4.34	0-0.1	6.91	10	10.15
PF02881	1jpn	1.90	52	19	85	0.063	0.119	1.89	All the same	3.94	2	5.78
PF01966	1ynb	1.76	158	12	91	0.057	0.333	5.85	0-0.2	-0.79	2	2.20
PF00191	1yii	1.42	178	28	66	0.076	0.273	3.59	0-0.2	-0.35	10	1.05
PF00317	1xje	1.90	79	23	90	0.056	0.178	3.17	0.5-2	10.01	0.5	13.16
PF00046	1puf	1.90	184	37	60	0.082	0.333	4.07	1, 2	6.07	2	8.60
PF00077	5fiv	1.90	48	27	108	0.049	0.093	1.89	2	-1.37	1	3.63
PF00042	1ecn	1.40	73	18	101	0.046	0.163	3.56	1, 2	6.89	2	7.19

[a]PDB ID; [b]Number of sequences; [c]Average sequences pairwise similarity (%); [d]Reference sequence length; [e]Random accuracy; [f]Accuracy for optimal α; [g]Improvement ratio over random prediction for optimal α; [h]Values for α=0; [i]α corresponding to the highest accuracy; [j]α corresponding to the highest X_d; [k]X_d highest value.

Table 2.2.4. Prediction parameters dependence on the number of analyzed contacts

Number of analyzed predicted contacts	Accuracy	Improvement ratio over random prediction
L	0.15±0.09	2.24±0.95
L/2	0.18±0.10	2.67±1.08
L/3	0.19±0.12	2.81±1.52
L/5	0.21±0.16	3.16±1.79
L/10	0.23±0.20	3.55±2.81

L is the length of the reference sequence. The value $\alpha = 0.5$ has been used.

We compared the dependences on α of: i) accuracy, ii) improvement ratio over random prediction, iii) number of correctly predicted contacts (C_{corr}); and, since our dataset is heterogeneous (see high standard deviations in Table 2.2.4), we normalized these parameters by the corresponding values at α=0 (wet prediction ratio). For the purpose of wet prediction ratio comparison at different values of α we found L/2 to be the most appropriate number of contacts. This choice is explained by the fact that the changes in prediction results influenced by α variation become hardly detectable if a smaller number of contacts (C_{total}) is considered for analysis since these changes are limited by low values of C_{total} and, consequently, of correctly predicted contacts (C_{corr}). On the other hand, the increase of C_{total} generally leads to decrease of prediction accuracy and to negligible differences in prediction results corresponding to different α values. Only in 2 out of the 50 families of our dataset best predictions correspond to α= 0 values (Table 2.2.3).

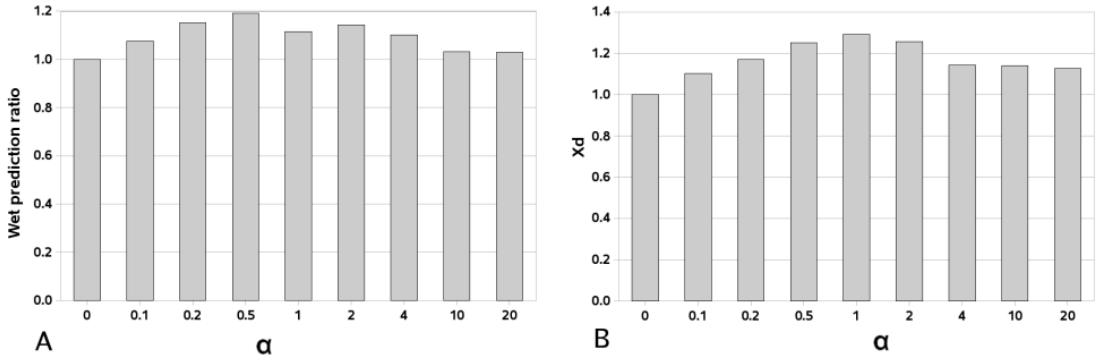

Figure 2.2.3. Dependence on α of relative prediction characteristics for the intradomain dataset. A) Wet prediction ratio. B) Relative harmonic weighted difference statistic (X_d).

Maximum values for wet prediction ratio and relative X_d (harmonic weighted difference statistic) averaged for the whole dataset are obtained when α=0.5 and α=1 (1.19 and 1.29, respectively; Figure 2.2.3 A, B). This means that, for these values of α, introduction of the WET similarity matrix improves prediction by 20-30% on average. Noticeably, the high values of $\alpha \in \{10, 20\}$ still make the

predictions on average better than by the single use of the DRY matrix. For optimal value α=0.5, absolute values of accuracy and improvement ratio over random prediction averaged for all 50 families increase by 1.4% and 0.19, respectively, in comparison to the single use of the DRY similarity matrix. For each family in the dataset there is an essentially higher increase of accuracy and improvement ratio over random prediction than on average. In some families, wet prediction ratio is improved more than twice (reference structures 1AF7, 1PDA, 8PAZ, 1DMR, 1AS0) and even 4.5 times (reference structure 1WVH) when $\alpha > 0$. Our results show a significant improvement (20-30% of increase in wet prediction ratio) in predictions by the introduction of the WET similarity matrix in comparison to the single use of the DRY matrix within a correlated mutations approach. We observe that for sequence separations |i-j| >6, 12, 24 our results follow the same trend. The obtained results for α=0.5 for different number of contacts (L, L/2, L/3, L/5, L/10) are shown in Table 2.2.5. We observe that the best predictions correspond to α=0.2 and 0.5 for most of sequence separation values and number of contacts. Wet prediction ratios for the whole range of analyzed α are presented in figure 2.2.4). In all cases, independently of sequence separation and number of contacts, the best predictions correspond to $\alpha > 0$.

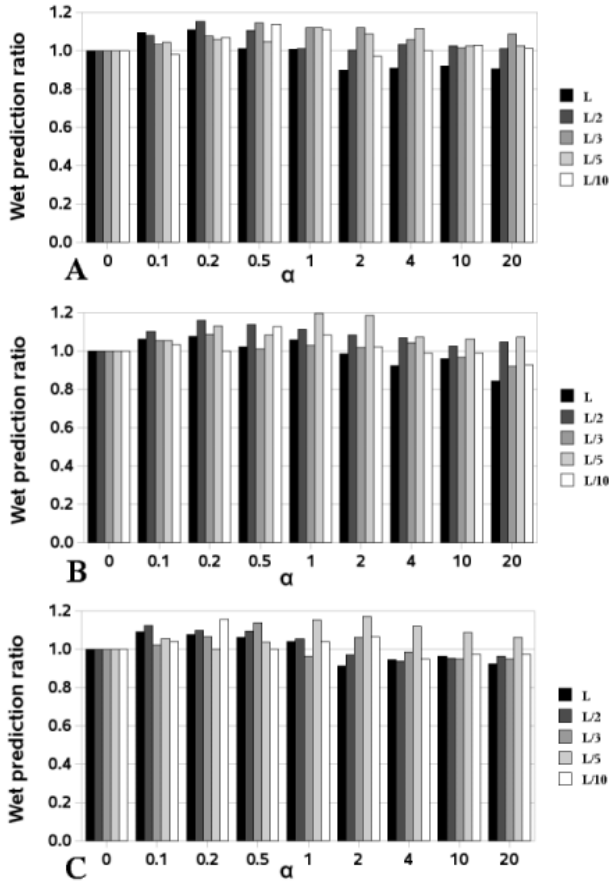

Figure 2.2.4. Dependence on α of wet prediction ratio for the intradomain dataset with sequence separation: A) 6. B) 12. C) 24.

Table 2.2.5. Accuracy, improvement ratio over random prediction and wet prediction ratio for different sequence separations

	Sequence separation=6			Sequence separation=12			Sequence separation=24		
Contacts number	Accuracy	R	Wet ratio	Accuracy	R	Wet ratio	Accuracy	R	Wet ratio
L	0.061	3.07	1.01	0.051	3.02	1.02	0.042	2.97	1.06
L/2	0.079	4.18	1.11	0.070	4.34	1.14	0.050	3.76	1.10
L/3	0.087	4.56	1.14	0.071	4.49	1.01	0.060	4.61	1.14
L/5	0.099	5.49	1.05	0.085	5.71	1.08	0.068	5.18	1.04
L/10	0.122	6.68	1.14	0.103	6.89	1.13	0.078	6.31	1.00

L is the length of the reference sequence. R is improvement over random prediction. The value $\alpha = 0.5$ has been used.

Table 2.2.6. Dataset used for interdomain contact predictions

Interacting partners	PFAM ID Family 1	PFAM ID Family 2	PDB ID[a]	N[b]	iden[c]	L_1[d]	L_2[e]	$X_{d\,dry}$[f]	Optimal$_{Xd}\alpha$[g]	$X_{d\,wet}\|_{opt\,\alpha}$[h]
Tyrosine kinase SH3 and SH2 domains	PF00018	PF00017	2src	19	35	57	83	1.86	0.2	3.25
Alcohol dehydrogenase N- and C-domains	PF08240	PF00107	1adg	89	23	128	143	3.52	0.2	3.64
Mg superoxide dismutase N- and C-domains	PF00081	PF02777	1ap5	23	44	82	107	4.76	0.2	5.04
Immunoglobulin heavy and light chains	PF00047	PF00047	12e8	116	36	107	114	13.56	0	13.56
Ortnithine transferase N- and C-domains	PF02729	PF00185	1duv	20	30	142	178	4.47	0.1	4.94
NFKB factor RHD and TIG domains	PF00554	PF01833	1svc	21	40	199	100	4.56	0.5	4.62
STAT alpha and binding domains	PF01017	PF02864	1bf5	32	38	180	251	4.30	0.2	4.42
Mur-ligase catalytic and C-terminal domains	PF01225	PF08245	1e8c	26	25	82	208	1.84	0.1	2.12
Dynamin central and N-domains	PF00350	PF01031	2aka	32	40	174	89	0.04	0.2	0.14
Trk C- and N-domains	PF02254	PF02080	1lnq	42	20	114	72	0.53	1	0.78

[a]PDB ID of the reference structure; [b]Number of sequences in the multiple sequence alignment; [c]Average percentage of sequences pairwise similarity; [d,e]Lengths of the reference sequences; [f]Values for α=0; [g]α value corresponding to the highest X_d; [h]X_d highest value.

Interdomain contacts prediction. The interdomain dataset used for our studies consisted of 10 different pairs of interacting domains (Table 2.2.6). From the analysis of the $(L_1+L_2)/2$ predicted interdomain residue contacts (L_1 and L_2 are the lengths of the sequences in each of the two domains) we

observed that in 9 out of 10 cases best predictions in terms of X_d were obtained when both the WET and DRY matrices were used. Relative X_d averaged for the whole dataset reaches a maximum value of 1.32 at α=0.2 and then decreases with the further increase of α (Figure 2.2.5). In one of the examples (SH2-SH3 domains interaction) the differences of distance distributions for different α values are dramatic (Figure 2.2.6). In this case the X_d value for predicted contacts at α=0 and α=0.2 changes almost twice (Table 2.2.6). These results point out that the use of the WET similarity matrix might improve the statistic X_d in comparison to the single use of the DRY similarity matrix.

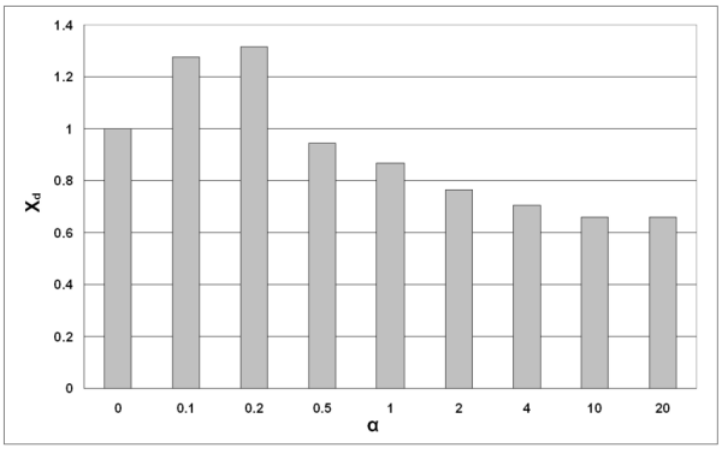

Figure 2.2.5. Predictions for interdomain dataset. Relative harmonic weighted difference statistic (X_d) dependence on α.

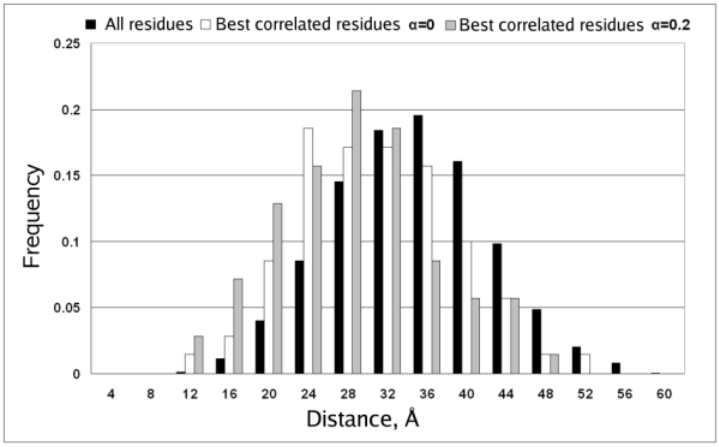

Figure 2.2.6. Proportion of residue pairs at distance bins for the interaction SH2-SH3.All residue pairs are shown in black, correlated pairs with α=0 in white, and correlated pairs with α=0.2 in grey. Reference structure used is PDB ID 2src.

Dependence of relative average X_d on α for interdomain contacts prediction (Figure 2.2.5) resembles the one obtained for intradomain prediction (Figure 2.2.3) but they differ in the optimal α and in the X_d corresponding to the higher α values. While in predictions of intradomain contacts all values of $\alpha > 0$ lead to the improvement of contact predictions, in the case of interdomain contacts

prediction the use of the WET similarity matrix yields higher X_d than the DRY alone when $\alpha \in \{0.1,0.2\}$. This might be due to the differences in distance distributions between the analyzed pairs of residues, which are closer to each other in the case of intradomain contacts. Nevertheless, introduction of the WET similarity matrix improves contact prediction compared to the single use of the DRY similarity matrix for both intra- and interdomain contacts. Although there are still significant limitations for practical use of the correlated mutations approach for interdomain contacts prediction, also mentioned by other authors [93,158], we believe that consideration of water by the use of "wet" similarity matrices could improve the results obtained by correlated mutations approaches.

2.2.5 Conclusions

This study is the first investigating the impact of inclusion of solvent into the concept of correlated mutations. With this work we further demonstrate our previous observations that relations between solvent and protein residues in protein interfaces differ from those in the whole protein. Recent work on bond preferences in inter- versus intraprotein interactions highlights the different architecture of protein interfaces and their unique bond preferences [172].

Two similarity matrices have been used in this work: the McLachlan matrix as the DRY similarity matrix and a WET similarity matrix derived by statistical analysis of the frequency of water contacts by residue type in protein interfaces in the whole PDB. Analysis of the DRY and WET similarity matrices shows that they are interdependent for some residue types, which could be explained by physico-chemical properties of individual amino acid residues. We analyze two datasets containing 50 domains and 10 domain pairs belonging to PFAM families. We sum the predictions obtained by the use of both matrices with different weight coefficients and find optimal combinations for best predictions. Our datasets are heterogeneous to propose one best weight value to be able to apply the optimized method to all domain families; however, the prediction of contacts obtained by the introduction of the WET similarity matrix is improved for most of the families in the datasets (for both intra- and interdomain) as well as on average (by 20-30%). Our analysis of solvent impact on contact prediction in proteins suggests that further development of the correlated mutations concept would benefit from taking into account solvent as an active participant in protein-protein interactions, which is usually overlooked in these studies.

CHAPTER 3

3.1 Characterization of fluorinated amino acids by QM and MD approaches

by Sergey Samsonov, Mario Salwiczek, Gerd Anders, Beate Koksch, and M. Teresa Pisabarro

Submitted to Proteins: Structure, Function, and Bioinformatics in 2009

3.1.1 Abstract

Non-canonical amino acids with newly designed side-chain functionalities represent powerful tools to improve structural, biological, and pharmacological properties of peptides and proteins. In this context fluorinated amino acids have increasingly gained importance. Despite the current wide use of fluorination in protein engineering, the basic properties of fluorine in protein environments are still not completely understood. Our aim has been to characterize the physico-chemical properties of fluorinated amino acids by using quantum mechanics (QM) and molecular dynamics (MD) approaches. We have analyzed geometry, charges and hydrogen bonding abilities of several ethane fluorinated derivatives at different QM theory levels, and have used them as simplified models for fluorinated amino acid sidechains. We have parametrized four fluorinated L-amino acids for the AMBER MD package: 4-monofluoroethylglycine (MfeGly), 4,4-difluoroethylglycine (DfeGly), 4,4,4-trifluoroethylglycine (TfeGly) and 4,4-difluoropropylglycine (DfpGly). We have characterized them in terms of molecular volumes, conformational preferences and hydration properties. The obtained results illustrate that fluorine and hydrogen atoms of fluoromethyl groups could be potential acceptors and/or donors of weak hydrogen bonds in protein environments. Hydration of the studied fluorinated amino acids was found to be more favorable than for their non-fluorinated analogues, and hydrophobicity was observed to increase with the number of fluorine atoms, which is in accordance with the experimental retention times we obtain for these amino acids. This study broadens our understanding on the properties of fluorine within protein environments, which is important in order to exploit the full potential of fluorine's unique properties for applications in the field of protein engineering.

3.1.2 Introduction

Because of its unique physico-chemical properties fluorinated compounds have gained a lot of interest in organic chemistry and biochemistry fields [173]. Impact of fluorination on bioactivity of substrates, inhibitors and catalysts has been monitored in a number of biochemical studies [174-177]. In protein chemistry, several studies on folding of polypeptides containing fluorinated amino acids [178-180], as well as on protein subunits association [181], and protein design [182] have suggested

that the use of fluorinated amino acids offers remarkable opportunities for improvement of structural, thermodynamic and kinetic properties of the engineered proteins [183,184]. This utility of fluorine could be attributed to its basic properties. Fluorine is highly electronegative, which makes its covalent bonds strongly polarized so that their contribution to the electric dipole interaction with the surrounding environment is very high. This also leads to polarization of the chemical bonds adjacent to fluorine covalent bonds. At the same time, this high electronegativity of fluorine results in its very poor ability of being electronic pair donor [185]. Furthermore, fluorine is the second smallest atom after hydrogen and almost isosteric to oxygen (Bondi radii of 1.47 Å, 1.20 Å and 1.57 Å, respectively) [186]. The possibility of substitution of these atoms by fluorine in different chemical groups is very powerful for drug design [187] and has a high potential in protein engineering. It is widely assumed that fluorine is a weak hydrogen bond acceptor [188,174]. Nevertheless, many studies have demonstrated the structural importance of the hydrogen bonds formed by fluorine atoms for establishing intra- and intermolecular interactions [189-191]. Though some detailed studies on hydrogen bonding abilities of fluorinated small molecules as methane derivatives [192,193] and 2-fluoroethanol [194] have been carried out, the basic properties and characteristics of fluorine-mediated hydrogen bonds in protein environments are still not clear. Fluorinated amino acids have not been yet parametrized for forcefields implemented in the currently most widely used MD packages like AMBER [8], CHARMM [195] and GROMACS [196]. In our previous studies fluorinated derivatives of ethylglycine (difluoroethylglycine, trifluoroethylglycine and difluoropropylglycine) have been incorporated into a α-helical coiled-coil system and their impact on the stability of coiled-coils interaction has been analyzed by using CD spectroscopy [178]. We concluded that there are two major effects of fluorine substitution on the behavior of the system: increased polarity and a change in steric demand of fluoromethyl groups. However, these experiments were not able to characterize the physico-chemical properties of non-standard amino acids at basic level of theory and to look in details at each of these two observed effects independently.

We have used a QM approach at several levels of theory to: i) characterize in terms of charge, geometry and size several fluorinated ethane derivatives as simplified models for fluorinated amino acid side-chains, ii) analyze their hydrogen bonding properties and compare them with canonical amino acids, iii) derive RESP charges for fluorinated amino acids used in libraries for the AMBER force field. We have also used a MD approach to carry out conformational analysis for fluorinated amino acids. We have compared the hydration properties of the fluorinated amino acids with non-fluorinated analogues

using free energy perturbation calculations, RDF-function and retention time experiments.

Our data show that fluoromethyl groups could play role of hydrogen bond acceptors and/or donors, though they are in general weaker than in canonical amino acids. The created fluorinated amino acids libraries, compatible with the AMBER force field, were used to obtain Ramachandran plots and rotamer libraries. The data on hydration of fluorinated amino acids, obtained both theoretically and experimentally, show that their hydrophobicity increases with the number of fluorine atoms and is lower than of their non-fluorinated analogues. The results obtained in our study contribute to widening our understanding on the properties of fluorine within protein environments, which is essential to be able to exploit the full potential of fluorine's unique properties for applications in the field of protein engineering.

3.1.3 Methodology

Quantum chemical calculations. We used ethane derivatives CH_3CH_3, CH_3CH_2F, CH_3CHF_2, $CH_3CF_2CH_3$, CH_3CF_3 for modeling fluorinated amino acid side-chains. The molecules were optimized at four levels of theory: 1) HF; 2) MP2 perturbation [102], proved to agree well with experimental data for fluoroethanol conformers [194]; 3) density functional theory BLYP [197,198] and 4) DFT/HF hybrid functional B3LYP [199], which has been shown to be one of the best choices for amino acid systems, especially rich in hydrogen bonds [200]. All methods were provided in GAMESS (US) [201]. Geometry optimizations employed the 6-311G**++ basis set. Inclusion of diffuse components is an obligatory requirement for hydrogen bond calculations [202]. Hydrogen bonding abilities of CH_3CH_3, CH_3CH_2F, CH_3CHF_2, $CH_3CF_2CH_3$, CH_3CF_3 were analyzed at the same theory levels. For comparative analysis of hydrogen bonds we used ethane derivatives and other hydrogen bond donors and/or acceptors typical for protein environments: water, amideD ($CH_3CONHCH_3$), imidazole, hydroxyl (CH_3CH_2OH), indole, methanethiol (CH_3SH), amideA (CH_3CONH_2), ketone (CH_3COCH_3) and furan. Geometry optimizations were carried out without any atomic constraints with an initial configuration obtained by minimization with the AMBER ff99 force field implemented in the MOE program [203]. We analyzed hydrogen bond donor-acceptor pairs that were interacting only via hydrogen bonds in order to minimize the impact of other types of interactions in our analysis. The following properties characterizing hydrogen bonds were measured: H-bond energy (difference between the energy of the hydrogen-bonded complex and the sum of the energies of the isolated binding partners) (E), BSSE corrected energy using the counterpoise correction approach (E_{BSSE}) [104], hydrogen bond length (d),

hydrogen bond angle (A-H-D), shift of the D-H bond (Δr), and the charge transfer on acceptor and hydrogen atom (Δq(A), Δq(H)).

Creation of non-canonical amino acids libraries for AMBER. The non-canonical L-amino acids ethylglycine (Abu), 4-monofluoroethylglycine (MfeGly), 4,4-difluoroethylglycine (DfeGly), 4,4,4-trifluoroethylglycine (TfeGly) and 4,4-difluoropropylglycine (DfpGly) and propylglycine were parameterized to be compatible with the ff03 AMBER force field [204] using a standard procedure implemented in the R.E.D. III program, which we used for RESP charge calculations [128]. Charges were derived for each amino acid in two conformations (helical and extended) with the *ab initio* Hartree-Fock method HF/6-31G* using GAMESS-US. We used the ff03 AMBER force field [204] with standard atomic parameters for fluorine derived from liquid simulations [205].

Volume and solvent accessible surface area calculations. For these calculations we used ethane derivatives optimized in B3LYP (6-311G**++) and amino acids from the created AMBER libraries (see above). Their volumes and solvent accessible surface areas (ASA) were calculated in Discovery Studio 1.7 [206] using VDW surface calculation. In the case of ethane derivatives, the volumes for fluoromethyl groups were calculated as the difference between the volumes for the entire molecules and a half of the ethane molecular volume. Volumes and ASA of the fluorinated amino acids side chains were calculated for the whole amino acids with subtraction of the corresponding values for Gly.

Conformational analysis. For the conformational analysis of the fluorinated amino acids we used the ff03 force field implemented in AMBER 8.0 [8]. Dipeptides Ace-FXR-Nme, where FXR corresponds to each of the studied amino acids (Abu, MfeGly, DfeGly, TfeGly, DfpGly, Ala, Val, Leu, Met, Phe) were minimized applying the pairwise generalized Born model of implicit solvent [207] in two steps: $5 \cdot 10^4$ cycles of steepest descent and $5 \cdot 10^4$ cycles of conjugate gradient with harmonic force restraints of 10^4 kcal/mol·Å on the atoms defining backbone dihedral angles (N, C, C_α) and a 10^{-4} kcal/mol·Å^2 convergence criterion. Backbone dihedral angles (φ, ψ) were varied from $-180°$ to $180°$ with a $10°$ step. The full energies calculated in the minimization were used for obtaining Ramachandran plots using Gnuplot [208]. For comparative quantitative characterization of secondary structure propensities we defined a *propensity index*, which is a covariance between the probabilities obtained from the secondary structure data from the whole PDB and the probabilities derived from our energy calculations. Using an in-house relational database of the PDB we calculated the density of dihedral angles distribution $\omega_i(\varphi, \psi)$ for each type of secondary structure elements i={α-helix, β-sheet and left α-helix}.

The probability of an amino acid to adopt a certain secondary structure conformation (*i*) was derived from energetic calculations as follows:

$$P_{(i)}(\varphi,\psi)=\exp[-(E(\varphi,\psi)-E_{min(i)})/RT] \quad (3.1.1),$$

where $E_{min(i)}$ is the energetic minimum observed in our calculations for this amino acid in this particular conformation (i). The propensity index for secondary structure type i is the defined as:

$$PI_i=\Sigma_{\varphi,\ \psi\in[-180;180]}\, w_i(\varphi,\ \psi)\cdot p_i(\varphi,\ \psi) \quad (3.1.2),$$

where:

$$w_i(\varphi,\psi)=\int_{\phi-5}^{\phi+5}\int_{\psi-5}^{\phi+5}\omega_i(\Phi,\Psi)\,d\Phi\,d\Psi \quad (3.1.3)$$

Therefore,

$$PI_i=\Sigma_{\varphi,\ \psi\in[-180;180]}\, w_i(\varphi,\psi)\cdot\exp[-(E(\varphi,\ \psi)-E_{min\,i})/RT] \quad (3.1.4)$$

A similar analysis was carried out for side-chains of the fluorinated amino acids. In this case three dihedral angles (φ, ψ, χ_1) for MfeGly, DfeGly, TfeGly and four dihedral angles (φ, ψ, χ_1,χ_2) for DfpGly were restrained as described above. χ_1 was defined by the N, C_α, C_β, C_γ atoms and χ_2 by C_α, C_β, C_γ, C_δ. The used scanning step for χ_1 and χ_2 was 5°, while φ, ψ were fixed at the energy minima values for α-helix, β-sheet and left α-helix areas in the Ramachandran plot. Similar approaches have been previously reported for both canonical and non-canonical amino acids [209,210].

Hydration energy calculations. We used free energy perturbation for the calculation of the hydration energy similarly to the work of Pendley *et al*. [211]. For the MD run in solvent the dipeptide Ace-FXR-Nme was placed in an octahedral TIP3 water box, the fluorine atoms were perturbed to hydrogen atoms using 12 values of thermodynamic integration parameter λ (0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, and 0.99078) with the following Gaussian integration. MD simulations were preceded by two energy-minimization steps: 500 cycles of steepest descent and 1000 cycles of conjugate gradient with harmonic force restraints on all dipeptide atoms, followed by 1000 cycles of steepest descent and 1500 cycles of conjugate gradient without constraints. This was followed by heating of the system from 0 to 300K for 10 ps. MD calculations were done at 300K and 10^6 Pa in isothermal isobaric ensemble (NPT). Periodic boundary conditions in NPT ensemble with Langevin temperature coupling with collision frequency parameter γ=1 ps^{-1} and

Berendsen pressure coupling with a time constant of 1.0 ps were applied. The SHAKE algorithm was used to constrain all bonds containing hydrogen atoms. A 2 fs time integration step was used. An 8 Å cut-off was applied to treat non-bonded interactions, and the Particle Mesh Ewald (PME) method was introduced for long-range electrostatic interactions treatment. MD trajectories were recorded each 2 ps. We analyzed the last nanosecond of a 2 ns MD trajectory. For *in vacuo* calculations, a 5 ns MD run was preceded by 1000 cycles of minimization. The same values for λ were used as in the explicit solvent simulations. The last three nanoseconds of the MD run were considered for analysis. The equilibration of $dV/d\lambda$ for solvated and *in vacuo* simulations is shown in Supplementary File 1.

The difference between the hydration energies (E_{hyd}) of fluorinated amino acids (f) and their non-fluorinated (nf) analogues is obtained from the thermodynamical cycle and is calculated using thermodynamic integration along the perturbation path:

$$\Delta E_{hyd,f} - \Delta E_{hyd,nf} = \Delta E_{pert\ in\ vacuo} - \Delta E_{pert\ in\ water} \quad (3.1.5)$$

The analysis of hydration in terms of the radial distribution function (RDF) was done with the PTRAJ module of AMBER 8.0.

Amino acids retention time calculations. An HPLC-assay was developed that aimed at investigating the impact of fluorination on the correlation of amino acid side chain volume and hydrophobicity. We used the Fmoc-protected analogues of Gly, Abu, Val, Leu, Ile as well as of MfeGly, DfeGly, TfeGly, and DfpGly and determined their retention times on a C18 column (Capcell C18, 5 µm). Approximately 10 µmol of the respective Fmoc-amino acid were dissolved in 5 mL of a mixture of 40 % acetonitrile (99.9%, HPLC gradient grade, Acros Organics) in deionized water containing 0.1% TFA (Uvasol®, Merck) and filtered over 0.2 µm. The retention times on the C18 column were determined. A linear gradient from 40% to 70% in 30 minutes was applied at room temperature and all experiments were performed as triplicates.

Statistical analysis. Statistical analysis of the data was carried out with the R-package [144].

3.1.4 Results and discussion

Geometry optimization of the fluorinated ethane derivatives. We optimized the geometry of ethane and its fluorinated derivatives CH_3CH_2F, CH_3CHF_2, CH_3CF_3, $CH_3CF_2CH_3$ at 4 levels of theory. The obtained values for charges and bond lengths are shown in Figure 3.1.1. Because of the fluorine atom being highly electronegative, fluorine substitution polarizes the C-F bond and the electronic density is partially drawn from carbon to fluorine. As the number of fluorine atoms increases, the charge of the

carbon atom (C2) bonded to fluorines increases by ~1 charge unit and becomes positive. As a consequence, the charge of C1 (the carbon atom not attached to fluorine) decreases by ~0.1 units. That is followed by polarization of the hydrogens bonded both to C1 (H_{C1}) and C2 (H_{C2}). The charge of H_{C2} increases about 0.4 units and becomes comparable to the charge of hydrogens in hydrogen bond donor groups in amino acids (amide, imidazole and hydroxyl, indole groups) and water (0.23-0.34 units, Table 3.1.1). This suggests that fluorinated groups may act as hydrogen bond donors. On the contrary, additional fluorination leads to decrease of fluorine atoms absolute charge values. The charges of fluorine atoms are comparable with the ones of standard hydrogen bond acceptor atoms of amino acid groups (amide, imidazole, ketone, Table 3.1.2) as well as of water and furan oxygen atoms.

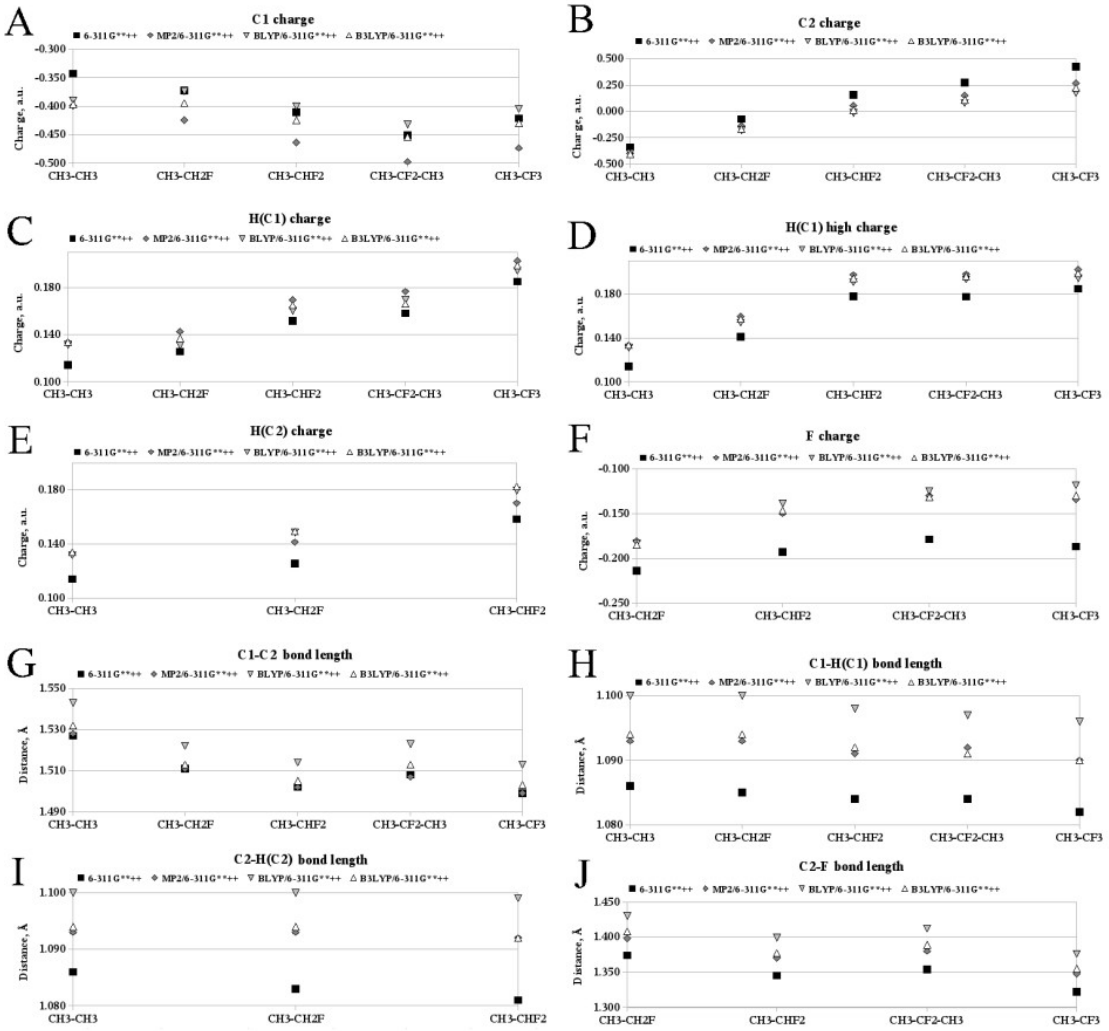

Figure 3.1.1. Charges and bond lengths in ethane and ethane derivatives. All methods were applied with the same basis set (6-311G**++). A) Charge of C1, a not-fluorinated carbon. B) Charge of C2, a fluorinated carbon. C) Charge of H(C1), hydrogen bound to C1. D) When 1 or 2 fluorine atoms are bonded to C2, the charge distribution on different H(C1) is not symmetric around the C1-C2 bond. This figure represents the highest charges on the H(C1) atoms. E) Charge of H(C2), hydrogen bound to C2. F) F charge. G) C1-C2 bond length. H) C1-H(C1) bond length. I) C2-H(C2) bond length. J) C2-F bond length.

The bond lengths change upon fluorination in the range of 10^{-2} Å for C-C and C-H bonds, and 10^{-1} Å for C-F bond. All bond lengths decrease with the increase of number of fluorine atoms, which is explained by increased polarization of the bonds. Although the electric dipole values of the different ethane derivatives do not change dramatically with fluorination, the directions of the dipole vectors are significantly distinct (Figure 3.1.2), which may have an important impact on the electrostatic properties of the molecules, containing these groups.

Table 3.1.1. Charges of hydrogen atoms in different hydrogen bond donor groups

Group/[a]Method	HF	MP2	BLYP	B3LYP
AmideA	-0.464	-0.340	-0.354	-0.379
Imidazole	-0.246	-0.177	-0.176	-0.192
Ketone	-0.362	-0.254	-0.263	-0.284
Furan	-0.143	-0.044	-0.053	-0.063

[a] All the methods were applied with the same basis set (6-311G**++).

Table 3.1.2. Charges of hydrogen bond acceptors

Group/[a]Method	HF	MP2	BLYP	B3LYP
AmideD	0.298	0.266	0.257	0.271
Imidazole	0.332	0.316	0.280	0.297
Hydroxyl	0.256	0.242	0.230	0.240
Indole	0.335	0.316	0.277	0.295
Water	0.259	0.248	0.242	0.249
CH3SH	0.027	0.033	0.037	0.039

[a]All the methods were applied with the same basis set (6-311G**++).

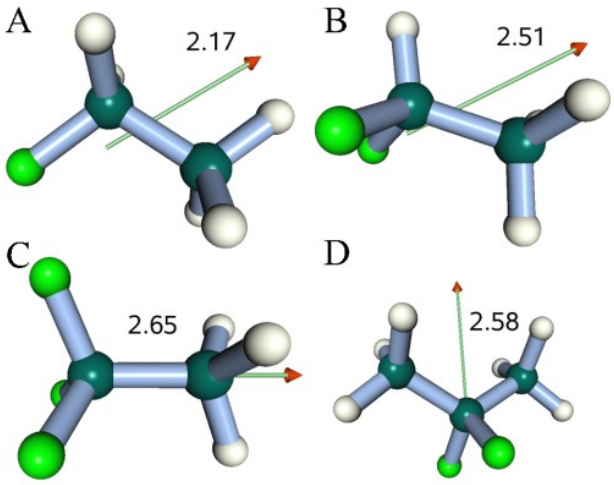

Figure 3.1.2. Electric dipoles calculated in B3LYP (6-311G**++).
A) Monofluoroethane. B) Difluoroethane. C) Trifluoroethane. D) 2,2-difluoropropane. Fluorine atoms are green, hydrogens are white, carbons are dark green. Dipoles are represented as arrows.

Hydrogen bond analysis. We carried out analysis of hydrogen bonding properties of the fluorinated ethane derivatives and compared them with hydrogen bond donors/acceptors typical in biological macromolecular systems: water, amideD ($CH_3CONHCH_3$), imidazole, hydroxyl (CH_3CH_2OH), indole, methanethiol (CH_3SH), amideA (CH_3CONH_2), ketone (CH_3COCH_3) and furan. Except for methanethiol and water, hydrogen bonding properties of these molecules have been also characterized in the work of Hao *et al.* [212]. However, direct comparison of the results of that work with our results is not feasible since the used methodology of the geometry optimization of hydrogen bonded complexes is substantially different.

All calculated data for hydrogen bonding pairs of donors and acceptors is represented in Tables 3.1.3-3.1.8. The data for the fluorinated ethane derivatives as hydrogen bond acceptors in comparison with the rest of analyzed acceptors are summarized in figure 3.1.3. The analysis of this data allows to rank these acceptors by hydrogen bonding properties in the following order: imidazole>amideA>ketone>CH_3CH_2F>furan~CH_3CHF_2~$CH_3CF_2CH_3$>CH_3CF_3. Obviously, the acceptor properties of the fluoromethyl group become less pronounced with the increase of number of fluorine atoms. All these analyzed hydrogen bonds yield positive D-H bond length shift (red-shifted hydrogen bonds) except for some bonds with methanethiol, which is a very weak hydrogen bond donor. This blue-shifting could be explained in terms of the redistribution of electronic density from both atoms (D-H) to a very electronegative fluorine atom [213]. The hydrogen bond energy, bond length and D-H bond length shift for the fluoromethyl groups are similar to the ones of furan and characterize the bonds as essentially weaker as the hydrogen bonds formed by imidazole, amideA or ketone acceptors. Directionality of the hydrogen bonds formed by the atoms of fluoromethyl groups is significantly broader, and the average angles have lower values compared to other analyzed groups. Hydrogens charge transfer in the formation of hydrogen bonds is also lower for fluorinated groups, while acceptor atom charge transfer is slightly positive and similar to furan and ketone groups. The strength of hydrogen bond decreases with additional fluorination of the group in terms of energy, bond length, and D-H bond length shift, but not in terms of charge transfer, which remains roughly the same for all fluoromethyl groups. These relatively low values for fluorine atom charge transfer are explained by the high electronegativity of fluorine atoms [185]. We also have analyzed hydrogen bonds formed by CH_3CHF_2 and $CH_3CF_2CH_3$ with water molecules where each water hydrogen atom interacts with a fluorine atom (Table 3.1.9). These interactions are energetically less favorable than in the case of only

one water hydrogen atom establishing a hydrogen bond with just a fluorine. To sum up, hydrogen bond acceptor properties of fluoromethyl groups are weak in comparison to typical acceptors in protein environments.

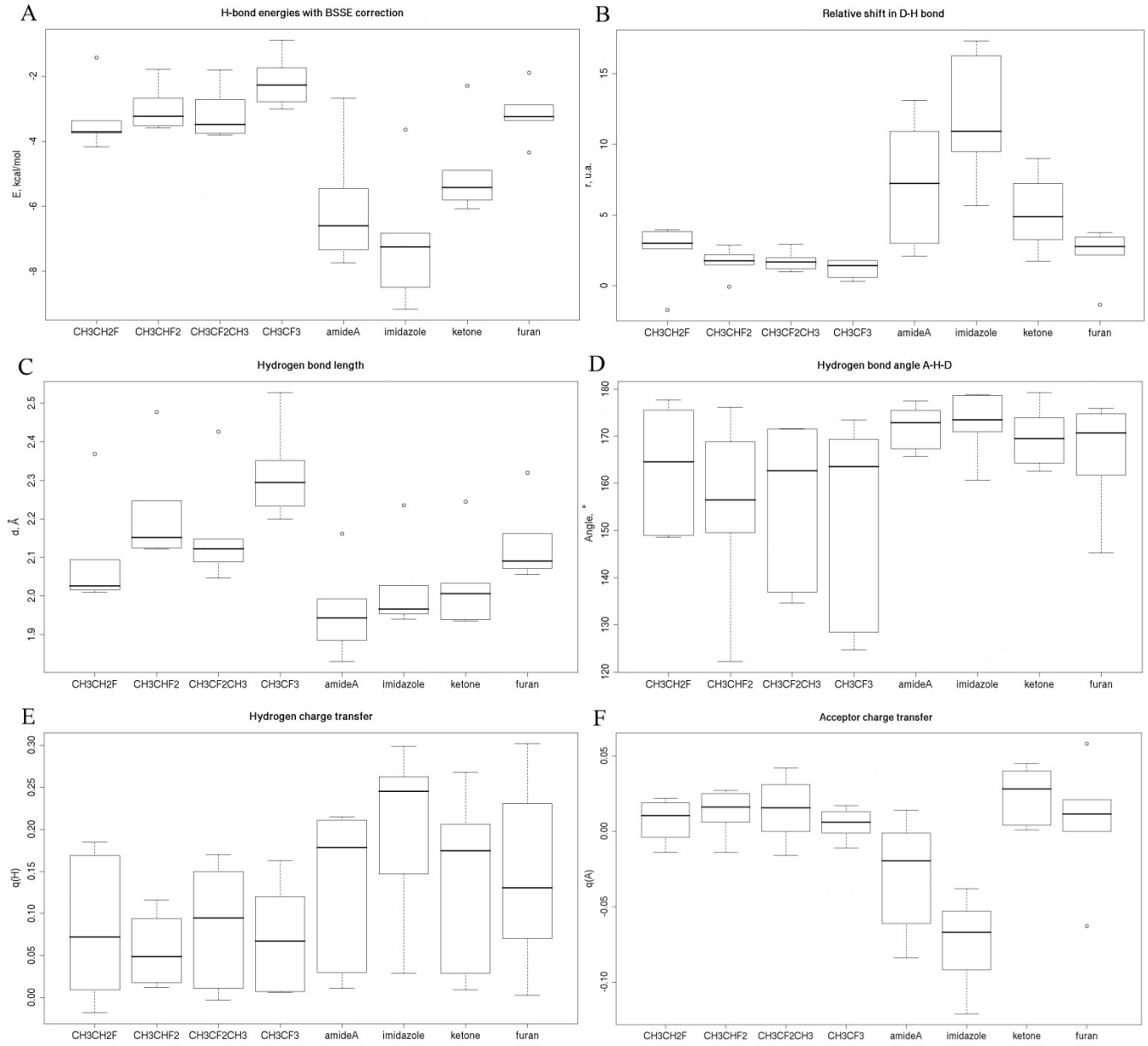

Figure 3.1.3. Hydrogen bond acceptors properties. The box plots contain data on all analyzed hydrogen bonds calculated in B3PLYP (6-311G**++). A) Hydrogen bond energy with BSSE correction. B) Relative shift in the covalent bond between hydrogen and heavy atom upon hydrogen bond formation. C) Hydrogen bond length. D) Hydrogen bond angle ∠(D-H-A). E) Hydrogen atom charge transfer. F) Acceptor atom charge transfer.

Table 3.1.3. Hydrogen bond acceptor properties of the fluorinated ethane derivatives. Hydrogen bond energies

	Donor	Water		AmideD		Imidazole		Hydroxyl		Indole		CH_3SH	
[a]Method	Acceptor	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}
HF	CH_3CH_2F	-3.43	-3.22	-2.71	-2.66	-3.87	-3.66	-3.15	-2.90	-3.26	-3.04	-1.35	-1.09
	CH_3CHF_2	-3.19	-2.92	-2.12	-1.95	-3.22	-2.92	-2.59	-2.20	-2.88	-2.49	-1.32	-0.97
	$CH_3CF_2CH_3$	-3.32	-3.01	-2.05	-1.93	-3.22	-2.94	-2.70	-2.29	-3.18	-2.73	-1.17	-0.85
	CH_3CF_3	-2.65	-2.32	-1.49	-1.26	-2.07	-1.75	-2.64	-2.19	-1.88	-1.38	-0.87	-0.52
MP2	CH_3CH_2F	-4.73	-3.67	-4.41	-3.37	-4.88	-4.17	-4.93	-3.74	-4.59	-3.74	-2.55	-1.43
	CH_3CHF_2	-4.43	-3.41	-3.60	-2.68	-4.58	-3.59	-4.38	-3.05	-5.02	-3.52	-3.73	-1.79
	$CH_3CF_2CH_3$	-4.95	-3.75	-3.58	-2.72	-4.62	-3.72	-4.66	-3.23	-5.26	-3.81	-3.93	-1.81
	CH_3CF_3	-3.95	-2.78	-2.80	-1.74	-3.12	-2.29	-4.20	-3.00	-3.48	-2.24	-2.44	-0.89
BLYP	CH_3CH_2F	-3.83	-3.41	-2.71	-2.60	-3.74	-3.45	-3.40	-2.93	-3.04	-2.97	-1.46	-0.99
	CH_3CHF_2	-3.07	-2.69	-1.93	-1.74	-2.93	-2.64	-2.69	-2.21	-2.80	-2.55	-1.11	-0.72
	$CH_3CF_2CH_3$	-3.55	-3.01	-1.94	-1.73	-3.44	-2.99	-2.82	-2.19	-3.25	-2.68	-1.12	-0.68
	CH_3CF_3	-2.65	-2.19	-1.34	-1.01	-1.86	-1.65	-2.50	-2.00	-1.54	-1.31	-0.72	-0.32
B3LYP	CH_3CH_2F	-4.16	-3.77	-3.00	-2.93	-4.26	-3.82	-3.81	-3.34	-3.78	-3.29	-1.68	-1.25
	CH_3CHF_2	-3.58	-3.20	-2.31	-2.12	-3.14	-2.82	-3.09	-2.63	-3.14	-2.64	-1.42	-1.03
	$CH_3CF_2CH_3$	-4.07	-3.52	-2.33	-2.06	-3.73	-3.23	-3.30	-2.65	-3.63	-3.14	-1.42	-0.96
	CH_3CF_3	-3.10	-2.65	-1.54	-1.31	-2.14	-1.86	-2.96	-2.49	-2.08	-1.50	-0.95	-0.52

[a] All the methods were applied with the same basis set (6-311G**++). [b] Hydrogen bond energy without BSSE correction. [c] Hydrogen bond energy with BSSE correction. Energies are in kcal/mol.

Table 3.1.4. Hydrogen bond acceptor properties of the fluorinated ethane derivatives. Hydrogen bond geometric characteristics

	Donor	Water			AmideD			Imidazole			Hydroxyl			Indole			CH_3SH		
[a]Method	Acceptor	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ
HF	CH_3CH_2F	2.084	154.831	2.5	2.222	172.571	3.0	2.103	178.140	2.3	2.119	155.562	1.9	2.153	176.423	2.4	2.482	175.159	0.0
	CH_3CHF_2	2.211	123.177	1.5	2.370	170.829	1.6	2.234	174.468	1.5	2.266	149.053	1.1	2.271	171.641	1.2	2.646	159.513	0.4
	$CH_3CF_2CH_3$	2.572	139.011	1.8	2.291	174.321	1.4	2.172	172.270	2.0	2.220	155.213	1.2	2.250	172.265	1.6	2.584	155.533	-0.5
	CH_3CF_3	2.509	125.460	1.1	2.451	169.959	0.5	2.311	170.529	0.9	2.472	126.931	0.9	2.236	169.709	0.6	2.762	162.810	-0.5
MP2	CH_3CH_2F	2.010	148.567	3.7	2.026	175.387	3.2	2.026	177.697	2.9	2.026	175.387	3.2	2.026	177.697	2.9	2.369	148.934	-2.3
	CH_3CHF_2	2.247	122.209	2.1	2.152	163.025	1.5	2.125	176.170	1.9	2.152	163.025	1.5	2.125	176.170	1.9	2.477	149.535	-0.1
	$CH_3CF_2CH_3$	2.129	137.015	2.8	2.047	171.537	1.9	2.089	170.470	2.0	2.047	171.537	1.9	2.089	170.470	2.0	2.427	134.670	1.6
	CH_3CF_3	2.352	124.691	1.6	2.233	166.900	1.8	2.200	169.361	0.6	2.233	166.900	1.8	2.200	169.361	0.6	2.528	160.231	0.4
BLYP	CH_3CH_2F	1.989	160.366	5.0	2.144	165.543	2.5	2.049	178.083	6.4	2.052	157.902	2.9	2.249	177.162	3.1	2.363	171.397	0.4
	CH_3CHF_2	2.204	130.982	2.3	2.365	155.522	1.6	2.205	172.500	5.0	2.168	151.905	1.5	2.254	173.347	0.6	2.506	165.807	-0.2
	$CH_3CF_2CH_3$	2.090	144.344	3.3	2.255	176.468	2.9	2.127	169.894	2.3	2.185	152.022	2.0	2.187	172.005	3.3	2.487	155.774	-0.1
	CH_3CF_3	2.314	132.532	1.6	2.457	165.471	1.2	2.261	169.937	0.4	2.592	122.560	0.3	2.234	173.871	1.6	2.831	154.534	-0.2
B3LYP	CH_3CH_2F	1.961	158.266	5.1	2.079	174.026	3.2	2.030	174.546	4.4	2.013	155.452	2.8	2.013	155.452	2.8	2.303	170.430	1.2
	CH_3CHF_2	2.174	129.169	2.7	2.312	155.892	1.8	2.184	171.791	2.8	2.139	150.998	1.6	2.139	150.998	1.6	2.475	158.554	0.2
	$CH_3CF_2CH_3$	2.080	142.302	3.0	2.189	176.309	1.2	2.100	171.442	3.3	2.128	152.521	2.1	2.128	152.521	2.1	2.420	154.250	-0.1
	CH_3CF_3	2.277	131.102	1.6	2.329	162.255	2.4	2.247	169.811	2.4	2.460	123.314	0.7	2.460	123.314	0.7	2.671	152.755	-0.5

[a]All the methods were applied with the same basis set (6-311G**++). [b]Hydrogen bond length. [c]Hydrogen bond angle. [d]D-H bond shift.

Table 3.1.5. Hydrogen bond acceptor properties of the fluorinated ethane derivatives. Hydrogen bond charge transfer characteristics: hydrogen and acceptor charge transfer

	Donor	Water		AmideD		Imidazole		Hydroxyl		Indole		CH_3SH	
[a]Method	Acceptor	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)
HF	CH_3CH_2F	0.023	0.007	0.115	0.006	0.199	-0.015	0.089	0.020	0.166	-0.006	-0.014	0.028
	CH_3CHF_2	0.016	-0.008	0.109	0.022	0.291	0.010	0.026	0.008	0.115	0.020	0.014	0.029
	$CH_3CF_2CH_3$	0.010	-0.012	0.163	-0.004	0.178	0.031	0.076	0.030	0.143	0.044	0.010	0.014
	CH_3CF_3	0.008	-0.009	0.111	0.014	0.235	0.006	0.014	0.004	0.107	0.013	0.001	0.020
MP2	CH_3CH_2F	0.009	0.005	0.088	0.016	0.185	-0.014	0.056	0.022	0.169	-0.004	-0.018	0.019
	CH_3CHF_2	0.012	-0.014	0.078	0.024	0.116	0.006	0.018	0.008	0.094	0.025	0.020	0.027
	$CH_3CF_2CH_3$	-0.003	-0.016	0.136	0.000	0.170	0.021	0.054	0.031	0.150	0.042	0.011	0.010
	CH_3CF_3	0.016	-0.011	0.120	0.012	0.163	-0.001	0.007	0.000	0.118	0.013	0.006	0.017
BLYP	CH_3CH_2F	-0.006	0.007	0.103	-0.001	0.190	-0.012	-0.030	0.020	0.044	0.016	0.141	0.000
	CH_3CHF_2	0.003	-0.010	0.051	0.022	0.102	0.007	0.002	0.025	0.008	0.006	0.116	0.020
	$CH_3CF_2CH_3$	-0.011	-0.011	0.116	0.006	0.171	0.029	0.002	0.014	0.036	0.027	0.155	0.043
	CH_3CF_3	0.001	-0.015	0.063	0.001	0.156	0.002	-0.001	0.019	0.005	0.008	0.125	0.015
B3LYP	CH_3CH_2F	-0.002	0.008	0.113	0.001	0.188	-0.012	0.052	0.017	0.178	-0.009	-0.025	0.022
	CH_3CHF_2	0.006	-0.012	0.056	0.023	0.103	0.011	0.015	0.010	0.114	0.018	0.008	0.027
	$CH_3CF_2CH_3$	-0.007	-0.012	0.133	-0.001	0.177	0.029	0.042	0.029	0.153	0.044	0.002	0.007
	CH_3CF_3	0.003	-0.015	0.087	0.011	0.157	0.002	0.007	0.011	0.112	0.013	-0.001	0.019

[a]All the methods were applied with the same basis set (6-311G**++). [b]Hydrogen charge transfer. [c]Acceptor charge transfer.

Table 3.1.6. Hydrogen bond properties of typical for proteins hydrogen bond acceptors. Hydrogen bond energies

	Donor	Water		AmideD		Imidazole		Hydroxyl		Indole		CH_3SH	
[a]Method	Acceptor	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}
HF	AmideA	-6.26	-6.06	-5.31	-5.36*	-7.31	-7.16	-5.68	-5.47	-6.20	-6.04	-2.48	-2.25
	Imidazole	-6.02	-5.73	-5.14	-5.17*	-6.89	-6.73	-5.49	-5.24	-5.80	-5.60	-2.48	-2.18
	Ketone	-5.30	-5.05	-4.37	-4.33*	-5.92	-5.68	-4.93	-4.65	-5.15	-4.89	-2.09	-1.78
	Furan	-2.77	-2.42	-1.89	-1.70	-3.31	-2.98	-2.55	-2.21	-2.20	-1.91	-0.96	-0.61
MP2	AmideA	-7.54	-6.25	-6.31	-5.46	-8.70	-7.73	-7.99	-6.94	-8.27	-7.33	-4.01	-2.68
	Imidazole	-7.80	-6.82	-8.42	-7.52	-10.05	-9.16	-8.27	-6.98	-9.58	-8.49	-5.19	-3.64
	Ketone	-6.55	-5.34	-5.72	-4.89	-6.89	-6.07	-6.89	-5.51	-6.92	-5.81	-3.33	-2.29
	Furan	-3.78	-2.88	-4.73	-3.36	-5.30	-4.34	-4.26	-3.14	-4.54	-3.34	-3.91	-1.89
B3LYP	AmideA	-6.91	-6.38	-5.05	-5.14	-7.05	-6.75	-6.20	-5.96	-6.46	-5.97	-2.62	-2.20
	Imidazole	-7.20	-6.88	-5.74	-5.89	-7.97	-7.59	-6.02	-5.95	-7.12	-6.48	-2.92	-2.56
	Ketone	-5.87	-5.37	-3.78	-3.96	-5.43	-5.14	-5.10	-4.85	-5.06	-4.73	-2.08	-1.68
	Furan	-2.86	-2.42	-1.47	-1.71	-3.05	-2.94	-2.30	-2.04	-2.71	-1.97	-0.51	-0.27
B3PLYP	AmideA	-7.41	-7.00	-5.61	-5.73	-7.79	-7.57	-6.81	-6.76	-6.94	-6.68	-2.94	-2.60
	Imidazole	-7.50	-7.25	-6.31	-6.44	-8.43	-8.07	-6.54	-6.48	-7.50	-6.90	-3.25	-2.89
	Ketone	-6.56	-5.92	-4.61	-4.55	-5.70	-5.24	-5.91	-5.45	-5.66	-5.06	-2.62	-2.05
	Furan	-3.30	-3.00	-1.97	-2.16	-3.46	-3.33	-2.75	-2.48	-2.99	-2.28	-0.88	-0.62

[a]All the methods were applied with the same basis set (6-311G**++). [b] Hydrogen bond energy without BSSE correction. [c] Hydrogen bond energy with BSSE correction. Energies are in kcal/mol.

Table 3.1.7. Hydrogen bond properties of typical for proteins hydrogen bond acceptors.Hydrogen bond geometric characteristics

	Donor	Water			AmideD			Imidazole			Hydroxyl			Indole			CH_3SH		
[a]Method	Acceptor	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ	[b]d, Å	[c]Angle, °	[d]Δr, mÅ
HF	AmideA	1.991	166.690	7.1	2.120	179.218	4.6	2.005	148.822	7.2	2.016	176.060	4.8	2.040	178.079	6.4	2.383	171.632	1.1
	Imidazole	2.094	171.049	7.9	2.117	178.906	8.5	2.131	178.955	9.8	2.116	170.668	6.8	2.181	178.774	8.6	2.497	163.719	1.7
	Ketone	2.038	166.113	5.5	2.177	178.557	3.9	2.072	178.678	4.6	2.063	166.899	4.5	2.100	174.664	4.3	2.451	170.328	0.4
	Furan	2.176	171.110	2.2	2.125	179.030	2.1	2.188	173.057	3.0	2.241	149.812	1.6	2.291	178.279	2.7	2.787	155.855	-0.3
MP2	AmideA	1.885	165.762	10.6	1.993	171.629	2.1	1.918	177.472	4.8	1.993	171.629	2.1	1.918	177.472	4.8	2.162	167.383	4.0
	Imidazole	1.978	170.993	11.7	2.027	178.668	10.0	1.940	178.841	17.8	2.027	178.668	10.0	1.940	178.841	17.8	2.236	160.720	7.6
	Ketone	1.936	164.360	8.7	2.022	170.927	3.3	1.990	179.276	4.8	2.022	170.927	3.3	1.990	179.276	4.8	2.245	162.617	2.3
	Furan	2.081	167.152	3.3	2.100	174.819	2.4	2.056	174.273	3.2	2.100	174.819	2.4	2.056	174.273	3.2	2.320	161.790	-1.8
B3LYP	AmideA	1.883	168.714	12.8	2.036	172.967	6.2	1.954	176.193	9.8	1.898	175.393	10.5	1.967	175.176	13.0	2.204	176.352	4.2
	Imidazole	1.953	174.639	16.7	2.090	179.119	13.7	2.000	175.822	18.4	1.996	167.467	12.3	2.033	179.547	16.7	2.292	166.024	7.1
	Ketone	1.924	166.400	9.4	2.117	171.709	5.1	2.043	171.394	5.0	1.966	166.053	8.9	2.048	176.941	5.3	2.334	168.417	3.3
	Furan	2.086	177.038	5.1	2.187	157.041	4.4	2.110	172.145	3.3	2.184	149.892	1.4	2.189	176.266	4.2	2.655	159.184	-0.9
B3PLYP	AmideA	1.860	168.728	12.7	1.994	173.872	6.1	1.937	176.201	11.5	1.860	174.925	10.7	1.860	174.925	10.7	2.202	177.625	5.0
	Imidazole	1.941	175.673	17.5	2.087	179.440	12.4	1.990	175.808	17.4	1.966	168.974	12.7	1.966	168.974	12.7	2.332	165.157	6.8
	Ketone	1.910	166.766	10.2	2.074	171.644	5.3	2.041	169.845	8.4	1.951	165.255	8.6	1.951	165.255	8.6	2.283	169.111	2.9
	Furan	2.048	176.820	4.5	2.167	176.929	3.8	2.121	172.220	5.9	2.177	149.612	2.5	2.177	149.612	2.5	2.612	110.817	1.2

[a]All the methods were applied with the same basis set (6-311G**++). [b]Hydrogen bond length. [c]Hydrogen bond angle. [d]D-H bond shift.

Table 3.1.8. Hydrogen bond properties of typical for proteins hydrogen bond acceptors. Hydrogen bond charge transfer characteristics: hydrogen and acceptor charge transfer

	Donor	Water		AmideD		Imidazole		Hydroxyl		Indole		CH_3SH	
[a]Method	Acceptor	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)	[b]Δq(H)	[c]Δq(A)
HF	AmideA	0.060	-0.040	0.207	-0.002	0.538	-0.091	0.198	-0.039	0.208	-0.073	0.020	-0.016
	Imidazole	0.145	-0.120	0.276	-0.082	0.607	-0.087	0.247	-0.100	0.198	-0.085	0.024	-0.028
	Ketone	0.059	-0.005	0.217	0.031	0.607	0.041	0.178	0.007	0.206	0.038	0.015	0.051
	Furan	0.095	-0.046	0.230	0.034	0.641	0.008	0.095	-0.001	0.197	0.016	0.002	0.055
MP2	AmideA	0.030	-0.030	0.183	-0.009	0.215	-0.084	0.174	0.014	0.211	-0.061	0.011	-0.001
	Imidazole	0.147	-0.121	0.263	-0.072	0.299	-0.062	0.252	-0.092	0.239	-0.053	0.029	-0.038
	Ketone	0.029	0.001	0.186	0.004	0.268	0.040	0.163	0.025	0.206	0.031	0.009	0.045
	Furan	0.094	-0.063	0.167	0.019	0.302	0.000	0.070	0.004	0.231	0.021	0.003	0.058
B3LYP	AmideA	0.016	-0.020	0.191	0.007	0.225	-0.068	-0.041	0.000	0.170	0.005	0.242	-0.055
	Imidazole	0.137	-0.112	0.226	-0.035	0.316	-0.034	-0.026	0.007	0.231	-0.060	0.255	-0.014
	Ketone	0.018	0.009	0.190	0.024	0.254	0.038	-0.026	0.045	0.098	0.016	0.262	0.046
	Furan	0.069	-0.028	0.140	0.021	0.293	0.012	-0.017	0.040	0.062	0.016	0.237	0.036
B3PLYP	AmideA	0.028	-0.026	0.207	0.007	0.236	-0.078	0.188	0.000	0.233	-0.059	-0.026	-0.007
	Imidazole	0.154	-0.124	0.236	-0.044	0.318	-0.045	0.262	-0.074	0.250	-0.023	-0.012	-0.001
	Ketone	0.028	0.006	0.205	0.026	0.251	0.041	0.109	0.017	0.234	0.041	-0.019	0.047
	Furan	0.085	-0.036	0.160	0.023	0.295	0.010	0.072	0.012	0.238	0.034	-0.015	0.043

[a]All the methods were applied with the same basis set (6-311G**++). [b]Hydrogen charge transfer. [c]Acceptor charge transfer.

Table 3.1.9.Characteristics of bifurcated hydrogen bonds of the fluorinated ethane derivatives with water

[a]Method	Acceptor	[b]ΔE, kcal/mol	[c]ΔE_{BSSE}, kcal/mol	[d]d, Å	[e]angle, °	[f]Δr, mÅ	[g]Δq(H)	[h]Δq(A)
HF	CH_3CHF_2	-2.72	-2.22	2.317	126.921	0.8	0.006	-0.004
	$CH_3F_2CH_3$	-3.01	-2.45	2.572	139.011	1.8	0.014	0.001
MP2	CH_3CHF_2	-3.94	-2.60	2.211	131.309	0.8	0.005	-0.010
	$CH_3F_2CH_3$	-4.32	-2.87	2.466	127.036	1.2	0.013	-0.005
BLYP	CH_3CHF_2	-2.58	-1.93	2.129	137.015	2.8	0.005	-0.009
	$CH_3F_2CH_3$	-3.06	-2.12	2.409	130.201	1.4	0.012	-0.009
B3LYP	CH_3CHF_2	-3.04	-2.40	2.524	128.645	0.9	0.004	-0.010
	$CH_3F_2CH_3$	-3.43	-2.58	2.090	144.344	3.3	0.014	-0.008

[a] All the methods were applied with the same basis set (6-311G**++). [b] Hydrogen bond energy without BSSE correction. [c] Hydrogen bond energy with BSSE correction. [d] Hydrogen bond length. [e] Hydrogen bond angle. [f] D-H bond shift. [g] Hydrogen charge transfer. [h] Acceptor charge transfer.

Most donor-acceptor pairs with fluoromethylated groups acting as hydrogen bond donor are

interacting via bifurcated hydrogen bonds (Tables 3.1.10-3.1.11).

Table 3.1.10. Hydrogen bond donor properties of the fluorinated ethane derivatives. Hydrogen bond energies.

	Donor	CH_3CH_3		CH_3CH_2F		CH_3CHF_2	
[a]Method	Acceptor	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}	[b]ΔE	[c]ΔE_{BSSE}
HF	AmideA	-0.37	-0.28	[d]-2.18	[d]-2.04	[e]-4.91	[e]-4.70
	Imidazole	-0.31	-0.23	-1.92	-1.69	[e]-3.62	[e]-3.32
	Ketone	-0.30	-0.20	[d]-1.81	[d]-1.65	-2.82	-2.56
	Furan	-0.18	-0.06	-1.60	-1.28	[e]-1.85	[e]-1.43
MP2	AmideA	-1.11	-0.66	[d]-3.08	[d]-2.42	[e]-6.40	[e]-5.18
	Imidazole	-1.05	-0.78	-3.85	-3.00	[e]-5.45	[e]-4.45
	Ketone	-1.06	-0.60	-2.77	-2.02	-3.50	-2.74
	Furan	-0.88	-0.79	-3.05	-2.11	[e]-3.39	[e]-2.30
BLYP	AmideA	-0.44	-0.32	[d]-1.98	[d]-1.67	[e]-4.73	[e]-4.65
	Imidazole	-0.11	-0.09	-1.75	-1.56	[e]-2.85	[e]-2.66
	Ketone	-0.20	-0.04	[d]-1.38	[d]-1.21	-2.30	-2.20
	Furan	No binding	-	-1.14	-1.00	[e]-1.10	[e]-1.04
B3LYP	AmideA	-0.24	-0.09	[d]-2.28	[d]-2.04	[e]-5.30	[e]-5.21
	Imidazole	-0.36	-0.20	-2.16	-1.90	[e]-3.46	[e]-3.26
	Ketone	-0.60	-0.30	[d]-1.91	[d]-1.45	-2.89	-2.52
	Furan	No binding	-	-1.52	-1.41	[e]-1.55	[e]-1.44

[a] All the methods were applied with the same basis set (6-311G**++). [b] Hydrogen bond energy without BSSE correction. [c] Hydrogen bond energy with BSSE correction. [d] Bifurcated hydrogen bond with one acceptor and 2 donors. [e] Bifurcated H-bond with an additional fluorine atom also H-bonded). Energies are in kcal/mol.

Table 3.1.11. Hydrogen bond donor properties of the fluorinated ethane derivatives. Hydrogen bond geometric characteristics and charge transfer.

[a]Method	[b]D/A	AmideA					Imidazole					Ketone					Furan				
HF		[c]d, Å	[d]Angle, °	[e]Δr, mÅ	[f]Δq(H)	[g]Δq(A)	[c]d, Å	[d]Angle, °	[e]Δr, mÅ	[f]Δq(H)	[g]Δq(A)	[c]d, Å	[d]Angle, °	[e]Δr, mÅ	[f]Δq(H)	[g]Δq(A)	[c]d, Å	[d]Angle, °	[e]Δr, mÅ	[f]Δq(H)	[g]Δq(A)
	CH_3CH_3	2.887	173.310	-1.5	0.090	0.022	3.174	171.141	-1.2	0.029	0.004	2.997	172.999	-1.2	0.071	0.042	3.180	168.351	-0.5	0.044	0.028
	CH_3CH_2F	3.038	96.545	-1.9	0.046	-0.004	2.816	139.011	-2.6	0.042	-0.024	3.062	97.193	-1.5	0.045	0.040	3.037	119.351	-1.3	0.018	0.029
	F-H bond	-	-	-	-	-	2.675	139.353	2.7	0.008	0.024	-	-	-	-	-	2.530	139.101	-0.1	0.043	0.028
	CH_3CHF_2	2.630	116.901	-1.9	0.065	0.003	2.772	119.265	-3.7	0.045	-0.009	2.507	143.364	-3.8	0.108	0.041	2.874	114.335	-2.9	0.033	0.027
	F-H bond	2.252	154.477	1.8	0.054	-0.006	2.634	123.213	-0.3	0.036	0.008	-	-	-	-	-	2.669	127.452	-0.2	0.028	0.018
MP2	CH_3CH_3	2.585	167.135	-0.3	0.099	0.033	3.066	179.349	-0.8	0.046	-0.009	2.620	153.658	-1.8	0.062	0.048	3.044	167.409	-0.3	0.040	0.034
	CH_3CH_2F	2.849	93.981	-1.5	0.043	0.012	2.619	138.555	-1.7	0.032	0.015	2.821	94.856	-1.2	0.044	0.044	2.880	120.789	-0.1	0.018	0.033
	F-H bond	-	-	-	-	-	2.506	139.872	0.3	0.037	0.024	-	-	-	-	-	2.424	139.389	-0.2	0.040	0.027
	CH_3CHF_2	2.510	115.450	-3.6	0.064	0.006	2.640	119.742	-3.3	0.040	0.005	2.465	130.671	-3.2	0.082	0.041	2.746	113.316	-2.1	0.031	0.030
	F-H bond	2.115	156.372	1.7	0.054	-0.007	2.535	123.007	-0.1	0.036	0.007	-	-	-	-	-	2.539	127.792	-1.2	0.029	0.017
BLYP	CH_3CH_3	2.897	167.213	-1.4	0.077	0.020	3.129	169.669	-0.2	0.033	0.004	3.054	162.787	-1.5	0.049	0.036	3.008	169.046	-1.6	0.029	0.038
	CH_3CH_2F	3.036	94.822	-4.5	0.024	0.006	2.757	138.940	-1.9	0.023	0.023	3.277	94.981	-1.2	0.019	0.042	3.065	120.795	-1.4	0.007	0.030
	F-H bond	-	-	-	-	-	2.612	141.211	0.1	0.027	0.018	-	-	-	-	-	2.447	145.269	-0.6	0.039	0.024
	CH_3CHF_2	2.555	126.094	-3.3	0.054	0.002	2.722	119.338	-3.0	0.020	0.022	2.422	150.411	-1.9	0.093	0.039	2.827	116.539	-1.0	0.020	0.035
	F-H bond	2.107	160.376	3.0	0.059	-0.005	2.604	124.079	0.0	0.034	0.006	-	-	-	-	-	2.754	126.905	-0.8	0.021	0.017
B3LYP	CH_3CH_3	2.735	168.348	-0.9	0.090	0.022	3.131	175.780	-2.0	0.039	-0.004	2.927	161.174	-0.3	0.055	0.040	3.226	167.326	-0.8	0.026	0.031
	CH_3CH_2F	3.072	94.825	-1.9	0.036	0.001	2.735	138.851	-6.8	0.021	0.021	3.015	95.726	-1.3	1.035	0.045	3.042	120.747	-1.0	0.014	0.029
	F-H bond	-	-	-	-	-	2.581	140.999	0.5	0.028	0.021	-	-	-	-	-	2.417	144.964	-0.8	0.040	0.027
	CH_3CHF_2	2.507	124.502	-2.4	0.068	0.001	2.735	119.339	-2.2	0.068	0.014	2.368	148.671	-3.4	0.101	0.044	2.804	115.664	-4.3	0.033	0.035
	F-H bond	2.080	159.135	3.3	0.055	-0.006	2.610	123.435	-2.9	0.033	0.005	-	-	-	-	-	2.698	126.690	-0.8	0.022	0.017

[a] All the methods were applied with the same basis set (6-311G**++). [b] Donors and acceptors. [c] Hydrogen bond length. [d] Hydrogen bond angle. [e] D-H bond shift. [f] Hydrogen charge transfer. [g] Acceptor charge transfer.

Furthermore, some fluoromethylated groups exhibit a dual donor/acceptor behaviour where the fluorine atom can in addition act as a hydrogen bond acceptor. Water does not behave as acceptor of hydrogen bond with fluoromethylated groups. The optimization shows that water acts as hydrogen bond donor and fluorine atoms are acceptors. Most of these hydrogen bonds are relatively weak in terms of their energies and lengths (Tables 3.1.10-3.1.11). However, they are much stronger than weak hydrogen bonds formed by methyl groups as donors, which are nevertheless considered to be important in protein environments and molecular recognition [214,215]. For example, the energy value of the hydrogen bond formed by the hydrogen atom of $CH_3C\mathit{H}F_2$ and a ketone oxygen is one order lower than the one of the hydrogen bond formed by the ethane hydrogen atoms and a ketone oxygen. Moreover, some hydrogen bonds established by hydrogen atoms from the fluorinated groups are comparable in terms of energies to the hydrogen bonds formed by other analyzed groups presented in canonical amino acids (Tables 3.1.3-3.1.8, 3.1.10-3.1.11). The hydrogen bonds formed by fluoromethyl group as a donor yield a negative D-H shift (blue-shifted hydrogen bonds). Charge transfer on the hydrogens is in the order of 10^{-2} units, which is one order less than in donor-acceptor pairs, where fluoromethyl groups act as acceptors. In general, hydrogen bond donor properties of fluoromethyl groups are weak and the groups prefer to participate in bifurcated hydrogen bonds.

Comparison of the results obtained by QM shows correlation (R^2 more than 0.7, see Table 3.1.12) between different levels of theory for all analyzed hydrogen bonding parameters. The best correlation is found for the energy calculations, while the worst is observed for HF method compared with other methods for hydrogen charge transfer and bond length calculations. In general, we conclude that all used QM methods yield qualitatively similar characterization of the studied hydrogen bonding properties.

Table 3.1.12. Correlation between different levels of theory (adjusted R^2) for hydrogen bond calculations

	Energy without BSSE correction			Energy with BSSE correction			D-H bond shift			Relative D-H bond shift		
Method	MP2	BLYP	B3LYP	MP2	BLYP	B3LYP	MP2	BLYP	B3LYP	MP2	BLYP	B3LYP
HF	0.872	0.964	0.971	0.920	0.974	0.973	0.778	0.893	0.907	0.829	0.920	0.932
MP2	-	0.900	0.900	-	0.951	0.946	-	0.879	0.850	-	0.885	0.861
BLYP	-	-	0.995	-	-	0.996	-	-	0.879	-	-	0.928
	H-bond length			H-bond angle			Hydrogen charge transfer			Acceptor charge transfer		
Method	MP2	BLYP	B3LYP	MP2	BLYP	B3LYP	MP2	BLYP	B3LYP	MP2	BLYP	B3LYP
HF	0.778	0.854	0.847	0.819	0.798	0.752	0.741	0.704	0.708	0.911	0.841	0.888
MP2	-	0.836	0.880	-	0.789	0.696	-	0.956	0.973	-	0.882	0.916
BLYP	-	-	0.954	-	-	0.767	-	-	0.989	-	-	0.986

Volumes and ASA of fluoromethylated groups and fluorinated amino acid side chains. We estimated volumes and ASA for the fluoromethylated groups from ethane derivatives (Figure 3.1.4 A). Fluorination of the methyl group leads to a volume and ASA increase of 21%, 33%, 47% and 15%, 27%, 39%, correspondingly, for each consecutive fluorine substitution.

We created libraries of 6 non-canonical L-amino acids compatible with AMBER 8.0: ethylglycine (Abu), 4-monofluoroethylglycine (MfeGly), 4,4-difluoroethylglycine (DfeGly), 4,4,4-trifluoroethylglycine (TfeGly), 4,4-difluoropropylglycine (DfpGly) and propylglycine. In comparison to canonical amino acids, side chain size increases as follows: Ala<Abu<MfeGly<DfeGly<TfeGly<Val<DfpGly<Ile<Leu<Met<Phe (Fig. 3.1.4 B). However, there are no canonical amino acids that could be isosterically substituted by the fluorinated ethylglycine derivatives since they all are branched except for Met, which chain is one carbon atom longer than DfpGly. Moreover, because of substantially different electrostatic properties dictated by fluorination, incorporation of fluorinated amino acids into protein environments could not be correctly estimated by taking into account only the size of the group.

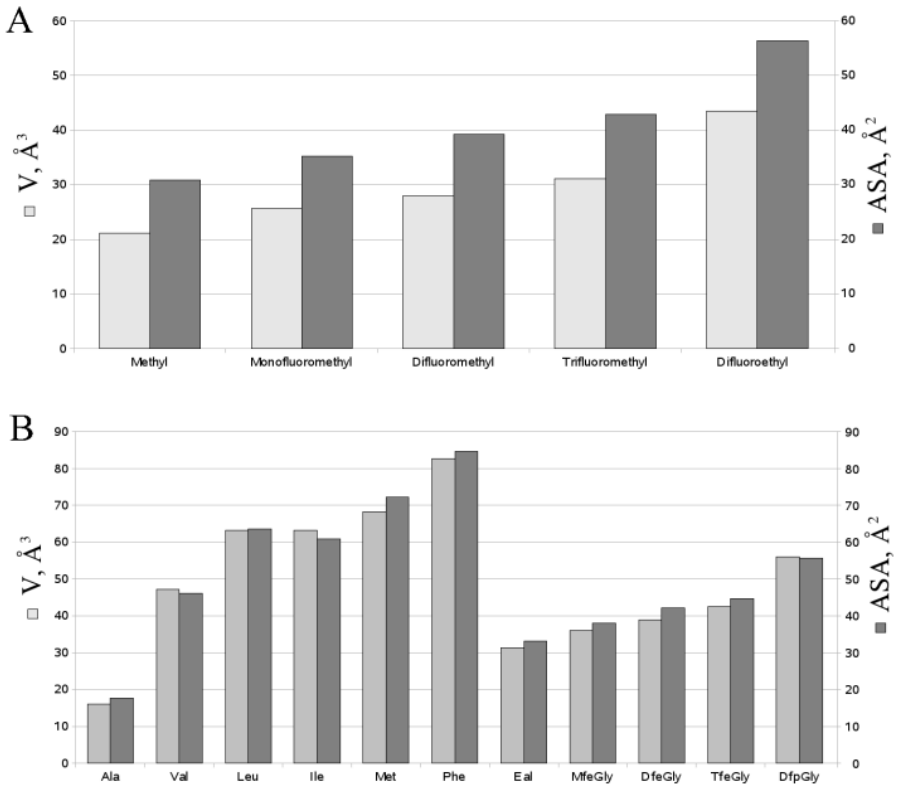

Figure 3.1.4. Volumes and solvent accessible surface areas (ASA). A). Fluoromethyl groups after B3LYP (6-311G**++) geometry optimization. B) Amino acids side chains created for AMBER libraries.

Conformational analysis of fluorinated amino acids. Ramachandran plots were calculated for the fluorinated amino acids and compared to other canonical hydrophobic amino acids (see Methods section). Their α-helical and β-strand propensities for amino acids can be observed in figure 3.1.5. As already described, Ala, Leu and Met have higher α-helical, while Val and Phe have higher β-strand propensities [216]. MfeGly and TfeGly show slight but clear preference for α-helical conformation in the Ace-FXR-Nme system, while DfeGly and DfpGly rather adopt a β-strand conformation. MfeGly, and TfeGly have substantially higher left α-helical propensities than Abu, DfeGly and DfpGly.

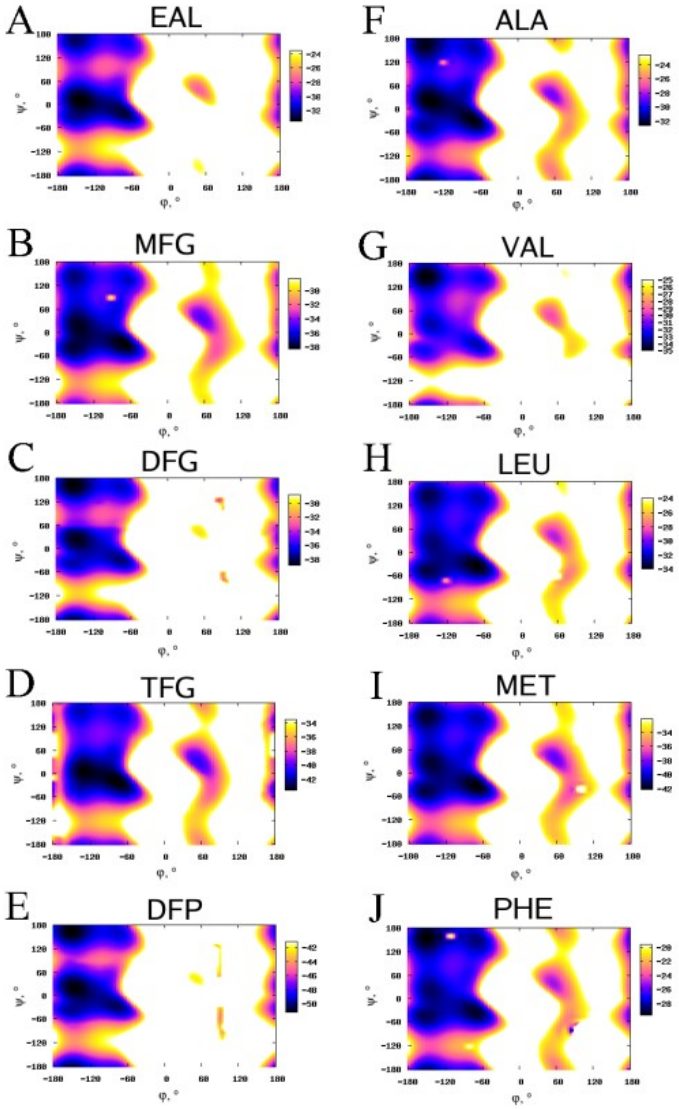

Figure 3.1.5. Ramachandran plots of the fluorinated amino acids in comparison to canonical hydrophobic amino acids.

For quantitative characterization of secondary structure propensities we analyzed *propensity indexes* (see Methods), which are shown in figure 3.1.6. The propensity indexes do not directly

characterize secondary structure propensities but rather the relations between the obtained potential energies for amino acids in the used the Ace-FXR-Nme dipeptide and summarized data for backbone dihedral angles in secondary structure elements from PDB structures. Propensity indexes are proportional to quadratic probabilities to adopt a certain conformation and, thus, demonstrate differences between amino acids more profoundly than probabilities. The obtained results reveal low β-strand propensity indexes for Abu, MfeGly and TfeGly. DfeGly and DfpGly have essentially higher β-strand propensity indexes, which, nevertheless, are still lower than for typical 'β-amino acids' like Val and Phe. As for α-helical propensity indexes, MfeGly and TfeGly have the highest values among the fluorinated amino acids and are comparable with typical 'α-amino acids' like Met, Ala. Abu, DfeGly and DfpGly have low α-helical propensity indexes. A big difference was found for the left α-helical propensity index, which shows significant increase for MfeGly and TfeGly. The analysis shows that secondary structure propensities are similar for two pairs of fluorinated amino acids: MfeGly and TfeGly; DfeGly and DfpGly. We suppose, that this finding could be partly attributed to their different electric dipole properties.

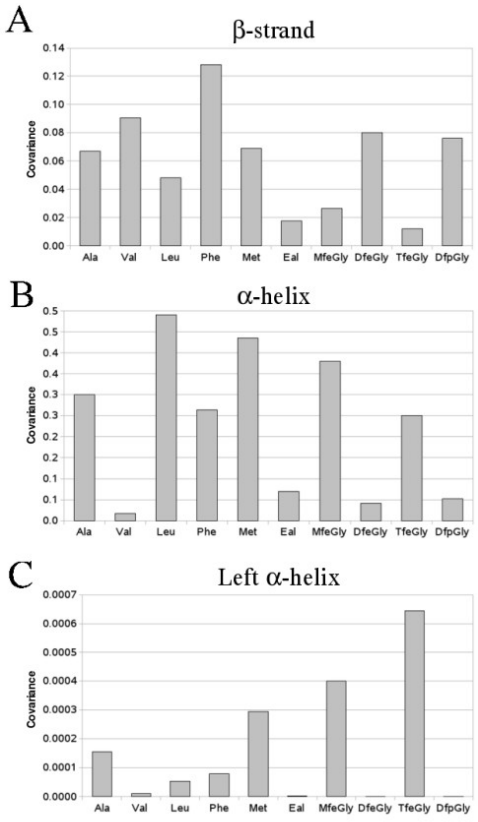

Figure 3.1.6. Covariance (propensity index) between probabilities obtained from calculated Ramachandran plots and PDB-derived secondary structure data. A) β-strand. B) α-helix. C) Left α-helix.

We characterized the side chain rotamers of Abu and its fluorinated derivatives using the Ace-FXR-Nme system and analyzed the changes occurring upon fluorination for three secondary structure types: α-helix, β-strand and left α-helix (Table 3.1.13). All these amino acids *i)* reveal three potential energy minima independently of the backbone dihedral angles values except for DfeGly in β-strand conformation, which has two minima; *ii)* have very similar values of χ_1 dihedral angle corresponding to energy minima (Figure 3.1.6). The differences in energy barriers between the minima could be explained in terms of electrostatic and dipole interactions of the fluoromethyl groups with backbone atoms and their steric demands.

Table 3.1.13. Side chain rotamers of Abu, MfeGly, DfeGly, TfeGly

	β-strand				α-helix				Left α-helix			
Amino acid	[a]χ_1, °	[b]E_{min}, kcal/mol	[c]p	[d]Δχ, °	[a]χ_1, °	[b]E_{min}, kcal/mol	[c]p	[d]Δχ, °	[a]χ_1, °	[b]E_{min}, kcal/mol	[c]p	[d]Δχ, °
Abu	-150	-29.7	0.09	51	-170	-31.8	0.20	44	-155	-27.9	0.17	27
	-65	-30.2	0.13	45	-60	-32.6	0.44	39	-55	-29.4	0.82	34
	65	-31.9	0.78	36	65	-32.4	0.36	37	60	-25.3	0.01	24
MfeGly	-175	-36.5	0.34	35	-175	-37.8	0.42	41	-160	-34.1	0.36	30
	-70	-35.8	0.17	42	-60	-37.5	0.33	39	-55	-34.7	0.63	23
	60	-36.8	0.49	39	65	-37.2	0.25	37	75	-30.2	0.01	21
DfeGly	-170	-38.3	0.46	45	-175	-37.4	0.36	41	-160	-31.7	0.91	31
	-75	-35.6	0.03	47	-65	-37.3	0.34	43	-65	-29.4	0.09	27
	65	-38.4	0.51	41	70	-37.2	0.30	37	-	-	-	-
TfeGly	-130	-39.8	0.17	116	-165	-41.9	0.22	51	-140	-37.8	0.24	30
	-70	-40.5	0.36	38	-65	-42.9	0.59	42	-60	-39.0	0.76	30
	70	-40.8	0.47	37	70	-41.8	0.19	30	70	-24.5	0.00	15

[a] Value for χ_1 dihedral angle corresponding to an energy minimum. [b] Energy minimum value. [c] Probability for the side chain to be in a conformation corresponding to χ_1. [d] Width of the area corresponding to $|E(\chi_1)-E_{min}|<1$ kcal/mol.

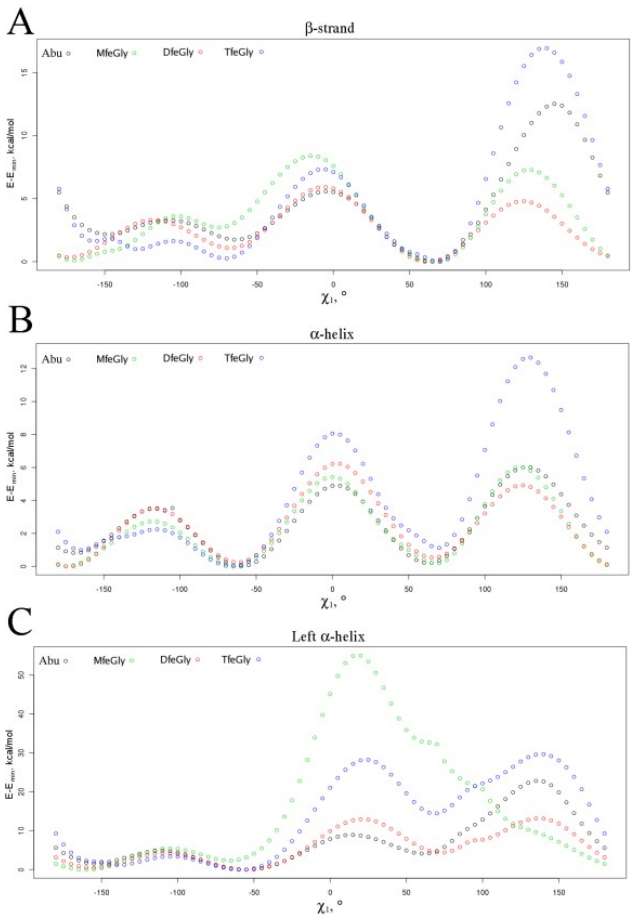

Figure 3.1.7. Side chain rotamers potential energy $E(\chi_1)$-$E_{min}(\chi_1)$ of the fluorinated ethylglycine derivatives in different backbone conformations. A) β-strand. B) α-helix. C) Left α-helix.

For example, for the energy barrier value between the minimum with $\chi_1 \in [60;70]$ and the corresponding maximum with $\chi_1 \in [120;150]$ in β-strand conformation there is an obvious increase of value for TfeGly compared to Abu and decrease for MfeGly and DfeGly. In minimum energy conformation none of the side chains interacts with the Ace-Xxx-Nme backbone, while the maximum energy conformation allows the side chain to interact with the backbone atoms (Figure 3.1.8). In particular, there is a repulsion between the trifluoromethyl group and the two electronegative atoms of the backbone (the carboxyl oxygen from TfeGly residue and the nitrogen from Nme group) only partly compensated by a very weak hydrogen bond between one of the fluorine atoms and the amide hydrogen of Nme group. These interactions lead to the increase of the barrier value. In case of Abu, there are no similar interactions since the methyl group is apolar. At the same time, one of the fluorine atoms of DfeGly forms a stronger hydrogen bond than in case of TfeGly with the amide hydrogen of Nme group, which lowers the energy barrier value. Similarly the barrier is lower in case of MfeGly

because of the hydrogen bond formation between the hydrogen of the monofluoromethyl group and the carboxyl oxygen from MfeGly, but in this case the hydrogen bond is weaker. In case of α-helix, the found energy maxima correspond to roughly the same χ_1 values but the energy barrier differences between Abu, MfeGly, DfeGly are lower since the distances between the fluoromethyl group and backbone atoms are longer. Another maximum, corresponding to $\chi_1 \sim 0$ appears to be due to both electrostatic repulsion and steric clash between the fluorine atoms and the oxygen and nitrogen atoms of the backbone. In the case of left α-helical conformation DfeGly has the most unfavorable interaction corresponding to this maximum with an energy barrier even higher than the one of TfeGly, because the fluorine atoms of DfeGly are more negatively charged than of TfeGly but the group volume is smaller. The third and also the lowest energy barrier ($\chi_1 \sim -120$) is very similar for all analyzed amino acids and, therefore, is almost independent of fluorination. In general, all the highest barriers are for left α-helical conformation and the lowest for α-helix, which is explained in terms of different proximities of backbone and side chain atoms. Obviously, dipole interactions should be also considered for this kind of analysis. As it was mentioned before, there is a significant difference in dipole vector directions of MfeGly, DfeGly and TfeGly.

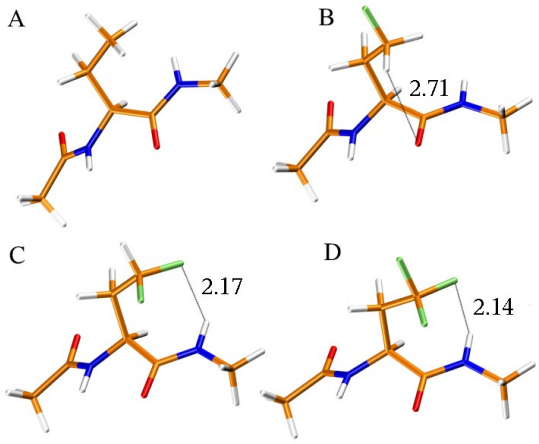

Figure 3.1.8. β-strand conformation for Ace-Xxx-Nme dipeptides, where Xxx= A) Abu. B) MfeGly. C) DfeGly. D) TfeGly.

For the DfpGly side chain we found 9 conformations for α-helix and 6 conformations for β-strand and 6 fox left α-helix, which corresponded to a probability ≥ 0.01 (Table 3.1.14). For the left α-helix $\chi_1>0$ values are conformationally forbidden because of steric clash between the side chain and the backbone as well as of electrostatic repulsion between the fluorine atoms and the backbone carbonyl oxygens (Figure 3.1.9). Some of the rotamers corresponding to the minima contain intramolecular hydrogen bonds between the fluorine atoms and the backbone amide hydrogens of the backbone.

However, the probabilities for a certain rotamer do not correlate with the probability to form a hydrogen bond, meaning that hydrogen bond interaction between the fluorine atoms of the difluorinated group and the backbone hydrogen atoms is not the decisive factor for a rotamer preference. We suppose that hydrophobic and dipole interactions make instead a substantial impact on the dynamic behavior of the side chain of DfpGly.

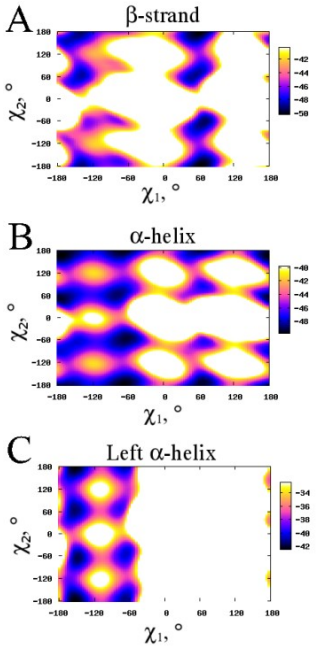

Figure 3.1.9. DfpGly side chain rotamers potential energy E(χ_1, χ_2) in different backbone conformations. A) β-strand. B) α-helix. C) Left α-helix.

Table 3.1.14. Side chain rotamers of DfpGly

Secondary structure	[a]χ_1, °	[b]χ_2, °	[c]E_{min}, kcal/mol	[d]p	[e]H-bond	Secondary structure	[a]χ_1, °	[b]χ_2, °	[c]E_{min}, kcal/mol	[d]p	[e]H-bond
β-strand	65	-175	-50.3	0.44	+	α-helix	-75	65	-48.8	0.08	-
	70	-60	-49.4	0.19	-		-175	-90	-47.9	0.03	-
	-140	-180	-49.2	0.15	+		70	85	-47.3	0.02	+
	55	65	-48.9	0.11	+		70	-65	-47.1	0.01	+
	-140	65	-48.7	0.09	+	Left α-helix	-150	-180	-42.5	0.43	-
	-125	-65	-46.3	0.01	-		-145	-60	-41.7	0.19	-
α-helix	-175	170	-49.9	0.23	-		-150	65	-41.5	0.16	-
	-65	-175	-49.7	0.20	+		-65	-170	-41.3	0.12	[g] +
	-175	55	-49.4	0.15	-		-65	-50	-40.6	0.07	+
	70	-170	-49.4	0.15	[f]+		-75	70	-39.6	0.02	-
	-60	-55	-49.3	0.13	+						

[a,b] Values for χ_1 and χ_2 dihedral angles corresponding to an energy minimum. [c] Energy minimum value. [d] Probability for the side chain to be in a conformation corresponding to (χ_1,χ_2). [e] Presence of a hydrogen bond with backbone. [f] Bifurcated hydrogen bond: two hydrogen atoms interact with one fluorine. [g] Bifurcated hydrogen bond: 2 fluorine atoms interact with one hydrogen.

Hydration of fluorinated amino acids. To characterize the hydrophobic properties of the fluorinated

amino acids we carried out free energy perturbation calculations *in vacuo* and in a TIP3 water box. MfeGly, DfeGly, TfeGly were perturbed to Abu, and DfpGly was perturbed to propylglycine, so that the fluorine atoms were perturbed into hydrogens. We obtained the differences between the energies of solvation for fluorinated amino acids and their non-fluorinated analogues (Table 3.1.15). The results show that these differences are negative for all fluorinated amino acids, meaning that hydration of the fluorinated amino acids is energetically more favorable than the hydration of their non-fluorinated analogues. Hydrophobicity of the mono-fluorinated side chain increases with the addition of fluorine atoms. Nevertheless, the difference in hydrophobicity between non-fluorinated side chains and di/trifluorinated is low.

Table 3.1.15. Hydration energies differences between fluorinated and non-fluorinated amino acids

Fluorinated residue	Non-fluorinated residue	[a]ΔE_{vacuo}	[b]ΔE_{water}	[c]$\Delta\Delta E_{hydration}$
MfeGly	Abu	3.8±0.29	6.1±0.58	–2.5±0.5
DfeGly	Abu	6.3±0.49	7.0±0.81	–0.9±0.7
TfeGly	Abu	9.1±0.87	9.6±0.95	–0.5±0.9
DfpGly	Propylglycine	20.2±0.55	21.1±0.81	–0.9±0.7

[a] Free energy of perturbation of a fluorinated residue to a non-fluorinated residue *in vacuo*. [b] Free energy of perturbation of a fluorinated residue to a non-fluorinated residue in TIP3 water. [c] Difference between hydration energies of a fluorinated and a corresponding non-fluorinated residue.

The obtained retention times for the canonical hydrophobic amino acids correlate very well, albeit not linearly between the volume of the side chains (calculated as described by Zhao *et al.* [217]) and their hydrophobicity (Figure 3.1.10). We fitted the plot for the non-fluorinated amino acids to yield an exponential equation to describe the correlation between the side chain volume and retention time ($rt = 8.0012\ e^{0.0086\ V(VdW)}$; R^2=0.998). The fluorinated amino acids, however, scatter around the curve and do not fit into the equation proposed above. The mono- and difluorination of Abu increases the volume of its side chain by roughly 6 Å^3 per fluorine atom. The hydrophobicity of the mono- and difluorinated side chains is lower than for Abu due to the polarization of the hydrogens in close proximity to the fluorine atoms and, thus, manifest a more favorable interaction with water. This effect is most dramatic for DfpGly. With a side chain very close to Leu in terms of steric size it represents an even more hydrophilic derivative than TfeGly. Nevertheless, a closer look at the fluorinated analogues of Abu (MfeGly, DfeGly, and TfeGly) also shows that the hydrophobicity increases more progressively by stepwise fluorination than it does by elongation or branching of the side chain without fluorination, which agrees with our calculations of free energies of hydration.

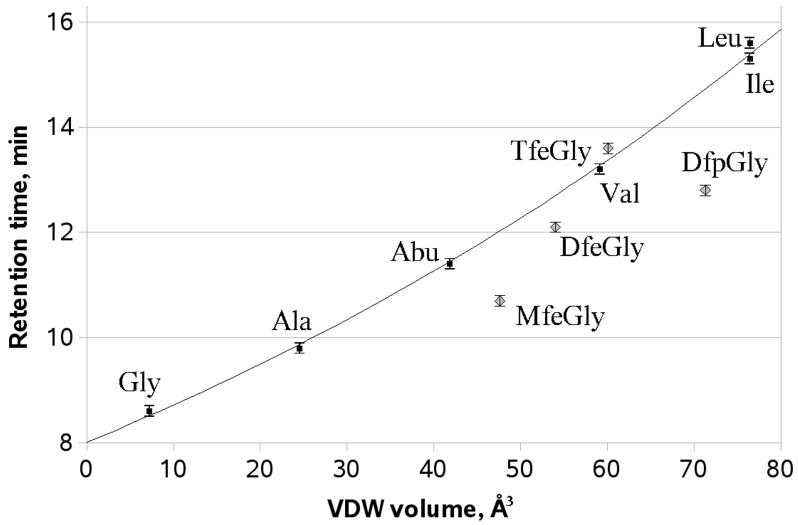

Figure 3.1.10. Retention times of the Fmoc-amino acids against the van der Waals volume of the side chains. Non-fluorinated amino acids are represented by black squares, the correlation between them is shown with a black line and fluorinated amino acids are represented by gray diamonds.

The hydration energy of 5,5,5,5',5',5'-hexafluoroleucine was found to be 1.1 kcal/mol higher than of leucine in another free energy perturbation study [211]**.** This, together with our experimental and computational findings, suggests that there are two opposing factors determining the solvation energetics for the fluorinated groups. On one hand, substitution of a hydrogen atom by fluorine increases the ASA of the chemical group and leads to the increase of solvation energy and hydrophobicity. On the other hand, the C-F bond is more polarized than the C-H bond and electrostatic interactions of the fluorinated group with solvent are energetically more favorable. With an increase of fluorination the polarity of the group decreases and that leads to weaker interactions with solvent. These results agree with the data obtained by Yin *et al.* for fluorinated ethane derivatives using the CHARMM force field [218].

Analysis of the radial distribution function (RDF) shows that the density of solvent in proximity to the side chains of fluorinated amino acids depends on the number of fluorine atoms. The solvent density in a 2 Å distance from the fluorine atom of MfeGly is about twice higher than for the fluorine atoms of DfeGly, TfeGly and DfpGly (Figure. 3.1.11 A), and it roughly corresponds to the hydrogen bond length between the fluorine of the fluoromethylated group and water hydrogens. At the same time, there is almost no difference between the RDF calculated for hydrogen atoms of fluoromethyl groups of Abu, MfeGly, DfeGly and water oxygen atoms (Figure 3.1.11 B), suggesting very weak hydrogen bonds. The same conclusion about these hydrogen bonds was drawn from our QM calculations. These results agree well qualitatively with the data obtained for fluorinated ethane

derivatives using the CHARMM force field [218] proving that *i)* fluorinated ethane derivatives, which we used for our QM calculations, represent a good model to study fluoromethylation in amino acid side chains; *ii)* parameterization for the AMBER force field of amino acids including fluoromethylated groups in their side chains and ethane derivatives for the CHARMM force field allows to describe similar hydrogen bonds properties.

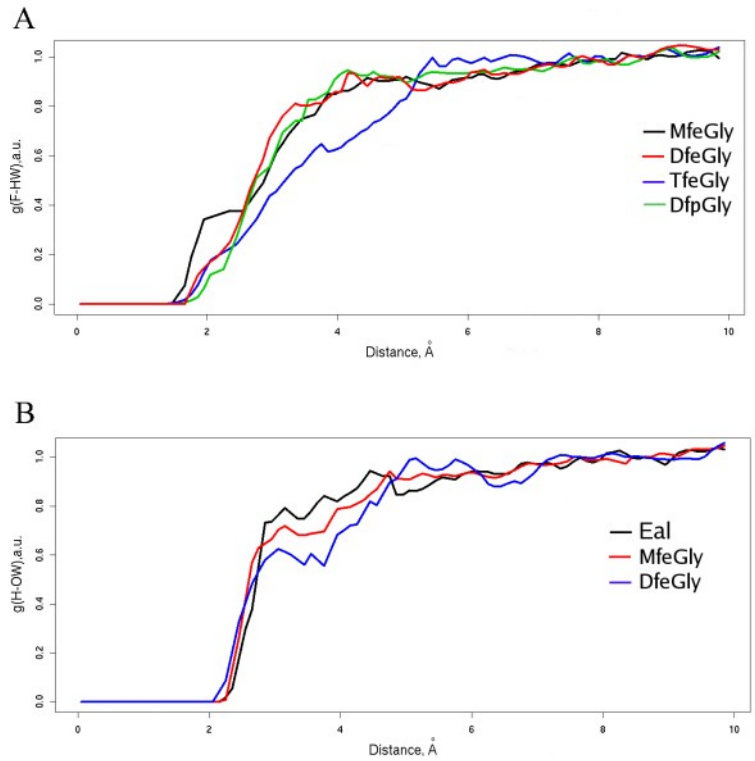

Figure 3.1.11. Radial distribution function (RDF) for A) fluorine atoms of the fluoromethylated group and water hydrogen atoms; B) hydrogen atoms of the fluoromethylated group and water oxygen atoms.

3.1.5 Conclusions

In this study we apply a QM approach to study the hydrogen bonding properties of fluorinated ethane derivatives as a simplification for fluorinated amino acid side chains, and we carry out a MD-based analysis of 4 fluorinated derivatives of L-ethylglycine. We show that polarization of covalent bonds upon fluorination leads to such a redistribution of electronic density within fluoromethylated groups, so that both fluorine and hydrogen atoms establish hydrogen bonds, which are weaker than in normal protein environments. Libraries for 4 fluorinated L-amino acids (4-monofluoroethylglycine, 4,4-difluoroethylglycine, 4,4,4-trifluoroethylglycine, 4,4-difluoropropylglycine) have been created for the AMBER MD package. Their side chain volumes and ASA have been estimated and compared to those of canonical amino acids. MD simulations performed with these libraries and a Ace-(fluorinated residue)-Nme dipeptide system have been used to characterize backbone and side chain conformational

preferences of these fluorinated amino acids. We find that MfeGly and TfeGly have high α-helical and especially pronounced left α-helical propensities, while DfeGly and DfpGly prefer a β-strand conformation. Side chain rotamer energy surfaces are explained by electrostatic properties of the fluoromethylated groups and their steric demand. Hydration analysis of Ace-FXR-Nme dipeptides shows that for the determination of the solvation properties of fluorinated amino acids there is a trade-off between polar properties of fluorine atoms and the increase of ASA upon fluorination, which is confirmed by the retention times obtained experimentally. The studied fluorinated amino acids are found to have more favorable hydration energy than their non-fluorinated analogues. The analysis of the RDF function for solvent in the MD simulations shows clearly the existence of weak hydrogen bonds between water molecules and fluorine atoms of side chains of fluorinated amino acids, and is in agreement with our QM data. Our results provide new insights into the understanding of the properties of fluorine within protein environments, which may assist in exploiting the full potential of fluorine's unique properties for applications in the field of protein engineering.

3.2 Position dependent effects of fluorinated amino acids on hydrophobic core formation of a heterodimeric coiled-coil

by Mario Salwiczek, Sergey Samsonov*, Toni Vagt, Elisabeth Nyakatura, Emanuel Fleige, Jorge Numata, Helmut Cölfen, M. Teresa Pisabarro, and Beate Koksch

Accepted for publication in Chemistry – A European Journal 2009, XXXXXXX

* first co-authorship (my contribution: computational part)

3.2.1 Abstract

Systematic model investigations of the molecular interactions of fluorinated amino acids within native protein environments substantially improve our understanding of the unique properties of these building blocks. A rationally designed heterodimeric coiled-coil peptide (VPE/VPK) and nine variants containing amino acids with variable fluorine content in either position a16 or d19 within the hydrophobic core were synthesized and used to evaluate the impact of fluorinated amino acid substitutions within different hydrophobic protein microenvironments. The structural and thermodynamic stability of the dimers were examined by applying both experimental (CD spectroscopy and analytical ultracentrifugation) and theoretical (MD simulations and MM-PBSA free energy calculations) methods. The coiled-coil environment imposes position dependent conformations onto the fluorinated side chains and thus affects their packing and relative orientation towards their native interaction partners. We find evidence that such packing effects exert a significant influence on the contribution of fluorine-induced polarity to coiled-coil folding.

3.2.2 Introduction

The widespread interest in peptides and proteins as highly potent pharmaceuticals [219] as well as bio-inspired materials [220] motivates attempts towards the *de novo* design of peptides and proteins with superior properties such as chemical and metabolic resistance as well as thermodynamic stability [221]. Moreover, endowing these biomolecules with novel functions that are not carried out by natural proteins [222] is perhaps one of the most interesting, albeit challenging prospects in protein engineering [223]. To this end, continuous efforts are made to expand the repertoire of genetically encoded amino acids through manipulation of the translational machinery in vitro and in vivo [224,225]. Also, pure synthetic and semi-synthetic approaches, i.e. the direct chemical modification of protein functional groups [226] as well as solid phase peptide synthesis [227], native chemical ligation [228], and expressed protein ligation [229] enable the incorporation of non-natural amino acids into peptide and protein sequences. In this context, fluorinated amino acids have increasingly gained

recognition as analytical probes and modulators for protein structure and stability [178].

Organic molecules containing C-F bonds display unique properties [185] that account for their ever-growing importance in medicinal chemistry [230]. Most prominent amongst these is a pronounced enhancement in steric size upon fluorination of alkyl groups that is combined with the very low polarizability of the fluorine atom. This often, although not generally [174], leads to a manifold increase in hydrophobicity and thus improves membrane permeability [231]. It has been anticipated that global replacement of hydrophobic amino acids in hydrophobic domains with fluorinated analogs would accordingly stabilize the structure of proteins. As summarized in a recent review [232] this has been proven to be a successful concept for the design of hyperstable α-helical coiled-coils. Along with enhanced self-association behaviour, some of these peptides display an increase in membrane binding affinity [181,233] that lead to the design of fluorinated peptides with enhanced antimicrobial activity [234,235]. The attempt towards a global replacement of leucine residues by fluorinated analogs within globular proteins, however, resulted in reduced thermodynamic stability [236,237]. In these cases, additional mutations were needed to compensate for the disadvantageous effects [238]. It was also shown that fluorination of aromatic side chains within proteins does not generally enhance secondary structure formation [239]. These findings suggest that properties other than hydrophobicity may also play an important role in directing the interactions of fluorine within native protein environments. Though a weak electron donor and thus poor hydrogen bond acceptor[240], carbon-bound fluorine has been shown to participate in favorable multipolar interactions within native protein environments [241]. It is also important to note that despite the fact that specific fluorine-fluorine interactions are able to promote ordered self-association [232], it has been proposed that they may also result in misfolding [242,236]. In addition, our previous studies suggest that hydrophobic interactions in proteins may be severely disturbed by fluorine-induced polarity [242]. In summary, it still seems rather difficult to predict the impact of fluorination on the structure and activity of peptides and proteins. To further investigate the impact of fluorine substitution in native protein environments, we designed a heterodimeric α-helical coiled-coil peptide containing one fluorinated amino acid at either of two positions within the hydrophobic core, which are different in terms of side chain packing. We find that the effect of fluorine-induced polarity highly depends on the microenvironment of the substitution.

3.2.3 Methodology

Materials. Fmoc-Glu(OtBu)- and Fmoc-Lys(Boc)-NovaSyn®-TGA resins (0,16 mmol g^{-1} and 0,21 mmol g^{-1}, respectively) were purchased from Novabiochem. Fmoc-L-amino acids, 2-(1*H*-benzotriazol-

1-yl)-1,1,3,3-tetramethyluronium tetrafluoroborate (TBTU), and 1-hydroxybenzotriazole (HOBt) were purchased from Fa. Gerhardt (Wolfhagen, Germany). 1-Hydroxy-7-azabenzotriazole (HOAt) was purchased from Iris Biotech, and Fmoc-protected (*S*)-2-aminobutyric acid (Abu) from Bachem. (*S*)-2-amino-4,4,4-trifluorobutyric acid (TfeGly) [243], (*S*)-2-amino-4,4-difluorobutyric acid (DfeGly) [244], and (*S*)-2-amino-4,4-difluoropentanoic acid (DfpGly) [245] were prepared according to literature procedures. Dimethylformamide (p.a., Acros), N,N-diisopropylethylamine (DIEA 98+%, Acros), N,N-diisoporopylcarbodiimide (DIC 99%, Acros), trifluoroacetic acid (TFA 99%, Acros), sodium perchlorate (p.a., Acros), triisopropylsilane (TIS 99%, Acros), piperidine (99% extra pure, Acros), acetonitrile (HPLC gradient grade, Acros), 1,8-diazabicyclo[5.4.0]undec-7-en (for synthesis, Merck), di-sodium hydrogenphosphate dihydrate (p.a, Merck), and sodium dihydrogenphosphate dihydrate (ultra >99%, Fluka) were used without further purification. Acetic anhydride (99%, Acros) was distilled prior to use. Deionized water for buffer solutions and HPLC was prepared using the MilliQ®-AdvantageA10®-System (Millipore). Water (solvent A) and acetonitrile (solvent B) for RP-HPLC were supplemented with 0.1 % TFA (Uvasol®, Merck).

Peptide Synthesis, Purification, and Characterization. Peptides were synthesized using a SyroXP-I peptide synthesizer (MultiSynTech GmbH, Witten, Germany) on a 0.05 mM scale according to standard Fmoc/tBu chemistry [246]. For standard couplings a four fold excess of amino acids and coupling reagents (TCTU/HOBt) as well as an eight fold excess of DIEA relative to resin loading was used. All couplings were performed as double couplings (30 minutes). Our initial attempts to synthesize the peptides produced poor purities and low yields of often less than 10 mg that we attributed to on-resin aggregation during the coupling step. Accordingly, we prevented aggregation by adding a chaotropic agent. For syntheses with the SyroXP synthesizer the amino acids are usually dissolved in a 0.5 M HOBt solution (in DMF). We additionally supplemented these solutions with either 0.8 M lithium chloride or 0.8 M sodium perchlorate with only the latter being effective in increasing the purity of the crude products. Its final concentration in the coupling mix was 0.23 M. The yield of the pure peptides could be increased to generally more than 20 mg. Fluorinated amino acids as well as the first subsequent amino acid were activated by means of DIC/HOAt (1/1) protocols (seven minutes preactivation) without the addition of base to prevent racemization [247]. The molar excess of amino acid and coupling reagents was reduced for fluorine-containing residues to 1.5 fold for the first and 0.8 fold for the second coupling. These couplings were performed manually until completion indicated by a negative Kaiser test [248]. Prior to deprotection possibly non-acylated N-termini were capped by

adding a mixture of acetic anhydride and DIEA (10% each) in DMF (3 x 10 min). A mixture of DBU and piperidine (2% each) in DMF was used for Fmoc-deprotection (4 x 5 minutes). Peptides were cleaved from the resin by treatment with 4 mL TFA/TIS/H2O (95/2.5/2.5). The resins were washed twice with TFA (1 mL) and dichloromethane (dry, 1 mL) and excess solvent was removed by evaporation. The peptides were precipitated with cold Et2O. Purification was carried out by RP-HPLC (Phenomenex® Luna C8, 10 µm, 250 mm x 21.2 mm) and the purity was confirmed by analytical HPLC (Phenomenex® Luna C8, 5 µm, 250 mm x 4.6 mm). All products were identified by high resolution ESI-MS (see Table 3.2.1). To identify the products high resolution mass spectra were recorded on the Agilent 6210 ESI-TOF mass spectrometer (Agilent Technologies, Santa Clara, CA, USA.) The samples were dissolved in acetonitrile/water (1/1) containing 0.1 % TFA and injected directly into the spray chamber using a syringe pump with flow rates of 10 to 50 μL/min. The spray voltage was 4.000V and the drying gas (N2) flow rate was set to 1 psi (1 bar). Peptide concentrations were determined using the absorbance of o-aminobenzoic acid (λ_{max} = 320 nm at pH 7.4) attached to the N-terminus of each peptide (see supporting information).

Table 3.2.1. Identification of the peptides by ESI-TOF mass spectrometry[a]

Peptide	Calc.$[M+4H]^{4+}$	Obs $[M+4H]^{4+}$
VPE	948.274	948.274
VPE-*N*YNO$_2$	970.527	970.529
VPK-*C*Abz	979.578	979.562
VPK	947.801	947.803
VPK(Leu16)	951.304	951.293
VPK(Abu16)	944.297	944.291
VPK(Abu19)	940.793	940.792
VPK(DfeGly16)	953.292	953.287
VPK(DfeGly19)	949.788	949.789
VPK(TfeGly16)	957.789	957.793
VPK(TfeGly19)	954.286	954.285
VPK(DfpGly16)	956.796	956.790
VPK(DfpGly19)	953.292	953.291

[a] If not stated otherwise all the peptides bear an N terminal Abz label. For VPK-*C*Abz the Abz label is attached to an additional lysine at the C terminus.

Determination of Peptide Concentration. Concentrations were estimated by UV spectroscopy on a Cary 50 UV/Vis spectrometer (Varian) using the absorption of o-aminobenzoic acid attached to each N-terminus. A calibration curve (Figure 3.2.1) was recorded using different concentrations of H_2N-Abz-

Gly-COOH · HCl (Bachem) in the buffer used for CD spectroscopy containing 6M guanidinium hydrochloride (Fluka). Disposable Plastibrand® PMMA cuvettes (Brand GmbH, Germany) with path lengths of 1 cm were used.

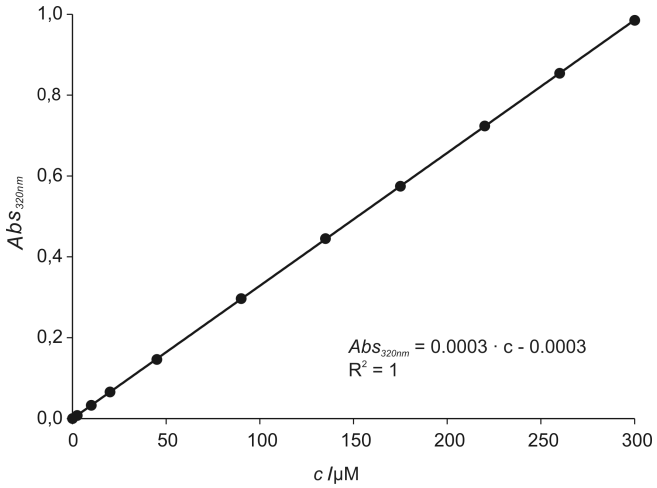

Figure 3.2.1. Calibration curve for the determination of peptide concentrations recorded at 20°C (100 mM phosphate buffer, 6M GdnHCl, pH 7.4).

Circular Dichroism. CD-spectra were recorded on a Jasco J-715 spectropolarimeter at 20 °C (Jasco PTC-348WI peltier thermostat). Overall peptide concentrations were 20 µM (10 µM VPE and 10 µM VPK) at pH 7.4 (100 mM phosphate buffer). CD-spectra were obtained in the far-UV range (190 nm - 240 nm) using 0.1 cm Quartz Suprasil® cuvettes (Hellma) equipped with a stopper. The nitrogen flow rate was set to 3 L/min. Ellipticity was normalized to concentration (c/mol L^{-1}), number of residues (n = 35, including the N-terminal label) and path length (l cm) using eq 3.2.1:

$$[\theta] = \frac{\theta_{obs}}{10000 \cdot l \cdot c \cdot n} \tag{3.2.1}$$

where θ_{obs} is the measured ellipticity in millidegrees and $[\theta]$ the normalized ellipticity in 10^3 deg cm^2 dmol^{-1} residue^{-1}. Melting curves were recorded using the signal at 222 nm applying a heating rate of 3 K min^{-1} from 20°C to 100°C. Each sample was prepared three times and both the baseline corrected spectra and the melting curves were averaged.

Fluorescence Resonance Energy Transfer. We carried out the FRET assay according to previously published procedures [249] using o-aminobenzoic acid (Abz: λ_{ex} = 320 nm, λ_{em} = 420 nm, Bachem) as the fluorescence label and 3-nitrotyrosine (YNO_2: λ_{abs} = 420 nm, Bachem) as the quencher [250]. Three peptides were synthesized: VPK carrying the Abz-label at either the N- or the C-terminus and VPE carrying YNO_2 at the N-terminus. Fluorescence spectra were recorded on a luminescence spectrometer

LS 50B (Perkin Elmer) using a 1 cm Quartz Suprasil® cuvette (Hellma) at 20°C. Three scans from 350 to 550 nm were performed averaged and the spectra were normalized to the respective maximum fluorescence.

Calculation of thermodynamic parameters. Thermodynamic parameters were determined by non-linear least square fitting of the normalized CD-melting curves to six parameters (a, b, $[\theta]_M(0)$, $[\theta]_D(0)$, ΔH_m, and T_m) assuming a two-state monomer-dimer equilibrium. The fits were performed as follows. Ellipticity can be calculated from the fraction unfolded (f_u) according to eq 3.2.2:

$$[\theta] = ([\theta]_M - [\theta]_D) \cdot f_u + [\theta]_D \tag{3.2.2}$$

where $[\theta]_M$ represents the linear temperature dependance of the ellipticity of the fully unfolded monomers M (eq 3.2.3), and VPK and VPE are mathematically regarded as equal [251]. $[\theta]_D$ is the linear temperature dependance of the ellipticity of the fully folded dimer D (eq 3.2.4):

$$[\theta]_M = a \cdot t + [\theta]_M(\cdot) \tag{3.2.3}$$

$$[\theta]_D = b \cdot t + [\theta]_D(0) \tag{3.2.4}$$

Here, t is the temperature in °C and $[\theta]_M(0)$ as well as $[\theta]_D(0)$ represent the hypothetical ellipticity values for the unfolded and the folded peptides at 0°C. The fraction unfolded can be expressed in terms of equilibrium constant (eq 3.2.5) after solving the equation for a bimolecular reaction D ⇄ 2M:

$$f_u = \frac{\sqrt{١٦K[D]_0 + K^2} - K}{٨[D]_0} \tag{3.2.5}$$

where K is the equilibrium constant and $[D]_0$ the concentration of the fully folded dimer. The temperature dependence of K is expressed by eq 3.2.6:

$$K = e^{-\Delta G / R \cdot T} \tag{3.2.6}$$

The Gibbs-Helmholtz equation can be used to express the temperature dependence of ΔG in terms of ΔH_m and T_m as given by eq 3.2.7:

$$\Delta G = \Delta H_m \cdot \left(1 - \frac{T}{T_m}\right) + \Delta C_p \cdot \left\{T - T_m - T \cdot \ln\left(\frac{T}{T_m}\right)\right\} \tag{3.2.7}$$

where ΔH_m is the enthalpy change at the melting temperature T_m, that is defined as the temperature at which fu = 0.5. ΔC_p is the change in heat capacity that was initially assumed to be zero for the purpose of fitting because due to the high interdependence of ΔH and ΔC_p these parameters cannot be fitted simultaneously. Equations 3.2.2 through 3.2.7 were combined and the data fitted directly. ΔC_p was calculated afterwards from the dependance of ΔH_m from T_m and the standard free energy of unfolding ΔG^o (1M standard state) was then calculated at 25°C according to eq 3.2.8:

$$\Delta G^{\phi} = \Delta H_m \cdot \left(1 - \frac{T}{T_m}\right) + \Delta C_p \cdot \left\{T - T_m - T \cdot \ln\left(\frac{T}{T_m}\right)\right\} - RT \ln 2[D]. \tag{3.2.8}$$

Errors were determined by a statistical analysis of the fitted parameters [252]. The error for the free energy of unfolding was calculated using eq 3.2.8 applying the minimum and maximum values for ΔH_m, ΔC_p, and T_m according to their individual errors. To prove the validity of the fit, ΔH_m and T_m were also determined manually using the Van't Hoff equation[253].

Analytical ultracentrifugation. Analytical ultracentrifugation (AUC) was performed on a XL-I (Beckman-Coulter, Palo Alto, CA) ultracentrifuge at 25 °C applying the UV-Vis absorption optics at 320 nm and using standard 12 mm double sector centerpieces. Sedimentation velocity experiments were performed at 60000 rpm and a sample concentration of 50 µM, sedimentation equilibrium experiments at 40000 rpm. The samples were dissolved in 100 mM phosphate buffer at pH 7.4 (ρ = 1.009942 g ml^{-1}, η = 0.9243 cP both at 25 °C). The partial specific volume of the samples was determined in a density oscillation tube (DMA 5000, Anton Paar, Graz) to be 0.730 ml g^{-1} for VPK and 0.594 ml g^{-1} for VPE. The partial specific volume of the VPE-VPK heterodimer was selected as the arithmetic average to be 0.662 ml g^{-1}. Apparent weight average molar masses were determined concentration dependent from sedimentation equilibrium experiments using the model independent MSTAR approach[254]. Sedimentation velocity data were evaluated using the program SEDFIT by P.

Schuck[255] yielding the diffusion corrected molar mass distribution *c*(M).

MD simulations and MM-PBSA free energy calculations. The crystal structure of the Sir4p C-terminal coiled-coil at 2.5 Å resolution (PDB ID: 1pl5) was used as template for modeling our parallel coiled-coil systems. To obtain the parent peptide model system (VPE-VPK) and its nine fluorine-substituted variants the length of the helices of the Sir4p coiled-coil was reduced to 34 aa, and the necessary side chain substitutions were carried out with the MOE program [203]. The structures were solvated in a TIP3P water octahedral box. MD simulations performed with AMBER 8.0[8] using the ff03 force field were preceded by two energy minimization steps: 500 cycles of steepest descent and 1000 cycles of conjugate gradient with harmonic force restraints on protein atoms, then 1000 cycles of steepest descent and 1500 cycles of conjugate gradient without constraints. This was followed by heating of the system from 0 to 300K for 10 ps, and a 30 ps MD equilibration run at 300K and 10^6 Pa in isothermal isobaric ensemble (NPT). Following the equilibration procedure, 5 ns of productive MD runs were carried out in NPT ensemble with Langevin temperature coupling with collision frequency parameter γ=1 ps^{-1} and Berendsen pressure coupling with a time constant of 1.0 ps. The SHAKE algorithm was used to constrain all bonds that contain hydrogen atoms. A 2 fs time integration step was used. An 8 Å cutoff was applied to treat non-bonded interactions, and the Particle Mesh Ewald (PME) method was introduced for long-range electrostatic interactions treatment. MD trajectories were recorded each 2 ps. For the analysis of the trajectories PTRAJ module was used. Non-standard amino acid residues were parameterized to be compatible with the Cornell force field using a standard procedure for non-natural amino acids [256-259] in the R.E.M. III program, which we used for RESP charge calculations [260]. For each amino acid charges were derived for two conformations (helical and extended) with the ab initio Hartree-Fock method HF/6-31G* using GAMESS-US [201] (the authors can provide derived charges information upon request). Energetic post-processing of the trajectories was done in a continuous solvent model as implemented in the AMBER 8.0 MM-PBSA module. The snapshots for the calculations were chosen as described by Lafont and coworkers [126]. Entropies were calculated using normal mode analysis. Significant comparison of the free energies of interaction between two coiled-coils is not possible because of the intrinsic flexibility of the helices termini. To avoid this additional source of noise in the MM-PBSA calculations only the central parts of the helices were analyzed (residues 10 to 25). Thus, taking into account the reduced size of our model system, only the comparison of relative values of energies with experimental data is reasonable.

3.2.4 Results and discussion

The aim of this study was to evaluate how fluorinated amino acids interact with native residues in a natural protein environment. A previously reported *de novo* designed α-helical coiled-coil interaction motif was shown to sufficiently fulfill the requirements for an appropriate model system [178]. Besides being of paramount biological importance [261,262], the coiled-coil's greatest advantage is that it provides two very well defined recognition surfaces [263]. Its primary structure is based on a repetitive pattern of seven amino acids, the heptad repeat $(abcdefg)_n$. Along the helical surface, the hydrophobic positions **a** and **d** and the mostly polar positions **b**, **c**, and **f** point in opposite directions. The **a-** and **d-**residues of the interacting helices are packed in a zipper-like fashion to form the hydrophobic core while all the other heptad positions are solvent exposed. The perfect interactions within the hydrophobic core provide the basis for a stable fold and drive oligomerization. In consequence the peptides associate to form a slightly left-handed superhelix. In dimeric coiled-coils positions **e** and **g** are preferably populated by charged residues that further contribute to stability and control the specificity of folding by forming interhelical salt bridges. Following this primary structure code coiled-coils of different length and oligomerization specificity can be designed de novo [263]. Because the packing of the hydrophobic side chains in a parallel coiled-coil, **a** against **a'** and **d** against **d'**, is not equivalent in terms of relative side chain orientation (*vide infra*) [264], a parallel design as presented below can be used to study the impact of fluorination within two different hydrophobic microenvironments.

The model system VPE/VPK was designed to provide the environment for specific interactions between a fluorinated and a non-fluorinated peptide. The peptide model fulfils two important criteria: 1) specificity for one distinct orientation of the peptide strands within the dimer and 2) heterodimerization. Figure 3.2.2 illustrates the design of the model peptide. The amino acid composition of the hydrophobic core is inspired by the GCN4 transcription factor, which has already been extensively characterized at high resolution [164]. Here, valine in all of the a- and leucine in all of the d-positions provide for a parallel orientation of the peptide strands in the coiled-coil dimer.

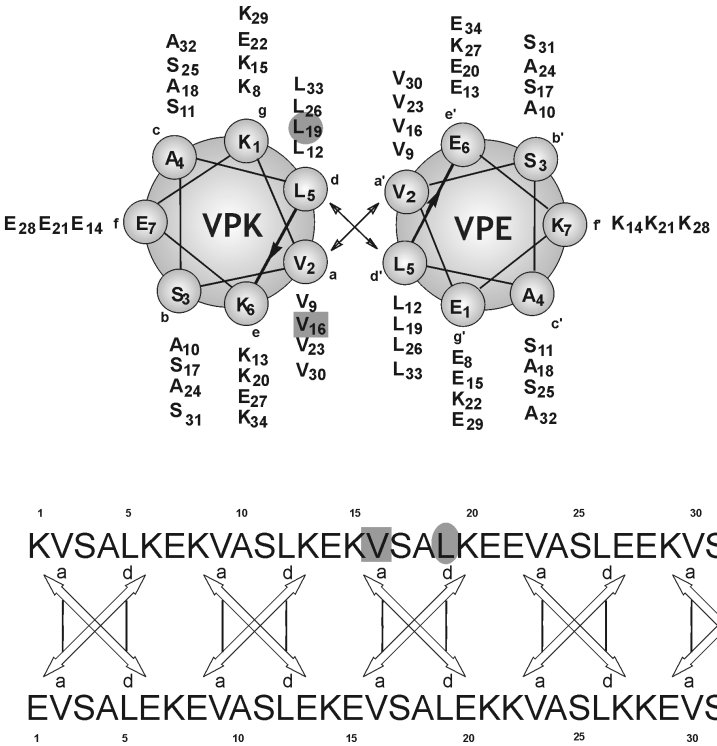

Figure 3.2.2. Amino acid sequence and helical wheel representation of the heterodimeric coiled-coil model system. Two series of peptides were synthesized - one that contains the fluorinated amino acid at position **a16** (grey box) and one that contains it at position **d19** (grey circle) within VPK. Each peptide carries Abz at its N-terminus (not shown).

Most important for the purpose of the study, heterodimerization is required to guarantee that the observed effects trace back to a single fluoroamino acid substitution per dimer. This condition is accomplished by introducing **e-g'** and **g-e'** pairs that engage in favorable electrostatic interactions in the heterodimer but would repel one another in both possible homodimers. The fully natural **VPE** peptide was then used as a template to screen the interactions with different fluorine-containing variants of the complementary interaction partner **VPK**.

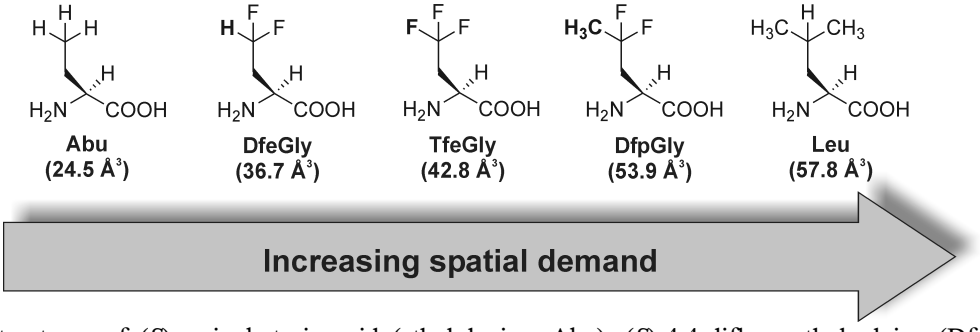

Figure 3.2.3. Structures of (*S*)-aminobutyric acid (ethylglycine, Abu), (*S*)-4,4-difluoroethylgylcine (DfeGly), (*S*)-4,4,4-trifluoroethylglycine (TfeGly), (*S*)-4,4-difluoropropylglycine (DfpGly) and native leucine. The VdW-volumes given in parentheses correspond to the alkyl groups that are attached to the β-carbon and were calculated according to Zhao et al. [217].

As mentioned above, the packing characteristics of the **a**- and **d**-positions in parallel coiled-

coils are different. Therefore, the **VPK** strand contains the fluorinated amino acids either at position **a16** or **d19**, which allows evaluating the impact of fluorination within two different hydrophobic microenvironments. Peptides containing leucine at the respective substitution site served as the reference peptides.

FRET and Analytical Ultracentrifugation. The parent peptides VPE and VPK were used to verify the parallel heterodimerization of the model system. In order to determine the relative orientation of the helices we applied a FRET-assay using o-aminobenzoic acid (Abz) as the fluorescence donor and 3-nitrotyrosine (YNO_2) as the acceptor [250]. Resonance energy transfer from Abz to YNO_2 only occurs when the donor and the acceptor are in close proximity. For a parallel alignment, this condition is fulfilled when donor and acceptor are attached to the respective N-termini of VPK and VPE.

Figure 3.2.4 A shows the fluorescence spectra of N-terminally Abz-labeled VPK (VPK-*N*Abz, where Abz is attached to the N-terminus) at different concentrations of N-terminally YNO_2-labeled VPE (VPE-*N*YNO_2). The spectra show a progressive decrease in fluorescence intensity as the concentration of VPE increases. A similar experiment in which the fluorescence donor Abz was present at the C-terminus of VPK (VPK-*C*Abz) shows much weaker quenching (Figure 3.2.4 B) and confirms that VPE and VPK preferentially form parallel heterooligomers. Furthermore, control experiments in the presence of a denaturant (GdnHCl) demonstrated that the quenching shown in figure 3.2.4 A is the result of specific folding rather than self-quenching. Accordingly to these experiments, the fluorescence should be recovered upon chemically induced unfolding. Figure 3.2.5 shows the fluorescence intensity of a mixture of VPK-NAbz (150 µg/mL) and VPE-NYNO2 (300 µg/mL) at different concentrations of guanidinium hydrochloride.

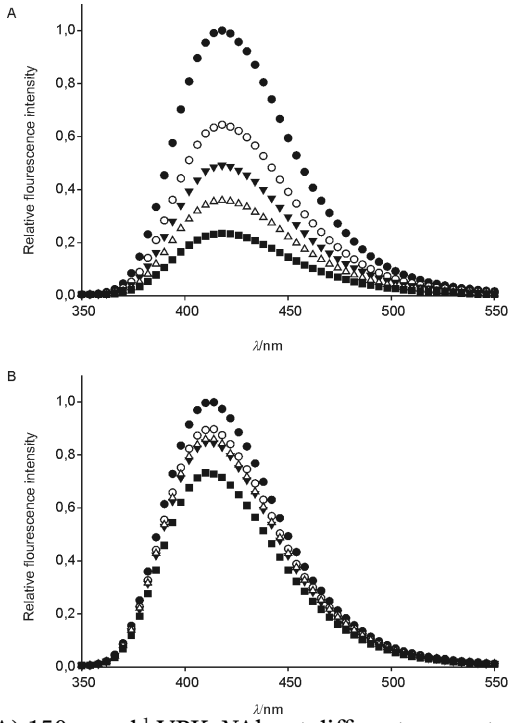

Figure 3.2.4. Fluorescence spectra of A) 150 µg ml^{-1} VPK-*N*Abz at different concentrations of VPE-*N*YNO2 and B) 150 µg mL^{-1} VPK-*C*Abz at different concentrations of VPE-*N*YNO$_2$: (●) 0 µg mL^{-1}, (○) 50 µg mL^{-1}, (▼) 100 µg mL^{-1}, (△) 150 µg mL^{-1}, and (■) 300 µg mL^{-1} (λ_{ex}=320 nm).

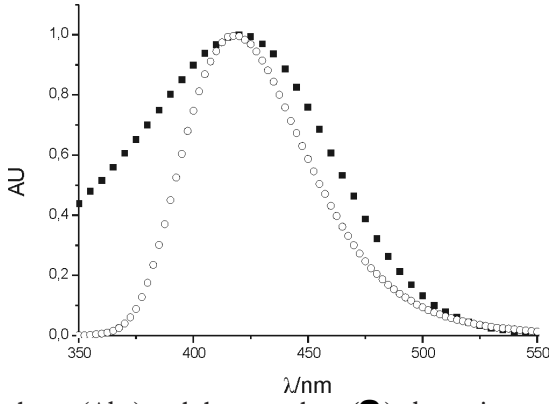

Figure 3.2.5. Spectral overlap of the donor (Abz) and the quencher: (●) absorption spectrum of 20 µM VPE-N-YNO2 and (○) fluorescence spectrum of 20 µM VPK-N-Abz at pH 7.4 (100 mM phosphate buffer). The spectra were normalized.

The oligomerization state of the VPE-VPK heterooligomers was determined by sedimentation velocity and equilibrium experiments. Sedimentation velocity experiments show artificial peak broadening due to insufficient removal of diffusion effects in the evaluation algorithm yielding a molar mass estimate of 7000 g mol^{-1} for the VPE-VPK heterodimer and a monomodal distribution confirming that only heterodimer is present in solution (Figure 3.2.6). This result was confirmed by the absolute molar mass determinations enabled by sedimentation equilibrium measurements, which yielded a M_w of 7600 g mol^{-1} from the extrapolation of five $M_{w,app.}$ to infinite dilution. This molar mass agrees very

well with the expected molar mass for the VPE-VPK heterodimer of 7580.82 g mol^{-1} and confirms the specific heterodimerization of the model system. The formal extrapolation to infinite dilution is necessary to remove the effects of charge and excluded volume on the determined apparent molar mass, which is found too low with increasing concentration due to these non-ideal effects. Although at infinite dilution monomer is to be expected, this formal extrapolation was possible for the investigated concentration range of 100 - 500 µM, since figure 3.2.6 B shows the absence of association or dissociation of the heterodimer in this concentration range.

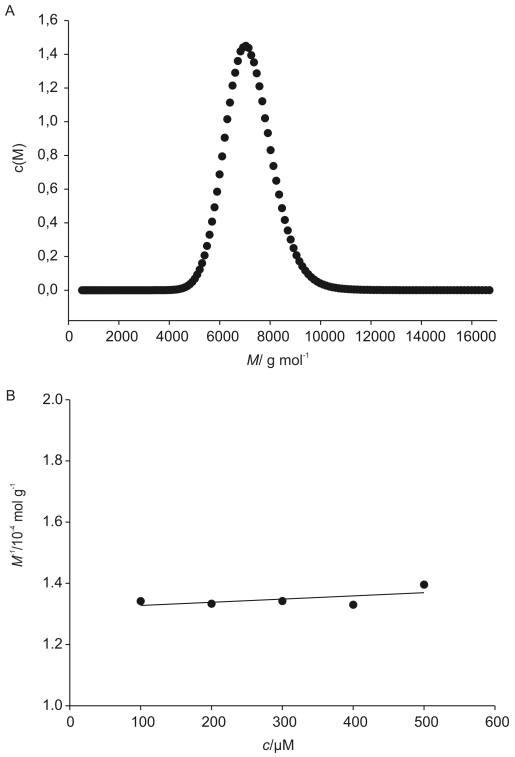

Figure 3.2.6 A) Diffusion corrected molar mass distribution c(M) of the VPE-VPK heterodimer determined for a 50 µM VPE-VPK sample. The peak is broadened due to insufficient removal of diffusion effects. **B)** Concentration dependence of the inverse apparent molar masses $M_{w,app}$. to yield
M_w = 7600 g mol^{-1} by formal extrapolation to infinite dilution (solid line).

CD spectroscopy and MD simulations. All CD-spectra of the equimolar mixtures of VPE and VPK-analogues display distinct minima at 208 and 222 nm at 20°C (Figure 3.2.7), indicating that all peptides form stable α-helical structures. Also, the intensities for all heteromers are very similar, which suggests that the substitution of leucine by Abu and its fluorinated analogues at either position **a16** or **d19** only causes minor structural perturbations. We carried out MD simulations to verify these findings and further support our studies. The results of these experiments show that the structures of all heterodimeric coiled-coils investigated here remain stable in solution.

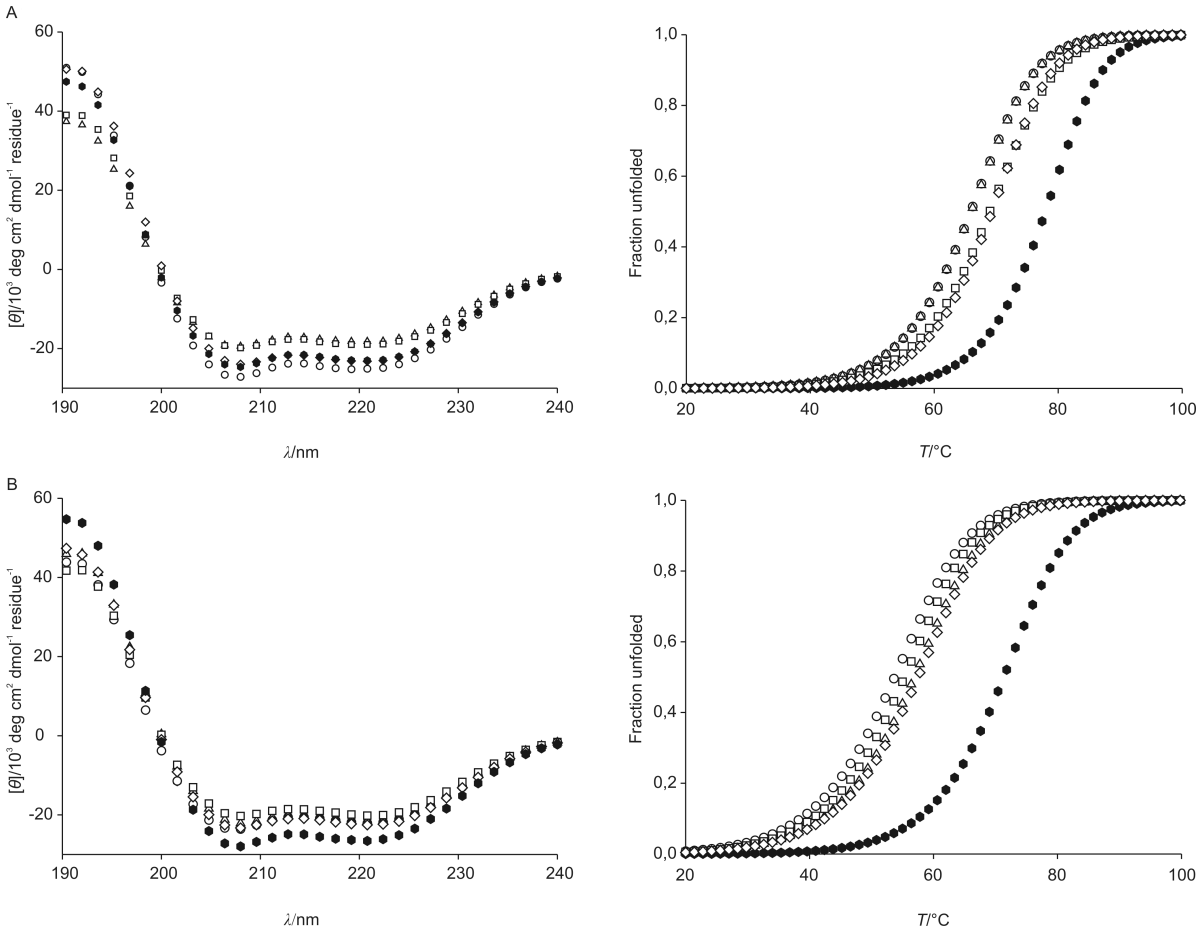

Figure 3.2.7. CD-spectra at 20°C and fitted thermal unfolding profiles of the 1:1 VPE-VPK mixtures substituted at A) position **a16** and B) position **d19** of VPK: (⬢) Leu, (○)Abu, (△)DfeGly, (□) TfeGly, and (◊)DfpGly. Overall peptide concentrations were 20 µM (10 µM in each monomer at pH 7.4, 100 mM phosphate buffer).

The root mean square deviation (RMSD) values for all atoms did not exceed 2.5 Å (2.0 Å for backbone atoms). Moreover, the distances between C_β atoms of the residues in **a**- and **d**- positions in each helix did not fluctuate substantially during simulation except for those at the N- and C-termini (Table 3.2.2). The dihedral angles of all residues are comparable to values for an ideal α-helix (-60º and -45º for φ and ψ, respectively), again with the exception of the N- and C-terminal residues. Although the coiled-coil structure was preserved in all systems, small deviations from ideal α-helical values were found for DfeGly16, DfeGly19 and Abu19. These results point to structural perturbations of the backbone that may, in part, account for the decreased thermodynamic stability of these dimers (*vide infra*).

Table 3.2.2. Distances between C_β atoms in **a**- and **d**-positions in MD simulations

Substitution	a2	d5	a9	d12	a16	d19	a23	d26	a30	d33
Leu16	6.8±1.5	4.7±0.7	5.5±0.4	5.2±0.3	5.2±0.3	4.5±0.3	5.6±0.4	5.0±0.4	6.4±1.2	8.6±3.2
Abu16	5.8±0.8	6.4±1.6	5.8±0.4	5.7±0.3	5.2±0.3	4.0±0.2	5.8±0.2	4.0±0.2	5.7±0.6	4.3±0.5
DfeGly16	5.4±0.4	4.1±0.2	5.6±0.5	4.6±0.6	5.0±0.4	4.1±0.2	5.7±0.3	4.0±0.2	6.0±0.5	5.2±0.9
TfeGly16	11.2±2.9	5.4±0.6	5.0±0.4	5.2±0.5	5.2±0.3	4.1±0.2	5.6±0.4	4.8±0.4	5.9±0.6	5.0±0.4
DfpGly16	5.6±0.4	4.2±0.4	5.6±0.5	4.9±0.4	5.2±0.3	4.0±0.2	5.8±0.2	4.0±0.2	6.0±0.6	5.2±0.6
Leu19	5.9±0.3	4.3±0.5	5.8±0.4	4.8±0.3	5.5±0.5	4.0±0.2	5.8±0.2	4.1±0.4	6.1±0.4	4.2±0.4
Abu19	4.9±0.6	5.3±0.5	5.2±0.3	4.6±0.5	5.4±0.5	4.1±0.4	6.0±0.2	4.4±0.6	5.8±0.6	4.7±1.1
DfeGly19	5.5±0.4	4.2±0.4	5.6±0.5	4.5±0.5	5.4±0.5	4.6±0.5	5.8±0.5	6.0±0.8	6.0±0.8	6.1±3.4
TfeGly19	5.9±0.3	4.1±0.3	5.4±0.4	5.3±0.3	5.1±0.4	4.2±0.4	5.7±0.2	3.9±0.2	6.2±0.4	5.2±0.5
DfpGly19	5.5±0.7	5.1±0.6	5.2±0.4	5.4±0.4	5.2±0.4	4.3±0.4	5.6±0.4	4.0±0.2	5.7±0.6	4.9±0.4

Thermodynamic Characterization. Temperature dependent circular dichroism spectroscopy was used to experimentally probe the thermodynamic stability of the dimers. Non-linear least squares fitting was carried out. The fitted parameters and their errors are shown in Table 3.2.3.

Table 3.2.3. Fitting parameters (and their errors)

VPE/VPK dimers	a^a	b^a	$[\theta]_M(0)^b$	$[\theta]_D(0)^b$	$T_m{}^c$	$\Delta H_m{}^d$	$\Delta G^{\circ d}$
Leu19	-0.02	0.21	-3.28 (0.4)	-31.30 (0.05)	344,48 (0.1)	61.69 (0.7)	11.66 (0.2)
Leu16	-0.02	0.19	-3.17 (0.6)	-27.60 (0.04)	351.09 (0.1)	75.54 (1.1)	13.83 (0.2)
Abu16	-0.01	0.24	-3.67 (0.2)	-31.09 (0.06)	339.04 (0.1)	62.19 (0.8)	11.48 (0.2)
DfeGly16	-0.02	0.17	-1.95 (0.2)	-22.30 (0.06)	339.12 (0.1)	61.99 (1.1)	11.46 (0.1)
TfeGly16	-0.02	0.16	-2.00 (0.3)	-22.95 (0.05)	342.12 (0.1)	61.37 (0.9)	11.51 (0.2)
DfpGly16	-0.01	0.18	-2.74 (0.3)	-27.58 (0.05)	342.45 (0.1)	67.15 (0.8)	12.27 (0.2)
Abu19	0.00	0.15	-4.54 (0.1)	-25.49 (0.05)	326.86 (0.1)	50.71 (0.4)	9.64 (0.1)
DfeGly19	0.00	0.13	-3.98 (0.1)	-24.60 (0.05)	330.05 (0.1)	52.67 (0.3)	9.99 (0.1)
TfeGly19	0.01	0.12	-4.49 (0.1)	-22.97 (0.05)	328.47 (0.1)	52.25 (0.4)	9.87 (0.1)
DfpGly19	0.00	0.15	-4.65 (0.1)	-26.04 (0.04)	330.64 (0.1)	52.45 (0.3)	10.00 (0.1)

[a] in 10^3 deg cm^2 dmol^{-1} residue^{-1} °C^{-1} (errors are smaller than 0.01), [b] in 10^3 deg cm^2 dmol^{-1} residue^{-1}, [c] in K, [d] in kcal mol^{-1}.

The heat capacity change upon unfolding is usually very small. Therefore, the second term of eq 3.2.7 was neglected for the purpose of fitting. An approximate value for ΔC_p was calculated afterwards from the ΔH_m against the melting points (Van't Hoff plot, Figure 3.2.8). The slope of the plot was calculated to be 0.94 ± 0.1 kcal/(mol · K), which corresponds to 0.013 kcal/(mol · K · residue).

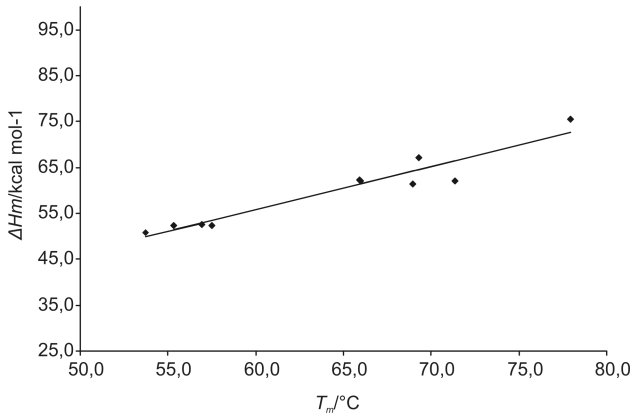

Figure 3.2.8. Van't Hoff Plot of ΔH_m against T_m for all dimers the slope of which yields ΔC_p.

The Van't Hoff equation was used in order to prove the validity of the above described fit. For this method, the baselines of the unfolding transitions were determined manually and T_m was determined at $f_u = 0.5$. ΔH_m was then calculated from $f_u(T)$ using the Van't Hoff equation that has been adapted to the dimer to monomer transition (eq 3.2.9):

$$\Delta H_m = ٦R(T_m)^{٢}\left(\frac{\delta f_u}{\delta T}\right)_{T=T_m} \tag{3.2.9}$$

The values for T_m and ΔH_m are in excellent agreement with those derived from the automated fitting of the denaturing data (Figure 3.2.9).

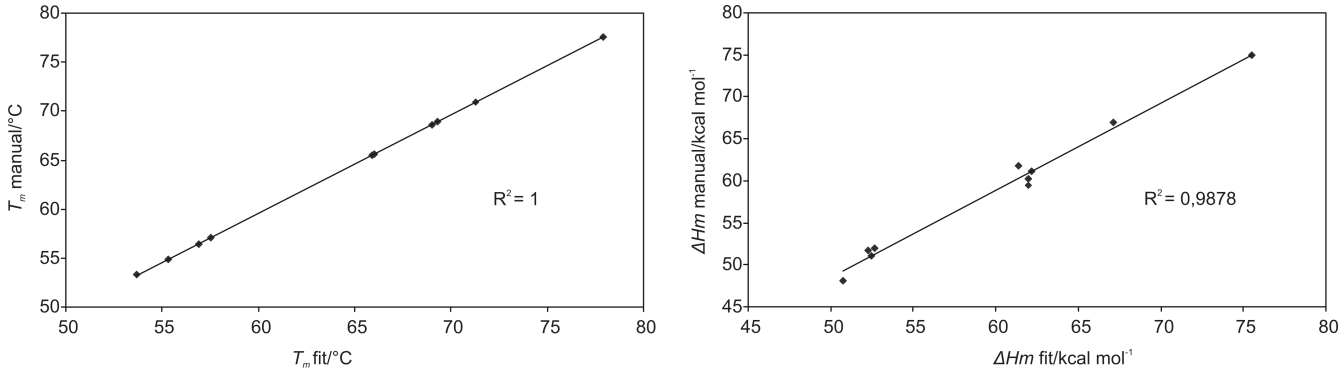

Figure 3.2.9. Plot of the manually determined T_m (left panel) and ΔH_m (right panel) values against those determined by non-linear fitting.

All of the dimers show cooperative thermal unfolding transitions upon heating from 20°C to 100°C (Figure 3.2.7). The thermodynamic parameters of unfolding are summarized in Table 3.2.4. In both positions **a16** and **d19** the substitution of Leu by Abu and its fluorinated analogues considerably decreases the thermodynamic stability of the dimer. Comparison of the stabilities relative to leucine

($\Delta G^{ø}$-$\Delta G^{ø}_{Leu}$), however, shows that in most cases substitution at position **a16** seems to be less tolerated than substitution at position **d19** (Figure 3.2.10). This loss in stability due to considerably removing hydrophobic surface area is partly attenuated by fluorination of the Abu side chain. Furthermore, while a pronounced increase in steric size of the fluorinated side chain by incorporation of DfpGly appears to further stabilize the folding motif at position **a16** the same substitution at position **d19** shows only marginal effects. Most strikingly, the findings for DfpGly contradict previous results for an antiparallel coiled-coil model, where this residue as a replacement for leucine was found to disturb folding even stronger than alanine in an **a-**position [178].

Table 3.2.4. Thermodynamic parameters for the unfolding of the heterodimers substituted at position **a16** and **d19** of VPK.

	Position **a16**		Position **d19**	
Amino Acid	T_m/°C[a]	$\Delta G^{ø}$/kcal mol^{-1} [b]	T_m/°C[a]	$\Delta G^{ø}$/kcal mol^{-1} [b]
Leu	77.9	13.8	71.3	11.7
Abu	65.9	11.5	53.7	9.6
DfeGly	66.9	11.5	56.9	10.0
TfeGly	69.0	11.5	55.3	9.9
DfpGly	69.3	12.3	57.5	10.0

[a] T_m is defined as the temperature at which the fraction unfolded is 0.5. Errors are typically not higher than 0.1 °C. [b] ΔGø values were calculated for the 1M standard state at 25°C using eq 8. The value for ΔC_p was determined from a Van't Hoff plot (see supporting information) to be 0.94 ± 0.1 kcal mol^{-1} K^{-1}. Errors for ΔGø are typically not higher than 0.2 kcal mol^{-1} for the a16 and 0.1 kcal mol^{-1} for the d19 substituted peptides.

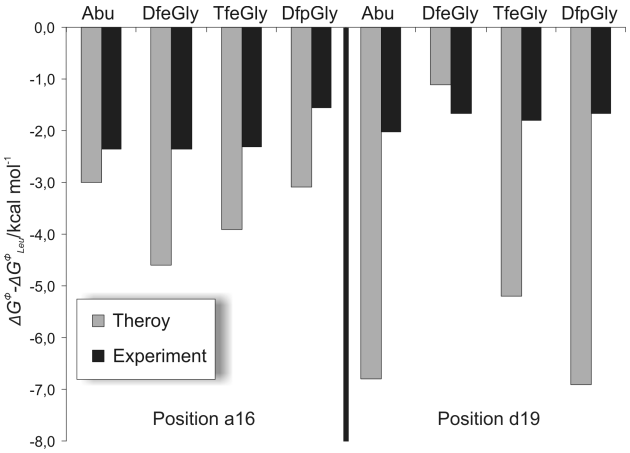

Figure 3.2.10. Relative stabilities of the **a16-** and **d19**-substituted dimers compared to the respective leucine variants as determined by thermal unfolding (black bars) and MM-PBSA analysis (grey bars).

The experimentally observed stability trends and those determined by the MM-PBSA energetic analysis are in agreement (Figure 3.2.10). The adjusted correlations between experimental and

calculated enthalpies and free energies are 0.35 and 0.58, respectively (with significance at the level of p-value 0.05, Figure 3.2.11). Despite relatively low correlation coefficient values, our computational results also qualitatively distinguish substitutions at positions **a16** and **d19** (Figure 3.2.10 and Table 3.2.5).

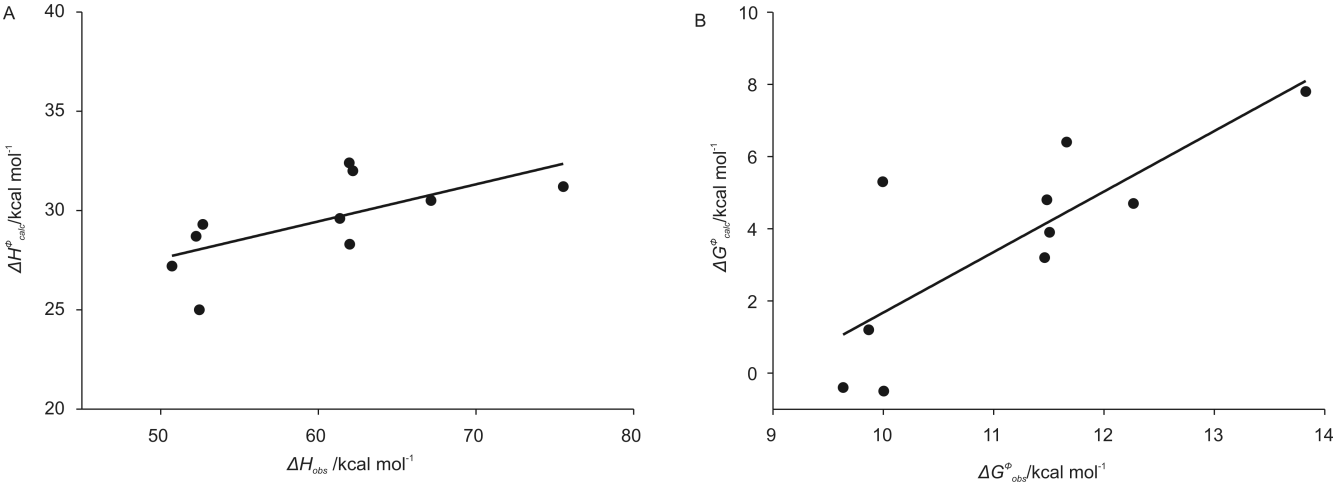

Figure 3.2.11. Correlation of the observed and theoretical thermodynamic parameters of folding: A) enthalpy (adjusted correlation coefficient: 0.35) and B) free energy of unfolding (adjusted correlation coefficient: 0.58).

Table 3.2.5. Calculated free energy components for the unfolding transition

Substitution	ELE[a] kcal mol^{-1}	VDW[b] kcal mol^{-1}	GAS[c] kcal mol^{-1}	PBSUR[d] kcal mol^{-1}	PBCAL[e] kcal mol^{-1}	PBELE[f] kcal mol^{-1}	ΔH kcal mol^{-1}	ΔG kcal mol^{-1}
Leu16	49.3	35.3	84.6	4.1	-57.6	-8.3	31.2	7.8
Abu16	81.5	33.3	114.8	3.9	-86.7	-5.3	32.0	4.8
DfeGly16	57.7	33.8	91.6	4.0	-67.2	-9.5	28.3	3.2
TfeGly16	47.6	35.5	83.1	4.1	-57.5	-9.9	29.6	3.9
DfpGly16	55.7	36.2	91.9	4.1	-65.6	-9.9	30.5	4.7
Leu19	65.5	37.5	103.1	4.2	-74.8	-9.3	32.4	6.4
Abu19	53.4	31.5	84.9	3.9	-61.7	-8.3	27.2	-0.4
DfeGly19	45.3	34.3	79.6	4.0	-54.2	-9.0	29.3	5.3
TfeGly19	58.5	32.3	90.8	3.9	-66.1	-7.6	28.7	1.2
DfpGly19	42.2	30.4	72.7	3.8	-51.5	-9.3	25.0	-0.5

[a]electrostatic energy; [b]van der Waals energy; [c]GAS = VDW + ELE; [d]hydrophobic component of solvation energy; [e]Poisson-Bolzmann reaction energy of the field; [f]PBELE = ELE + PBCAL (full electrostatic energy)

The theory supports the experimental finding that an increasing spatial demand of the fluorinated side chain at position **a16** increases the stability of the dimer, while this trend is not reflected for identical substitutions at position **d19**. However, direct quantitative comparison of experimental and MM-PBSA data is not possible because of three factors: 1) there is an overestimation of the entropic component of the free energy because a single trajectory was used for the entropy

calculation of the bound and unbound complex components [265], 2) the length of the model system was considerably reduced for the energy calculations, which results in a generally stronger impact on stability upon substitutions, and 3) entropy component are still the least accurate of MM-PBSA energy calculations [81].

The general decrease in stability that was observed for the fluorinated peptides may be attributed to several factors. Recent investigations reveal that fluorine-containing amino acids exhibit weaker helix forming propensities than their native counterparts [266]. Reliable thermodynamic scales for helix propensity are essentially measured using isolated helices[267]. Coiled coil stability, however, is substantially determined by interhelical interactions of the **a-** and **d-** positions within the hydrophobic core [268]. For example, Abu in a monomeric helix favors helix formation by approximately 0.08 kcal mol^{-1} [81] compared to Leu but its substitution for Leu within the hydrophobic core of our coiled-coil destabilizes the folding motif by more than 2 kcal mol^{-1}. Our MD-simulations reveal mostly non-significant effects of fluorination on the conformational preferences of the amino acids within this coiled-coil environment. We certainly do not rule out that introduction of fluorine affects helix propensity, but we would assign it less importance in the case of strongly interacting coiled-coil residues. The stability of coiled-coils generally correlates with hydrophobicity and with the spatial demand of hydrophobic side chains in positions **a** [269] and **d** [270]. In addition, the packing characteristics of side chains in both positions are significantly different (Figure 3.2.13) [264]. This difference may explain the general differences between relative stabilities of positions **a** and **d** as shown in Figure 3.2.10. The most striking dissimilarity between the positions is the relative orientation of the C_α-C_β vectors of interacting residues within the dimer. For **a**-positions they point away from each other, whereas they point towards each other for **d**-positions. Interestingly, this happens in all simulated coiled-coil systems, suggesting a key role in the packing differences of **d**- and **a**-positions. For **a**- and **d**- positions the dihedral angles defined by both side chains (i.e. C_α-C_β-C_β'-C_α') were found to be significantly different during the MD simulations (-96±7° and 91±14°, respectively). Also there is an observable difference in C_β- C'_β distances, which is roughly 1 Å shorter for **d**- than for **a**-positions (see supporting information). Figure 3.2.12 exemplarily illustrates the different packing for TfeGly at both substitution positions according to the MD simulations.

The fluorinated amino acids used in this study share a common structural feature, i.e. fluorine substitution at the γ-carbon of the side chain, which results in a significant polarization of the β-methylene groups. According to the different packing characteristics at **a**- and **d**- positions described

above, these β-methylene groups and their corresponding dipoles are closer to their hydrophobic interaction partners at the **d**- than at the **a**-positions. We conclude that fluorine-induced polarity may accordingly have varying degrees of importance for the stability of coiled-coil interactions at these positions. Apparently, the impact is stronger at position **d19** because, unlike for position **a16,** the increase in volume of the fluorinated side chains by methylation (DfpGly) is not able to gain further stability (see Figure 3.2.10 and Table 3.2.5 for experimental and calculations results, respectively). This is because the highly polarized β-methylene group in position **d19** points towards the interaction partner in the opposite strand, while it points away from it at position **a16**. The interpretation that the impact of fluorine-induced polarity in amino acid side chains may depend on the packing and orientation of coiled-coil helices gains further support from the finding that DfpGly in an antiparallel coiled-coil [178] destabilizes the folding motif much stronger than observed here (Figure 3.2.12). This is because the side chains in antiparallel coiled-coils are generally more tightly buried within the core [264] and we concluded that the highly polarized γ-methyl group of DfpGly strongly disturbs hydrophobic interactions. The differences in packing of position **a** in antiparallel and parallel coiled-coil dimers are outlined in figure 3.2.12.

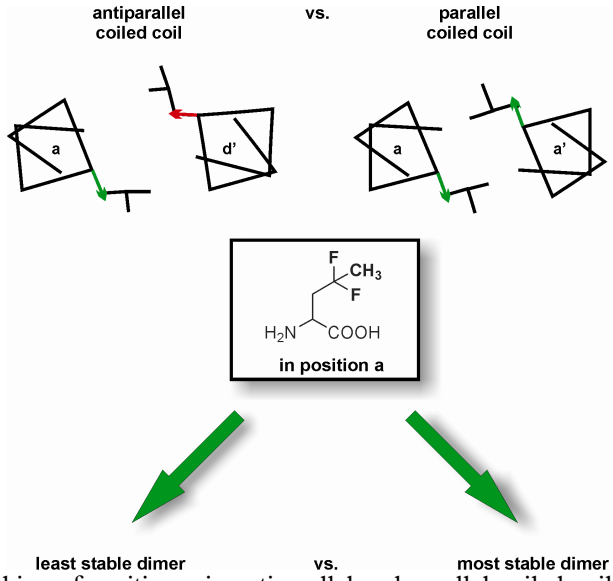

Figure 3.2.12. Differences in packing of position a in antiparallel and parallel coiled-coil dimers and consequences on the stability of DfpGly substitutions.

Our findings for the parallel and the previously reported antiparallel system suggest that the orientation and flexibility of fluorinated side chains within a certain protein environment are additional factors that strongly determine the impact of fluorine-induced polarity. These conclusions are also supported by very recent MD studies by Pendley et al, which also reveal an important role of

electrostatics in the stability of parallel coiled-coil systems containing fluorinated amino acid residues (5,5,5,5’5’5’-hexafluoroleucine) in the hydrophobic core [211].

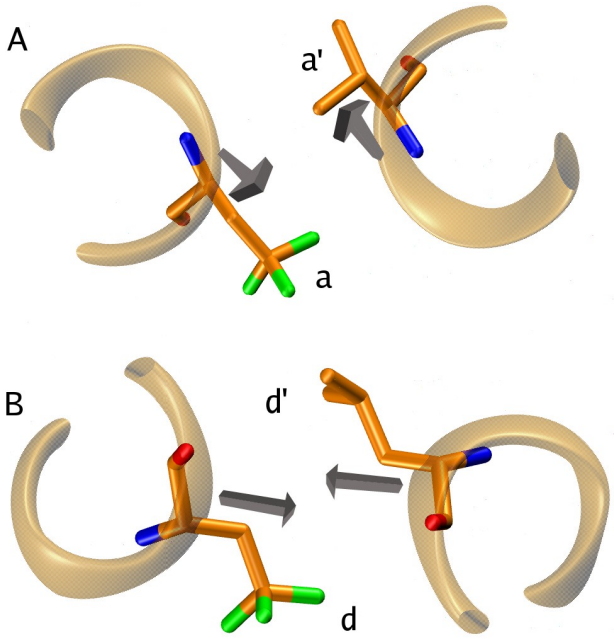

Figure 3.2.13. Packing of TfeGly against its direct interaction partner. A) position **a16** and B) position **d19**. The C_β atoms of the interacting side-chains are closer in the **d**-position (B) than in the **a**-position (A). The displayed C_α-C_β vectors highlight the significantly different packing characteristics of the side chains in **a**- and **d**- positions.

3.2.5 Conclusions

We have shown that the effect of fluorine at different positions within a heptad repeat on the stability of an α-helical coiled-coil can be rather ambiguous. Its effects highly depend on the microenvironment of a certain substitution that, in our case, is defined by both the substitution position and by helix orientation. Although the coiled-coil model is a rather specific folding motif, our results imply that the packing and orientation of fluorinated side chains are very important in determining their interactions with native protein environments. The conception or notion that the introduction of fluorine into proteins necessarily leads to stabilization is clearly disputable according to our results. Changes in fluorine content and position of fluorination can considerably change the polarity and steric properties of an amino acid side chain and, thus, can influence the properties that a fluorinated amino acid develops within a native protein environment. This study shows that not only the fluorine itself, but also the characteristics of the environment determine the consequences of fluorine-induced polarity and steric demand of fluorinated side chains. Such systematic investigations will pave the way towards its directed application in protein engineering, for fine tuning of protein stability, their interactions with peptidic ligands as well as for therapeutical applications.

3.3 Binding of fluorinated peptide substrates in the catalytic center of chymotrypsin

by Matthias Hakelberg, Sergey Samsonov, M. Teresa Pisabarro, and Beate Koksch

to be published

my contribution: computational part

3.3.1 Introduction

Chymotrypsin (bovine γ chymotrypsin: EC 3.4.21.1) is a peptidase participating in proteolysis. Chymotrypsin cleaves substrate peptides after Tyr, Trp and Phe residues, which side-chains fit into a hydrophobic pocket of the enzyme [3]. Chymotrypsin's catalytic mechanism is carried out by a catalytic triad. A catalytic triad commonly refers to the three amino acid residues found inside the active site of certain protease enzymes: Ser, Asp and His. More generally, catalytic triad can refer to any set of three residues that function together and are directly involved in catalysis. The residues of a catalytic triad can be far from each other in the primary structure however are brought close together in the tertiary structure. In chymotrypsin the triad consists of Ser195, Asp102 and His57. Ser195 hydroxyl group binds covalently to the carboxyl carbon atom of Phe, Trp or Tyr, while Asp102 and His57 then hydrolyze the bond [1].

The goal of this work has been to study impact of fluorinated amino acids substitutions in a chymotrypsin's peptidic substrate depending on sequential positions of the substitutions. For this we have used a MD approach and MM-PBSA energy calculations implemented in AMBER. The obtained data have been used to explain experimentally observed trends.

3.3.2 Methodology

Docking. The structure of the receptor was retrieved from Protein Data Bank (bovine alpha-chymotrypsin, ID: 4CHA, 1.68 Å resolution). His57 of the catalytic triad was considered to be protonated at Nє atom. The system was minimized in AMBER 8.0 [8] sander module in TIP3P octahedral water box and periodic boundary conditions for 5000 cycles using steep descent (3000 cycles) and conjugate gradient (2000 cycles) algorithms.

The ligands were built in AMBER xleap module:

Wt : Ace-Ala-Phe-Ala-Ala-Nme

Dfe1: Ace-**DfeGly**-Phe-Ala-Ala-Nme

Tfe1: Ace-**TfeGly**-Phe-Ala-Ala-Nme

Dfp1: Ace-**Dfp**-Phe-Ala-Ala-Nme

Dfe3: Ace-Ala-Phe-**DfeGly**-Ala-Nme

Tfe3: Ace-Ala-Phe-**TfeGly**-Ala-Nme

Dfp3: Ace-Ala-Phe-**Dfp**-Ala-Nme

Dfe4: Ace-Ala-Phe-Ala-**DfeGly**-Nme

Tfe4: Ace-Ala-Phe-Ala-**TfeGly**-Nme

Dfp4: Ace-Ala-Phe-Ala-**Dfp**-Nme,

[DfeGly- difluoroethylglycine, TfeGly- trifluoroethylglycine, DfpGly- difluoropropylglycine.]

For docking, atomic potential grid was calculated in autogrid4 (Autodock 4.0 [132]) with the 0.375Å spacing in the box of size of 40x40x40 Å centered on the OG atom of catalytic Ser195. The docking was done by autodock4 program ($5 \cdot 10^7$ energy evaluations, 150 members in population and 50 individual runs). The results were clustered using g_cluster program in GROMACS [196].

The obtained binding poses were considered as productive if the distance between the main chain carbon atom of ligand's Phe and the side-chain hydroxyl oxygen atom of Ser195 was less than 3.0 Å and the Phe carboxyl oxygen atom pointed into the oxyanionic hole formed by the main chain nitrogen atoms of Ser195 and Gly193. A productive docked pose with the minimal energy was taken as initial structure for a MD simulation.

MD simulations. The system was minimized in AMBER 8.0 sander module in TIP3P octahedral water box and periodic boundary conditions for 5000 cycles using steep descent (3000 cycles) and conjugate gradient (2000 cycles) algorithms. This was followed by heating of the system from 0 to 300 K for 10 ps, and a 30 ps MD equilibration run at 300 K and 10^6 Pa in isothermal isobaric ensemble (NPT). Following the equilibration procedure, 5 ns of productive MD runs were carried out in periodic boundary conditions in NPT ensemble with Langevin temperature coupling with collision frequency parameter $\gamma = 1$ ps^{-1} and Berendsen pressure coupling with a time constant of 1.0 ps. The SHAKE algorithm was used to constrain all bonds that contain hydrogen atoms. A 2 fs time integration step was used. An 8 Å cutoff was applied to treat nonbonded interactions, and the particle mesh ewald (PME) method was introduced for long-range electrostatic interactions treatment. MD trajectories were recorded each 2 ps. Non-standard amino acid residues were parameterized to be compatible with the Cornell force field using a standard procedure for non-natural amino acids in the R.E.M. III program, which we used for RESP charge calculations [260]. For each amino acid charges were derived for two conformations (helical and extended) with the ab initio Hartree-Fock method HF/6-31G* using GAMESS-US [201].

MM-PBSA calculations. For MM-PBSA free energy calculations only the parts of the trajectories with

a productive binding mode were considered. In order to avoid bias towards certain conformations, 10 frames corresponding to electrostatic energy weighted intervals were chosen for the calculations as it is described in Lafont et al [126]. The values for the individual energy components as well as their standard deviations were averaged from three MD runs.

Measuring residues mobility. Average fluctuations of the ligand residues at the positions for mutations were obtained using PTRAJ module of AMBER 8.0. The obtained values were averaged with weights for three MD runs. The weights represented the times of a productive conformation in each run.

3.3.3 Results and Discussion

Docking. For each peptide docking at least 3 (3-9) productive conformations (Figure 3.3.1) out of overall 50 were found. In most cases a productive conformation corresponded to a minimal docking energy. Docking results were clustered by RMSD of main-chain and CB atoms for X-Phe-X-X part of the docked peptides with a 1.2Å cutoff. The number of clusters varied from 11 to 27, with the biggest cluster containing from 7 to 35 members. Productive conformations are predominantly presented in the same clusters. However, because of the flexibility of a productive mode (mostly due to the mobility of the N-terminal residue) productive conformations could have substantial RMSD values (up to 1.3 Å) between each other (also shown in the fluctuation analysis, *vide infra*).

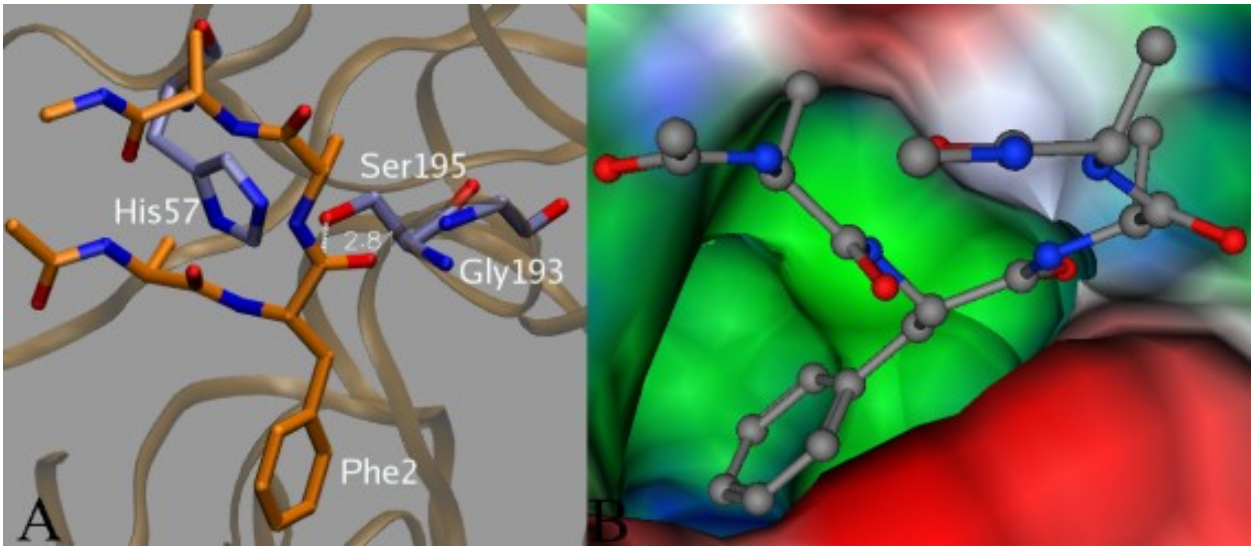

Figure 3.3.1. Productive conformation for Wt peptide obtained in docking. A) Ribbon and licorice representation. Ligand and receptor residues are shown in orange and green, respectively. B) The ligand is shown in balls and sticks representation, the receptor is shown as a surface coloured by electrostatic potential.

MD. In all simulations the peptide ligand remained in the catalytic site, though a productive conformation could have been disrupted during the simulation. The calculated energies and average

residue fluctuations are presented in the Figures 3.3.2 and 3.3.3 and discussed for positions 1, 3, 4. The corresponding binding energies obtained by docking do not correlate with total MM-PBSA energies, but with the MM vacuo component (adjusted correlation coefficient R=58%, p-value=0.05).

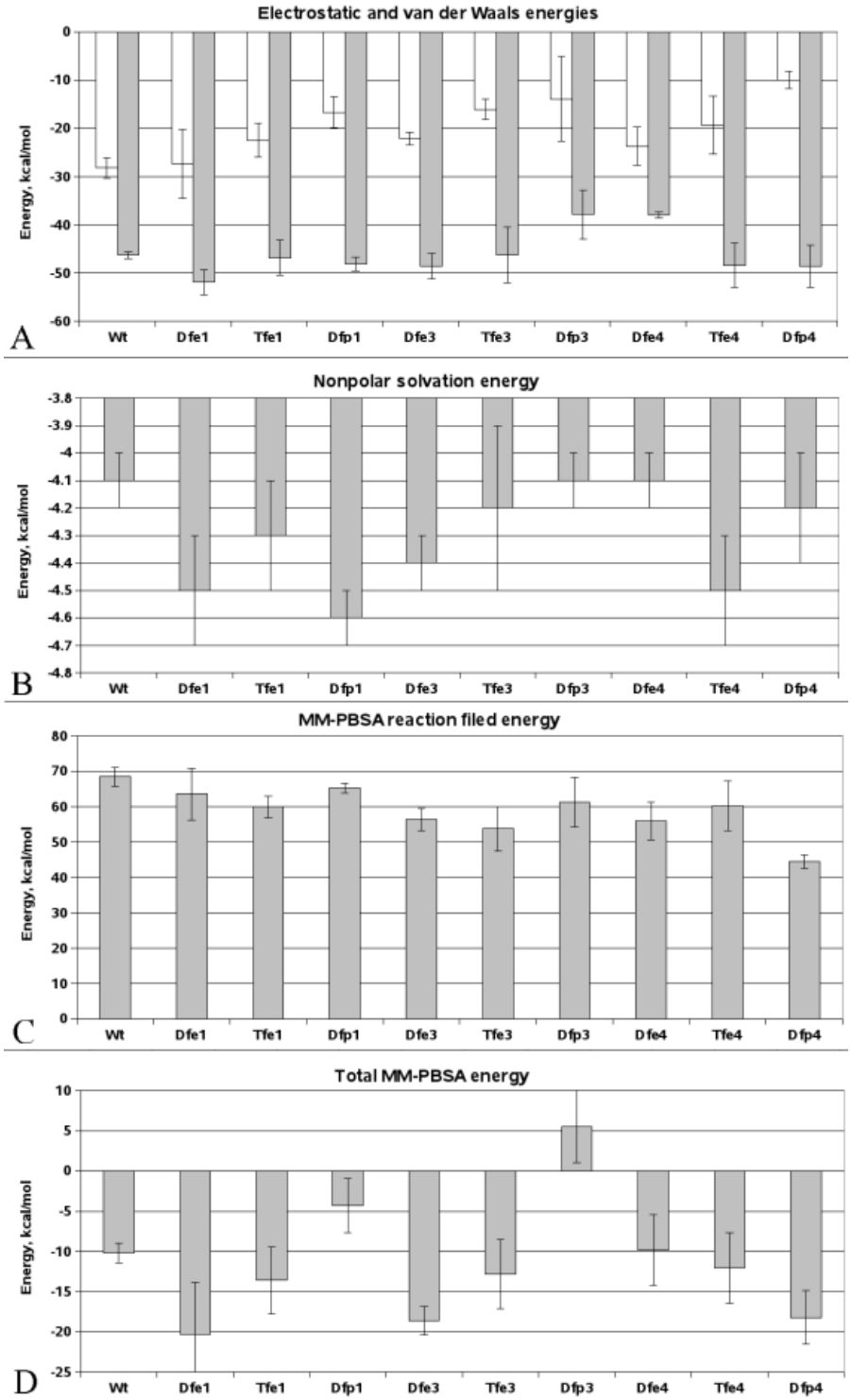

Figure 3.3.2. MM-PBSA binding energy components. A) Electrostatic and van der Waals components B) Non-polar solvation energy C) MM-PBSA reaction field energy D) Total MM-PBSA energy.

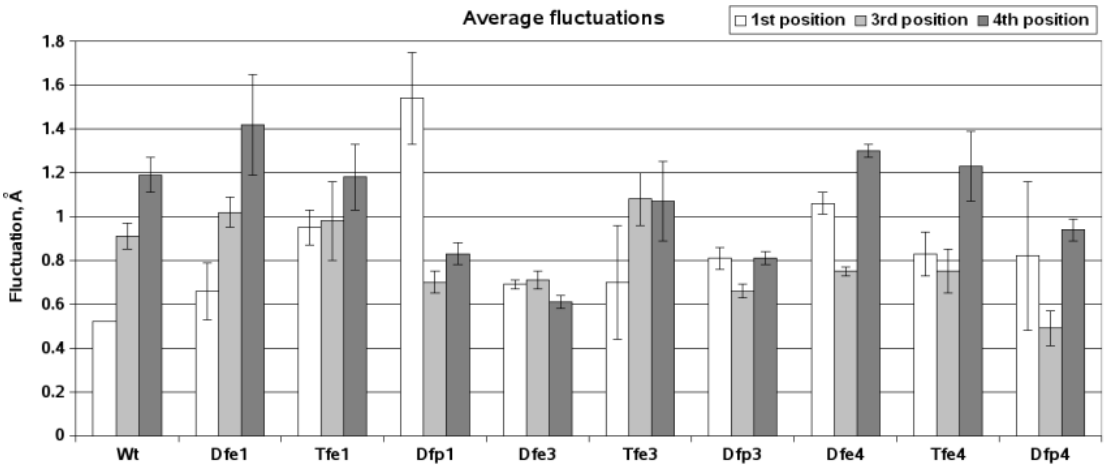

Figure 3.3.3. Mobility of the peptide residues at mutated positions.

Position 1. The introduction of DfeGly and TfeGly stabilizes the interaction at position 1 (Figure 2D), disallowing the flip of the first residue in the peptide. In the fluorinated peptides van der Waals energies as well as non-polar energies of solvation are lower, which is due to an increase of solvent accessible surface area (Figure 2A, 2B). Because of the bigger hydrophobic moiety exposed to the solvent in case of DfpGly in the first position, the residue demostrates an increased mobility (Figure 3) and flips.

Position 3. The introduction of DfeGly and TfeGly makes hydrophobic contacts for the position 3 more favorable. Additional stabilization results from contacts made by the Ala1 side chain. In general, these mutations make the peptide less mobile, though twists of Ala1 are allowed without any disruption of the productive conformation. DfpGly mutant disrupts the productive conformation rapidly, Dfp3 is the only peptide with not-favorable binding energy in the productive conformation (Figure 2).

Position 4. The bigger becomes the moiety of a fluorinated substitute (in the row DfeGly, TfeGly, DfpGly), the more stabilized are hydrophobic contacts and a peptide binding in general (Figure 2A, 2D). Residues in this position are the most mobile compared to the ones in positions 2 and 3 (Figure 3).

Fluctuations analysis. The linear regression analysis shows that total energy is dependent on quadratic fluctuation of the residue in the 1st position (adjusted R= 17%, 66%, 82% for all, all peptides without Dfp3 and all peptides without Wt and Dfp3, respectively), for other dependencies p-values are always higher than 0.05, meaning that impact of residues in the position 1 could be energetically decisive for binding.

3.3.4 Conclusions

Introduction of a sterically more demanding hydrophobic moiety into the ligand leads either to destabilization of interactions between a receptor and a ligand if a moiety is exposed to solvent (Dfp1, Dfp3) or to a hydrophobic interaction energy gain by formation of new contacts (Dfe1, Tfe1, Dfe3, Tfe3, Dfp4).

All the analyzed substitutions are characterized by decrease of electrostatic energy. 6 out of 9 studied fluorinated peptides improve the binding energy in productive binding mode (Dfe1, Tfe1, Dfe3, Tfe3, Tfe4, Dfp4), Dfe4 do not significantly affect the binding, and Dfp1 and Dfp3 make the binding weaker compared to Wt. Comparison of the obtained energy values shows that the position 1 in substrate peptide affects binding energetics the most.

Our results agree and help to explain the experimental data obtained by use of analytical HPLC with a fluorescence marker.

CHAPTER 4

Selection of a buried salt bridge by phage display

by Toni Vagt , Christian Jäckel, Sergey Samsonov , M. Teresa Pisabarro, and Beate Koksch

Bioorganic & Medicinal Chemistry Letter, 2009 Mer 21 [Epub ahead of print]

my contribution: computational part

4.1 Abstract

The α-helical coiled-coil is a valuable folding motif for protein design and engineering. By means of phage display technology, we selected a capable binding partner for one strand of a coiled-coil bearing a charged amino acid in a central hydrophobic core position. This procedure resulted in a novel coiled-coil pair featuring an opposed Glu-Lys pair arranged staggered within the hydrophobic core of a coiled-coil structure. Structural investigation of the selected coiled-coil dimer by CD spectroscopy and MD simulations suggest that a buried salt bridge within the hydrophobic core enables the specific dimerization of two peptides.

4.2 Introduction

The α-helical coiled-coil is one of the most widespread structural motifs in nature and is found in motor proteins, transcription factors, viral fusion proteins, and many more[271]. Due to extensive investigations within the last decades, the coiled-coil is also one of the best understood protein folding motifs[272]. The structural simplicity and distinctive coherence between sequence and structure allows the *de novo* design of coiled-coils with special features and makes it a useful tool in protein engineering[232,273]. In such cases, one of its most valuable features is the formation of highly specific dimerization domains, which introduces the possibility of targeting viral fusion proteins or transcription factors in therapeutic applications. Furthermore, the coiled-coil has proven to be an excellent model system for the systematic investigation of protein folding [178,221].

Coiled coils typically consist of two to five right-handed α-helices that wrap around each other to form a left-handed superhelix. The primary structure of each helix comprises the so-called heptad repeat, a periodicity of seven residues commonly denoted $(a\text{-}b\text{-}c\text{-}d\text{-}e\text{-}f\text{-}g)_n$. Typically, nonpolar residues occupy positions **a** and **d**, forming a hydrophobic surface which initiates oligomerization under aqueous conditions. Charged amino acids in positions **e** and **g** form a second interaction domain which favors coiled-coil formation by interhelical ionic interactions, while positions **c**, **b** and **f** are solvent-exposed and thus often populated by polar residues.

Despite the important role of the hydrophobic character of positions **a** and **d** for the formation and stability of the coiled-coil folding motif, an analysis of protein databases revealed that in natural occurring proteins up to 20 % of these positions are populated by polar and charged residues[274]. In spite of their destabilizing effect, these amino acids play a decisive role in the oligomerization state and orientation of the monomers within the coiled-coil [275,276]. As the construction of specific interacting peptide domains is one of the main goals of coiled-coil design, modification of the hydrophobic core by charged amino acids represents a promising strategy for the design of coiled-coil heteromers [277]. Buried salt bridges have already been used successfully in the design of highly specific interacting coiled-coil dimers [278-280]. However, these approaches are commonly based on lysine or arginine analogues with shortened side chains in order to minimize steric mismatches within the hydrophobic core. However, folding motifs which are applicable for synthesis *in vivo*, desirable for many applications, typically require a composition of naturally occurring amino acids. Otherwise, charged interactions within the hydrophobic core provide a promising design principle to direct oligomerization specificity in a manner which keeps the **e**- and **g**- positions free for the introduction of further specifications. To combine both features (control of heteromerization specificity by charged interactions within the hydrophobic core and accessibility by *in vivo* synthesis) within one system, we were interested in the determination of peptides built up of canonical amino acids which specifically interact with coiled-coil strands bearing charged amino acids in hydrophobic core positions.

In contrast to the often used strategy of rational peptide design, we used saturation mutagenesis to construct an extensive library of potential coiled-coil pairs which was screened using phage display technology[281]. This technique links the phenotype of a peptide which is displayed on the surface of a bacteriophage with the genotype encoding this peptide within the phage particle. While saturation mutagenesis enables generation of a phage displayed peptide library, physical linkage between phenotype and genotype enables the easy identification of individual peptides which are selected by binding preference to a given target. Phage display has already proven to be a powerful tool for the screening of protein-DNA, protein-protein, and protein-peptide interactions [282-285]. Previously, this technique was successfully applied in the determination of specific coiled-coil pairing, demonstrating its suitability for peptide design [286,287].

4.3 Methodology

Library construction and phage display. The VPE-library was expressed on the surface of bacteriaphage M13 as pIII fusion using ***E. coli*** ER2738 (New England Biolabs #E4104S), **V**CSM13

helper phage (Stratagene #200251), **p**Comb3H [288] phagemid vector (GenBank database accession number: AF268280, Barbas laboratory, TSRI). Library construction was performed by annealing of two complementary oligonucleotides that have the four library codons randomized by applying the NNK-strategy [289,290].

The following randomized oligonucleotides, which possess phosphate at the 5'-end, were purchased from *biomers.net GmbH* (Ulm, Germany) and applied for library construction (codons are in reading frame):

sense-strand: 5'-CG GCC GAG GTT AGC GCG CTG GAA AAG GAG GTG GCC AGT TTA GAG AAA GAG NNK AGT GCC NNK NNK AAG AAA NNK GCG AGC CTG AAA AAG GAG GTA AGT GCG TTA GAA GGC CAG GC-3'

anti-sense-strand: 5'-TG GCC TTC TAA CGC ACT TAC CTC CTT TTT CAG GCT CGC MNN TTT CTT MNN MNN GGC ACT MNN CTC TTT CTC TAA ACT GGC CAC CTC CTT TTC CAG CGC GCT AAC CTC GGC CGC CT-3'

N stands for A, G, C, and T; K stands for G and T; M stands for A and C

Vector digest. 10 µl pComb3H phagemid vector (1 µg/µL) were incubated with 60U SfiI (recombinant, New England Biolabs #R0123S) and BSA (100 µg/mL) in NE-buffer 2 (total volume 100 µL) for 4.5 h at 50 °C. The reaction was worked up by agarose gelelectrophoresis, and both the phagemid DNA as well as the small DNA fragment (insert) were isolated from the gel, DNA bands were cut out, and DNA was isolated using a Qiaquick gel extraction kit (Qiagen). After elution from the column with 50 µL deionized water, the DNA concentration was determined and the DNA was stored at -20°C.

Annealing of the library encoding oligonucleotides. 3 × 675 ng of both of the randomized oligonucleotides were incubated at 95°C for 10 min in a volume of 200 µL annealing buffer (10 mM TrisHCl, 2 mM $MgCl_2$, 50 mM NaCl, pH 7.5). The annealing samples were cooled down slowly to 15°C within 80 min. After ethanol precipitation and dissolving in 20 µL deionized water, all DNA samples were purified by agarose gelelectrophoresis. DNA bands were cut out and DNA was isolated using a Qiaquick gel extraction kit (Qiagen). After elution with 20 µL deionized water all DNA samples were pooled, and concentration was determined.

Ligation and transfection of ER2738. 1 µg annealed library DNA was incubated with 1.4 µg digested (o)-phagmid DNA and 4000U T4 DNA Ligase (New England Biolabs #M0202S) over night at 16 °C in a total volume of 200 µL T4 DNA Ligase Reaction Buffer. Ligation reactions without insert DNA

served as negative controls and ligations of vector DNA with the non-randomized DNA insert, which was cut out of the vector before, served as positive controls. Ligated phagemid DNA was purified by ethanol precipitation and dissolved in 15 µL deionized water. For transfection of electrocompetent ER 2738 all 15 µL of ligated phagemid were mixed with 300 µL freshly thawed electrocompetent cells and transferred into a 2 mm electro cuvette. After electroporation with 2.5 kV, the cuvette was immediately flushed with 1 mL prewarmed SOC medium. The cuvette was rinsed with additional 4 mL SOC media and all five fractions were pooled. Cell cultures were shaken for 1h at 37°C and 200 rpm. After addition of 10 mL prewarmed SB medium, 5, and 0.5 µL of each culture, respectively, were plated on carbenicillin agar plates (10 µg/ml) and allowed to grow over night at 37°C. Colonies were counted and a library size of 2.3×10^6 was calculated.

Production of library phage. 3 µL carbenicillin (100 mg/mL) were added to the 15 mL cell culture (see above). After 1h incubation at 37°C/ 200rpm agitation, additional 4.5µl carbenicillin (100 mg/ml) were added, and the incubation proceeded for 1h at 37°C / 200 rpm. Cell cultures were transferred to 183 mL prewarmed SB media, containing 92.5 µL carbenicillin (100 mg/mL) in centrifuge bottles, and 2 mL VCSM13 helper phage (Stratagene #200251) were added. After incubation for 2 h at 37 °C/ 200 rpm, 280 µL kanamycin (50 mg/mL) were added and phage were produced over night at 37°C/ 200 rpm agitation. Overnight cultures were centrifuged at 3000 g/ 4°C for 30 min. The supernatant was transferred to precooled centrifuge bottles, containing 0.2 volumes 20% PEG 8000/ 2.5M sodium chloride solution, and the phage precipitation proceeded for 30 min on ice. The phage were centrifuged for 30 min at 15000g/ 4°C. The supernatant was discarded, and bottles were drained by inverting on a paper towel for 10 min. Phage pellets were resuspended with 2 ml PBS and passed through a 0.22µm filter. For storage at 4°C, sodium azide was added to a final concentration of 0.02% (w/v).

Library panning. For the immobilization of target peptides, 30µL of streptavidin-coated magnetic beads (M-280, Dynal Biotech) were incubated with 500 µL 10 µM biotinylated peptide in PBS for 45 min at RT. Particles were washed twice with 500 µL 0.1% Tween 20 in PBS. 500 µL 5% non-fat dried milk in PBS was added and the sample was incubated for 45 min at RT. The milk-PBS suspension was removed and the target-phage binding on magnetic particles was performed with 500µL phage solution for 1.5 h at RT. Particles were washed 4 x with 500 µL Tween20 in PBS (PBS buffer contained 0.1 % Tween 20 in round 1; 1 % Tween 20 in rounds 2-5; in round 5 two washing steps with 1 M GndHCl in PBS were added) and once with 500µL TBS (4min incubation). For Panning with VPK-E_{19} only PBS containing 0.1 % Tween20 (round 1-4) and 0.2 % Tween20 (round 5) was used.

Reinfection. *E. coli* for phage infection were prepared by inocculation of 7.5 µL electrocompetent ER2738 in 5 mL prewarmed SB media and growing for 2 h at 37°C/ 200rpm. The elution of bound phage from magnetic particles proceeded with 25 µL freshly prepared trypsin solution (10 mg/mL in TBS) for 30 min at RT and the reaction was quenched with 75 µL SB media. For reinfection, 100 µL phage solution was transferred to 5 mL *E. coli* culture and shaken for 30 min at 37°C/ 200rpm agitation. After removing 10 µl cell culture for output-titering (see below), 5 ml prewarmed SB media and 2.5 µL carbenicillin (100mg/mL) were added and samples were shaken for 1h at 37°C/ 200rpm. Cell cultures were transferred to 90 mL prewarmed SB media, containing 46 µL carbenicillin (100 mg/ mL) in centrifuge bottles and 1 ml helper phage were added. After incubation for 1.5 h at 37°C/ 200 rpm, 140 µL kanamycin (50mg/mL) were added and phage were produced over night at 37°C/ 200 rpm agitation.

Determination of phage titers. Input-titering was performed by infecting 50 µL *E. coli* with 1 µL of a $1x10^{-6}$ dilution of the phage preparation and incubation for 15 min at RT. All 50 µL were plated on carbenicillin agar plates and bacteria were allowed to grow over night at 37°C. Output-titering was performed by plating 50 µL of a 10^{-2} and af 10^{-3} dilution of the 5 mL reinfection cell culture (see above) on carbenicillin agar plates. After growing over night at 37°C bacteria colonies were count and output/input ratios for all samples were calculated.

Peptide synthesis, purification and characterization. Fmoc-Glu(OtBu)- and Fmoc-Lys(Boc)-NovaSyn-TGA resins (0.16 mmol/g and 0.21 mmol/g, respectively) were purchased from Novabiochem. Fmoc-L-amino acids, 2-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethyluronium tetrafluoroborate (TBTU), 1-hydroxybenzotriazole (HOBt) were purchased from Fa. Gerhardt (Wolfhagen, Germany) and 1-hydroxy-7-azabenzotriazole (HOAt) from Iris Biotech. Dimethylformamide (p.a.), triisopropylsilane (TIS 99%), N,N-diisopropylethylamine (DIEA 98+%), N,N-diisopropylcarbodiimide (DIC, 99%), trifluoroacetic acid (TFA, 99%), sodium perchlorate (p.a.), piperidine (99% extra pure) and biotine were purchased from Acros. 1,8-diazabicyclo[5.4.0]undec-7-en and trifluoroacetic acid (Uvasol) were obtained from Merck. Peptides were synthesized on a SyroXP-1 peptide synthesizer (MultiSynTech GmbH, Witten, Germany) using standard Fmoc/tBu chemistry and TCTU/HOBt as coupling reagents at a 0.05 mM scale. For standard couplings a fourfold excess of amino acids and coupling reagents as well as an eight fold excess of DIEA relative to resin loading was used. All couplings were performed as double couplings (30 min.). The coupling mixture contained 0.23 M $NaClO_4$ to prevent on-resin aggregation. Fmoc-deprotection was performed 4 × 5 min using 2

% DBU and 2 %piperidine in DMF. After synthesis peptides were cleaved from the resin with TFA/TIS/H_2O (95/2.5/2.5). Purification was carried out by RP-HPLC (Phenomenex® Luna C_8, 10 µm, 250 nm × 21.2 mm). Purity was determined by analytical HPLC (Phenomenex® Luna C_8, 5 µm, 250 nm × 4.6 mm).

To identify the products high resolution mass spectra were recorded on the Agilent 6210 ESI-TOF mass spectrometer (Agilent Technologies, Santa Clara, CA, USA.). The samples were dissolved in acetonitrile/water (1/1) containing 0.1 % TFA and injected directly into the spray chamber using a syringe pump with flow rates of 10 to 50 μL/min. The spray voltage was 4.000V and the drying gas (N2) flow rate was set to 1 psi (1 bar).

Circular Dichroism. CD-spectra were recorded in 100 mM phosphate buffer pH 7.4 at an overall peptide concentration of 20 µM on a Jasco J-715 spectropolarimeter at 20°C (Jasco PTC-348 WI peltier thermostat). The ellipticity was normalized to concentration (c/mol×l^{-1}), number of residues (n) and path length (l/cm) using the following equation:

$$[\Theta] = \frac{\Theta_{obs}}{1 \cdot \cdots \cdot l \cdot c \cdot n} \tag{4.1}$$

where Θ_{obs} is the measured ellipticity in mdeg and [Θ] the normalized ellipticity in 10^3 deg × cm^2 × dmol^{-1} × residue^{-1}. Each sample was prepared three times and spectra were averaged.

Determination of Peptide Concentration. Concentrations were estimated by UV spectroscopy on a Cary 50 UV/Vis spectrometer (Varian) using the absorption of o-aminobenzoic acid attached to each N-terminus. A calibration curve (Figure 4.1) was recorded using different concentrations of H_2N-Abz-Gly-COOH·HCl (Bachem) in the buffer used for CD spectroscopy containing 6M guanidinium hydrochloride (Fluka). Disposable Plastibrand® PMMA cuvettes (Brand GmbH, Germany) with path lengths of 1 cm were used.

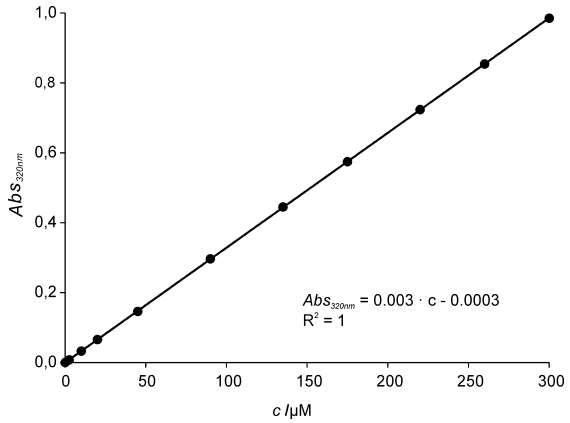

Figure 4.1. Calibration curve for the determination of peptide concentrations recorded at 20°C (100 mM phosphate buffer, 6 M GdnHCl, pH 7.4).

MD simulations and MM-PBSA free energy calculations. The crystal structure of the Sir4p C-terminal coiled-coil at 2.5 Å resolution (PDB ID: 1pl5) was used as template for modeling our parallel coiled-coil systems. To obtain the parent peptide model system (VPE-VPK) the length of the helices of the Sir4p coiled-coil was reduced to 34 aa, and the necessary side chain substitutions were carried out with the MOE program[203]. The structures were solvated in a TIP3P water octahedral box, and periodic boundary conditions under constant temperature (300K) and constant pressure (10^6 Pa NTP) were applied. MD productive runs of simulations of 5 ns were performed with AMBER 8.0 [8] using the ff03 force field. Energetic post-processing of the trajectories was done in a continuous solvent model as implemented in the AMBER 8.0 MM-PBSA module. The snapshots for the calculations were chosen as described by Lafont and coworkers[126]. Entropies were calculated using normal mode analysis. Significant comparison of the free energies of interaction between two coiled-coils is not possible because of the intrinsic flexibility of the helices termini. To avoid this additional source of noise in the MM-PBSA calculations only the central parts of the helices were analyzed (residues 10-25).

Phage Display. Sequencing of randomly picked clones after the last round of panning against the VPK wild type (Table 4.1) and VPK-E_{19} (Table 4.2) resulted in the following amino acid pattern in the randomized positions of VPE:

Table 4.1. Peptides used in this study synthesized by SPPS.

Peptide	Sequence
Bio-VPK	Biotin-GSGKVSALKEKVASLKEKVSALKEEVASLEEKVSALK-OH
Bio-VPK-E_{19}	Biotin-GSGKVSALKEKVASLKEKVSAEKEEVASLEEKVSALK-OH
VPK	Abz-KVSALKEKVASLKEKVSALKEEVASLEEKVSALK-OH
VPK-E_{19}	Abz-KVSALKEKVASLKEKVSAEKEEVASLEEKVSALK-OH
VPE	Abz-EVSALEKEVASLEKEVSALEKKVASLKKEVSALE-OH
VPE-LLLL	Abz-EVSALEKEVASLEKELSALLKKLASLKKEVSALE-OH
VPE-LLLK	Abz-EVSALEKEVASLEKELSALLKKKASLKKEVSALE-OH

Abz: o-Aminobenzoic acid.

Table 4.2. Identification of the synthesized peptides by ESI-TOF mass spectrometry.

Peptide	Calc.$[M+4H]^{4+}$	Obs $[M+4H]^{4+}$
Bio-VPK-E_{19}	1028.8112	1028.8219
Bio-VPK	1024.8216	1024.8302
VPK-E_{19}	951.7824	951.7886
VPE	948.2665	948.7715
VPE-LLLL	951.2847	951.2795
VPE-LLLK	955.0375	955.0466

Oligomerization state of the VPK-E_{19}/VPE-LLLK heteromer. In order to determine the oligomerization state of the VPK-E_{19}/VPE-LLLK heteromers CD-spectra of mixtures of both peptides at different molar ratios were recorded. Plotting of the mean residue ellipticity at 222nm versus the mole fraction of VPK-E_{19} showed a minimum at a nearly 1:1 molar ratio of VPK-E_{19} and VPE-LLLK suggesting the formation of heterodimers (Figure 4.2).

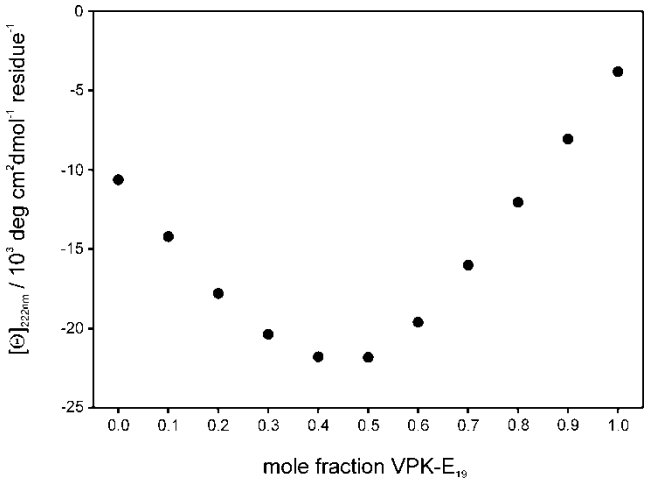

Figure 4.2. Mean residue ellipticity at 222 nm versus the mole fraction of VPK-E_{19}. Spectra were recorded at 20 °C in 100 mM phosphate pH 7.4. The total peptide concentration was held constant at 20 µM.

4.4 Results and Discussion

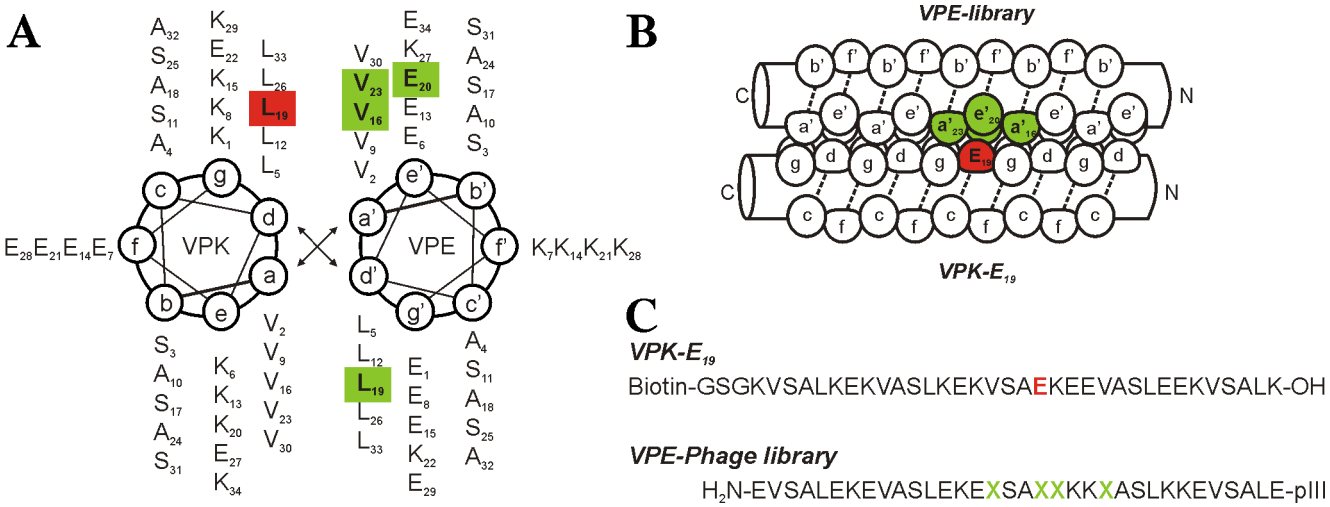

Figure 4.3. A) Helical wheel presentation of the parallel VPK-VPE heterodimer. B) Schematic side view of VPK-E_{19} and the VPE-library. C) Glu_{19} in VPK-E_{19} is highlighted in red, while the four randomized amino acid positions in VPE which directly interact with Glu_{19} are highlighted in green. Amino acid sequences of VPK-E_{19} and the VPE-phage library.

The starting point for our studies was the coiled-coil pair VPE/VPK, reported by our group to form parallel heterodimers [291]. VPK features lysine in positions **e** and **g** while **e'** and **g'** in VPE are occupied by negatively charged glutamic acid to further the formation of heterodimers by ionic

interactions under physiological conditions (Figure 4.3 A). A parallel arrangement as well as a preference for dimer formation is dictated by valine in position **a** [263]. In order to place a negatively charged side chain within the hydrophobic core, Leu_{19} in VPK was replaced by glutamic acid (Figure 4.3 B). As expected, this substitution results in significant destabilization of the VPK/VPE coiled-coil. Neither homodimerization of VPK-E_{19} nor the formation of VPK-E_{19}/VPE dimers could be observed by CD spectroscopy (Figure 4.4). Subsequently, the four amino acid positions within VPE which directly interact with Glu_{19} in VPK-E_{19} were fully randomized and the resulting VPE-library was fused to the minor coat protein pIII on the surface of the filamentous bacteriophage M13 (Figure 4.3 B). The DNA fragment that encodes for the VPE-peptide, including the four randomized positions a_{16}, d_{19}, e_{20} and a_{23}, was inserted into the phagemid vector pComb3H to the 5'-end at the gene that encodes for the C-terminal part of the truncated minor coat protein pIII [288]. After successful cloning of the randomized DNA into M13, amplified phage that present the peptide library were used in the selection for binding partners of the Glu_{19}-substituted VPK-variant. VPK-E_{19} used for the selection procedure carries an N-terminal biotin label for immobilization on streptavidin-coated magnetic beads. Coiled coil pairing selectivity was then used to determine the best binding partner in the library, which is able to compensate for the destabilizing effect of the charge within the hydrophobic domain of VPK-E_{19}.

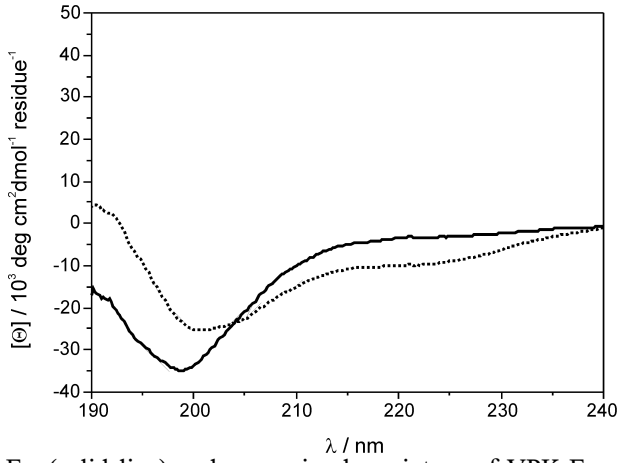

Figure 4.4. CD spectra of VPK-E_{19} (solid line) and an equimolar mixture of VPK-E_{19} and wild type VPE (dotted line). Spectra were recorded at 20 °C in 100 mM phosphate buffer pH 7.4 at an overall peptide concentration of 20 µM.

In a control experiment, 5 rounds of panning against wild type VPK resulted in a variety of sequences which can be summarized as the motifs VPK-$Z_{16}L_{19}X_{20}L_{23}$ and $L_{16}L_{19}X_{20}Z_{23}$, where X indicates predominantly hydrophobic residues and Z marks aromatic residues (Table 4.3, 4.4). Despite the surprising variety of selected VPE-variants, the hydrophobic character of the amino acids which were found in the positions of the hydrophobic core clearly correlate with the design principles of the

α-helical coiled-coil.

Table 4.3**.** Results of the sequencing of 10 randomly selected clones after 5 rounds of panning against VPK wild type.

Frequency	Position a‘$_{16}$	Position d‘$_{19}$	Position e‘$_{20}$	Position a‘$_{23}$
2×	Leu	Leu	Leu	Tyr
1×	Leu	Leu	Leu	Phe
1×	Leu	Leu	Tyr	Tyr
1×	Leu	Leu	Gln	Tyr
1×	Tyr	Leu	Lys	Leu
2×	Tyr	Leu	Leu	Leu
1×	Phe	Leu	Leu	Leu
1×	Leu	Leu	Leu	Leu

Table 4.4. Results of the sequencing of 9 randomly selected clones after 5 rounds of panning against VPK-E_{19}.

Frequency	Position a‘$_{16}$	Position d‘$_{19}$	Position e‘$_{20}$	Position a‘$_{23}$
2×	Leu	Leu	Leu	Tyr
1×	Leu	Leu	Leu	Phe
1×	Leu	Leu	Tyr	Tyr
1×	Leu	Leu	Gln	Tyr
1×	Tyr	Leu	Lys	Leu
2×	Tyr	Leu	Leu	Leu
1×	Phe	Leu	Leu	Leu
1×	Leu	Leu	Leu	Leu

In contrast, panning against VPK-E_{19} resulted in greater sequence convergence. Sequencing of nine clones yielded only two different phenotypes: 8 of 9 clones matched the sequence of VPE-LLLK, in which leucine occupies positions $\mathbf{a_{16}}$, $\mathbf{d_{19}}$ and $\mathbf{e_{20}}$ and lysine position $\mathbf{a_{23}}$. The remaining clone differs only in position e_{20}, where methionine was found instead of leucine. Overall, leucine represents the most common amino acid within the hydrophobic core of naturally occurring coiled-coil peptides and its selection in positions $\mathbf{a}_{16}$ and $\mathbf{d_{19}}$ is unsurprising. The increase in hydrophobic surface area by the substitution of Glu_{20} with leucine or methionine obviously stabilizes the coiled-coil structure even further.

As expected, CD analysis of VPK-E_{19} and VPE-LLLK in isolation yields CD spectra characteristic for predominantly random coil structures (Figure 4.5 A). In both peptides, the charged residue within the hydrophobic core, in addition to the repulsive interactions between the **e**/**g’** and **g**/**e’**

positions which are a feature of the original design strongly discourages the formation of homodimers. In addition, a 1:1 mixture of VPK-E_{19} and wild type VPE does not result in coiled-coil formation (Figure 4.4). In contrast, an equimolar mixture of VPK-E_{19} and VPE-LLLK shows a strong α-helical CD signal. Plotting the mean residue ellipticity at 222 nm versus the mole fraction of VPE-LLLK shows a minimum at a 1:1 molar ratio of VPK-E_{19} and VPE-LLLK, suggesting the presence of heterodimers (Figure 4.2). An additional experiment confirmed the importance of Lys_{23} for the formation of the coiled-coil heteromer. The control peptide VPE-LLLL differs from VPE-LLLK only in position $\mathbf{a_{23}}$, in which lysine was replaced by leucine. While this single substitution enables homodimerization of VPE-LLLL, no formation of heteromers in combination with VPK-E_{19} could be observed. Instead, an equimolar mixture of the two peptides shows a CD spectrum which exactly matches the sum of the CD spectra recorded for each isolated peptide (Figure 4.5 B). This indicates that no interaction between VPK-E_{19} and VPE-LLLL takes place. Obviously, Lys_{23} is the key element for the formation of the heteromeric coiled-coil dimer VPK-E_{19}/VPE-LLLK.

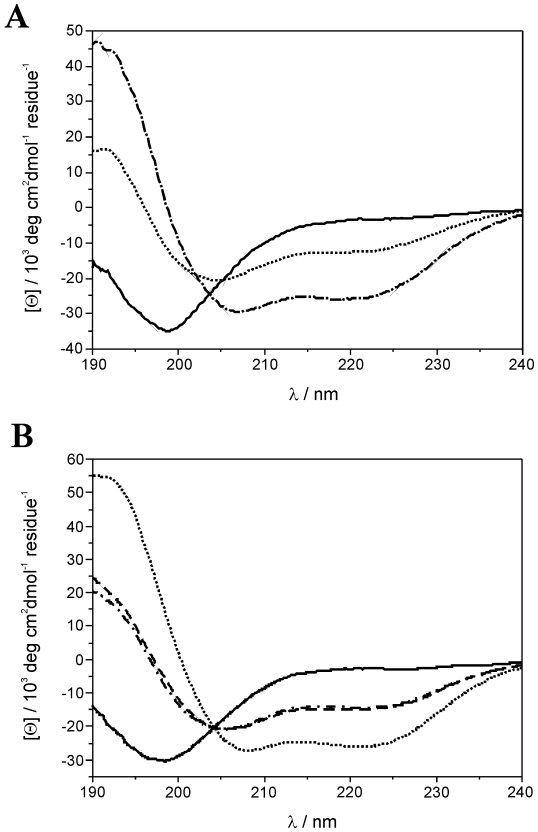

Figure 4.5. A) CD spectra of VPK-E_{19} (solid line), VPE-LLLK (dotted line), and an equimolar mixture of VPK-E_{19} and VPE-LLLK (dashed-dotted line). B) CD-spectra of VPK-E_{19} (solid line) and an equimolar mixture of VPK-E_{19} and VPE-LLLL (dashed line). The sum of the single spectra of VPK-E_{19} and VPE-LLLL is presented as dashed-dotted line and the spectrum of VPE-LLLL as dotted line. Spectra were recorded at 20 °C in 100 mM phosphate buffer pH 7.4 at an overall peptide concentration of 20 µM.

Our results suggest that the formation of an interhelical salt bridge between Glu_{19} (VPK-E_{19}) and Lys_{23} (VPE-LLLK) represents one of the driving forces for dimerization of VPK-E_{19}/VPE-LLLK and that it determines the specificity of this interaction (Figure 4.6). Computer modelling supports this assumption. MD simulations (see Supporting Information for details) demonstrate that the heterodimer is stable in solution. The distances between C_β atoms of the residues in **a**- and **d**-positions in each helix do not fluctuate significantly in the simulation and the backbone dihedral angles remain close to ideal α-helical values. The distance between the carbonyl oxygen of Glu_{19} and amide nitrogen of Lys_{23} was calculated to be less than between 3Å (95.9% of the simulation time) and 4Å (99.7% of the simulation time), which strongly suggests the existence of a salt bridge between these functional groups. Furthermore, decomposition of the MM-GBSA energy values show that Lys_{23} contributes the most significant value of all amino acids to the overall energy of the coiled-coil interaction (-5 kcal/mol, which is about 20% of the overall free energy of unfolding).

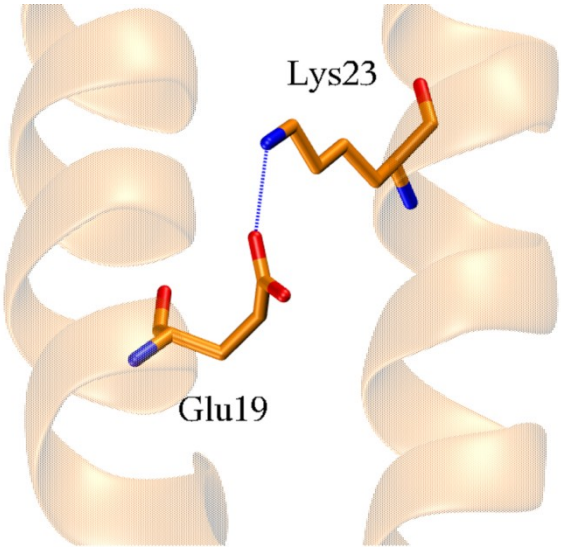

Figure 4.6. Structure of the proposed salt bridge between Glu_{19} of VPK-E19 and Lys_{23} of VPE-LLLK.

4.5 Conclusions

Buried salt bridges within the hydrophobic core of coiled-coil peptides created by rational design to construct specific interacting coiled-coil pairs have been shown before. Nevertheless, such structures typically are limited to directly opposed a and a' or d and d' positions. Due to the small distance between these positions in the parallel coiled-coil, aspartic acid and positively charged non proteinogenic amino acids with shortened side chains are necessary for stable coiled-coil formation [278-280]. However, by using saturation mutagenesis and selection, we were able to find a coiled-coil pair characterized by a buried salt bridge between lysine and glutamic acid. Here, the charged amino

acids are not placed in positions directly opposed, but staggered within the hydrophobic core. This arrangement between i and i'+4 positions apparently enables the formation of a salt bridge between the ammonium function of the lysine side chain and the carboxylate function of the glutamic acid side chain without disruption of the coiled-coil structure. Instead, this interaction results in a highly specific dimerization domain made up of canonical amino acids which could potentially be used in protein design and material engineering [277]. The selection of VPE-LLLK not only delivers a new highly specific coiled-coil heterodimer but also demonstrates the potential of the phage display-based screening system to select specific interaction partners for uncommon amino acids within the hydrophobic core of coiled-coils.

REFERENCES:

1. Leninger. Principles of Biochemistry. 2004;pp1100.

2. Zubay G, Parson W, Vance D. Principles of Biochemistry. 1994;pp989.

3. Berg J, Tymoczko J, Stryer L. Biochemistry. 2002;pp1050.

4. van_Holde K, Curtis Johnson, Ho P. Principles of Physical Biochemistry. 2005;pp752.

5. Cramer CJ. Computational Chemistry. 2004;pp596.

6. Mount D. Bioinformatics: Sequence and Genome Analysis . 2001;pp564.

7. McCammon J, Harvey S. Dynamics of Proteins and Nucleic Acids. 2004;pp234.

8. Amber 8. Case D, Darden T, Cheatham T, Simmerling C, Wang J, Duke R, Luo R, Merz K, Wang B, Pearlman D, Crowley M, Brozell S, Tsui V, Gohlke H, Mongan J, Hornak V, Cui G, Beroza P, Schafmeister C, Caldwell J, Ross W, Kollman P. University of California, San Francisco 2004-2004.

9. Wikipedia, http://www.wikipedia.org/.

10. UniProt, http://www.uniprot.org/.

11. NCBI HomePage, http://www.ncbi.nlm.nih.gov/.

12. Jones S, Thornton J. Principles of protein-protein interactions. Proc Natl Acad Sci U S A 1996;93:13-20.

13. Nooren I, Thornton J. Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 2003;325:991-1018.

14. Nooren I, Thornton J. Diversity of protein-protein interactions. Embo J 2003;22:3486-3492.

15. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177-2198.

16. Argos P. An investigation of protein subunit and domain interfaces. Protein Eng 1988;2:101-113.

17. Tsai C, Lin S, Wolfson H, Nussinov R. Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. Crit Rev Biochem Mol Biol 1996;31:127-152.

18. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155-164.

19. Park J, Lappe M, Teichmann S. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol 2001;307:929-938.

20. Aloy P, Ceulemans H, Stark A, Russell R. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989-998.

21. Keskin O, Tsai C, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. Protein Sci 2004;13:1043-1055.

22. Stein A, Russell R, Aloy P. 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res 2005;33:D413-417.

23. Davis F, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics 2005;21:1901-1907.

24. Gong S, Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, Choi H, Cho Y, Han K, Lee S, Choi H, Lappe M, Holm L, Kim S, Oh D, Bhak J. PSIbase: a database of Protein Structural Interactome map (PSIMAP). Bioinformatics 2005;21:2541-2543.

25. Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser D, Oh D, Kim D, Bhak J. A protein domain interaction interface database: InterPare. BMC Bioinformatics 2005;6:207.

26. Ogmen U, Keskin O, Aytuna A, Nussinov R, Gursoy A. PRISM: protein interactions by structural matching. Nucleic Acids Res 2005;33:W331-W336.

27. Teyra J, Doms A, Schroeder M, Pisabarro M. SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. BMC Bioinformatics 2006;7:104.

28. Glaser F, Steinberg D, Vakser I, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 2001;43:89-102.

29. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003;325:377-387.

30. Dill K, Truskett T, Vlachy V, Lee B. Modeling water, the hydrophobic effect, and ion solvation. Annual Review of Biophysics and Biomolecular Structure 2005;34:173-199.

31. Ladbury J. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. Chem Biol 1996;3:973-980.

32. Raschke TM. Water structure and interactions with protein surfaces. Curr Opin Struct Biol 2006;16:152-159.

33. Dolenc J, Baron R, Missimer J, Steinmetz M, van Gunsteren W. Exploring the conserved water site and hydration of a coiled-coil trimerisation motif: a MD simulation study. Chem Bio Chem 2008;9:1749-1756.

34. England J, Lucent D, Pande V. A Role for Confined Water in Chaperonin Function. Journal of the American Chemical Society 2008;130:11838-11839.

35. Qvist J, Davidovic M, Hamelberg D, Halle B. From the Cover: A dry ligand-binding cavity in a solvated protein. Proceedings of the National Academy of Sciences 2008;105:6296-6301.

36. Palencia A, Cobos E, Mateo P, Martinez J, Luque I. Thermodynamic dissection of the binding energetics of proline-rich peptides to the Abl-SH3 domain: implications for rational ligand design. J Mol Biol 2004;336:527-537.

37. Petrone P, Garcia A. MHC-peptide binding is assisted by bound water molecules. J Mol Biol 2004;338:419-435.

38. Ogata K, Wodak S. Conserved water molecules in MHC class-I molecules and their putative structural and functional roles. Protein engineering 2002;15:697-705.

39. Rhodes MM, Reblova K, Sponer J, Walter NG. Trapped water molecules are essential to structural dynamics and function of a ribozyme. Proc Natl Acad Sci U S A 2006;103:13380-13385.

40. Loris R, Langhorst U, De Vos S, Decanniere K, Bouckaert J, Maes D, Transue T, Steyaert J. Conserved water molecules in a large family of microbial ribonucleases. Proteins 1999;36:117-134.

41. Wiggins P. High and low density water and resting, active and transformed cells. Cell biology international 1996;20:429-435.

42. Wiggins P. Life depends upon two kinds of water. PLoS ONE 2008;3:e1406.

43. Cooper A. Heat capacity effects in protein folding and ligand binding: a re-evaluation of the role of water in biomolecular thermodynamics. Biophys Chem 2005;115:89-97.

44. Hamelberg D, McCammon JA. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. J Am Chem Soc 2004;126:7683-7689.

45. Lu Y, Yang CY, Wang S. Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes. J Am Chem Soc 2006;128:11830-11839.

46. Barillari C, Taylor J, Viner R, Essex JW. Classification of water molecules in protein binding sites. J Am Chem Soc 2007;129:2577-2587.

47. Wong S, Amaro R, McCammon A. MM-PBSA Captures Key Role of Intercalating Water Molecules at a Protein-Protein Interface. Journal of Chemical Theory and Computation 2009;5:422-429.

48. Papoian G, Ulander J, Eastwood M, Luthey-Schulten Z, Wolynes P. Water in protein structure prediction. Proc Natl Acad Sci U S A 2004;101:3352-3357.

49. Park S, Saven J. Statistical and molecular dynamics studies of buried waters in globular proteins. Proteins 2005;60:450-463.

50. Nakasako M. Water-protein interactions from high-resolution protein crystallography. Philosophical Transactions of the Royal Society B: Biological Sciences 2004;359:1191-1206.

51. Helms V. Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. Chemphyschem 2007;8:23-33.

52. Damjanovic A, Garcia-Moreno B, Lattman E, Garcia A. Molecular dynamics study of water penetration in staphylococcal nuclease. Proteins 2005;60:433-449.

53. Chen X, Weber I, Harrison R. Hydration Water and Bulk Water in Proteins Have Distinct Properties in Radial Distributions Calculated from 105 Atomic Resolution Crystal Structures. J. Phys. Chem. B 2008;112:12073-12080.

54. Johnson M, Malardier-Jugroot C, Murarka R, Head-Gordon T. Hydration Water Dynamics Near Biological Interfaces. The Journal of Physical Chemistry B 2009;113:4082-4092.

55. Pizzitutti F, Marchi M, Sterpone F, Rossky P. How protein surfaces induce anomalous dynamics of hydration water. The journal of physical chemistry. B 2007;111:7584-7590.

56. Makarov VA, Andrews BK, Smith PE, Pettitt BM. Residence times of water molecules in the hydration sites of myoglobin. Biophys J 2000;79:2966-2974.

57. Priya H, Shah J, Asthagiri D, Paulaitis M. Distinguishing Thermodynamic and Kinetic Views of the Preferential Hydration of Protein Surfaces. Biophys. J. 2008;95:2219-2225.

58. Qiu W, Kao Y, Zhang L, Yang Y, Wang L, Stites W, Zhong D, Zewail A. Protein surface hydration mapped by site-specific mutations. Proceedings of the National Academy of Sciences 2006;103:13979-13984.

59. Furse K, Corcelli S. The Dynamics of Water at DNA Interfaces: Computational Studies of Hoechst 33258 Bound to DNA. J. Am. Chem. Soc. 2008;130:13103-13109.

60. Zhang L, Wang L, Kao Y, Qiu W, Yang Y, Okobiah O, Zhong D. From the Cover: Mapping hydration dynamics around a protein surface. Proc Natl Acad Sci U S A 2007;104:18461-18466.

61. van Dijk A, Bonvin A. Solvated docking: introducing water into the modelling of biomolecular complexes. Bioinformatics 2006;22:2340-2347.

62. Bui H, Schiewe A, Haworth I. WATGEN: An algorithm for modeling water networks at protein-protein interfaces. Journal of Computational Chemistry 2007;28:2241-2251.

63. Jiang L, Kuhlman B, Kortemme T, Baker D. A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. Proteins 2005;58:893-904.

64. Teyra J, Pisabarro M. Characterization of interfacial solvent in protein complexes and contribution of wet spots to the interface description. Proteins: Structure, Function, and Bioinformatics 2007;67:1087-1095.

65. Jorgensen WL. Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. J. Am. Chem. Soc. 1981;103:335-340.

66. Jorgensen W, Chandrasekhar J, Madura J, Impey R, Klein M. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics 1983;79:926-935.

67. Mahoney M, Jorgensen W. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. The Journal of Chemical Physics 2000;112:8910-8922.

68. Berendsen H, Postma J, van Gunsteren W, Hermans J. Intermolecular Forces. 1981;331.

69. Berendsen H, Grigera J, Straatsma T. The missing term in effective pair potentials. The Journal of Physical Chemistry 1987;91:6269-6271.

70. Bernal J, Fowler R. A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions. The Journal of Chemical Physics 1933;1:515-548.

71. Horn H, Swope W, Pitera J, Madura J, Dick T, Hura G, Gordon T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. The Journal of Chemical Physics 2004;120:9665-9678.

72. Abascal J, Sanz E, Fernandez G, Vega C. A potential model for the study of ices and amorphous water: TIP4P/Ice. The Journal of Chemical Physics 2005;122:234511.

73. Abascal J, Vega C. A general purpose model for the condensed phases of water: TIP4P/2005. The Journal of Chemical Physics 2005;123:234505-234505.

74. Stillinger F, Rahman A. Improved simulation of liquid water by molecular dynamics. The Journal of Chemical Physics 1974;60:1545-1557.

75. Rick S. A reoptimization of the five-site water potential (TIP5P) for use with Ewald sums. The Journal of Chemical Physics 2004;120:6085-6093.

76. Nada H, van der Eerden J. An intermolecular potential model for the simulation of ice and water near the melting point: A six-site model of H2O. The Journal of Chemical Physics 2003;118:7401-7413.

77. Silverstein K, Haymet A, Dill K. A Simple Model of Water and the Hydrophobic Effect. J. Am. Chem. Soc. 1998;120:3166-3175.

78. Izvekov S, Voth G. Multiscale coarse graining of liquid-state systems. The Journal of Chemical Physics 2005;123:134105.

79. Lippow SM, Tidor B. Progress in computational protein design. Curr Opin Biotechnol 2007;305-311 .

80. Vizcarra CL, Mayo SL. Electrostatics in computational protein design. Curr Opin Chem Biol 2005;622-626 .

81. Weis A, Katebzadeh K, Soderhjelm P, Nilsson I, Ryde U. Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. Journal of Medicinal Chemistry 2006;49:6596-6606.

82. Hu X, Kuhlman B. Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences. Proteins 2006;62:739-748.

83. Joughin B, Green D, Tidor B. Action-at-a-distance interactions enhance protein binding affinity. Protein science : a publication of the Protein Society 2005;14:1363-1369.

84. Kang S, Saven J. Computational protein design: structure, function and combinatorial diversity. Current Opinion in Chemical Biology 2007;11:329-334.

85. Wang L, Jiang T. On the complexity of multiple sequence alignment. J Comput Biol 1994;1:337-348.

86. Pirovano W, Heringa J. Multiple sequence alignment. Methods in molecular biology (Clifton, N.J.) 2008;452:143-161.

87. Thompson J, Higgins D, Gibson T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673-4680.

88. Thompson J, Gibson T, Higgins D. Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] 2002;Chapter 2:

89. Higgins D, Thompson J, Gibson T. Using CLUSTAL for multiple sequence alignments. Methods Enzymol 1996;266:383-402.

90. Higgins D, Sharp P. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 1988;73:237-244.

91. Chenna R, Sugawara H, Koike T, Lopez R, Gibson T, Higgins D, Thompson J. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003;31:3497-3500.

92. Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309-317.

93. Halperin I, Wolfson H, Nussinov R. Correlated mutations: Advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 2006;832-845.

94. Eyal E, Frenkel-Morgenstern M, Sobolev V, Pietrokovski S. A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. Proteins 2007;67:142-153.

95. Shackelford G, Karplus K. Contact prediction using mutual information and neural nets. Proteins: Structure, Function, and Bioinformatics 2007;69:159-164.

96. Horner D, Pirovano W, Pesole G. Correlated substitution analysis and the prediction of amino acid structural contacts. Brief Bioinform 2007;bbm052.

97. Kundrotas P, Alexov E. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics 2006;7:503.

98. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 2005;102:1029-1034.

99. Ortiz A, Strauss C, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606-2621.

100. Qiu J, Hue M, Ben-Hur A, Vert J, Noble WS. A structural alignment kernel for protein structures A structural alignment kernel for protein structures. Bioinformatics 2007;23:1090-1098.

101. Kedem K, Chew L, Elber R. Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. Proteins 1999;37:554-564.

102. Moeller C, Plesset M. Note on an Approximation Treatment for Many-Electron Systems. Physical Review 1934;46:618-622.

103. Hohenberg P, Kohn W. Inhomogeneous Electron Gas. Physical Review 1964;136:B864.

104. Boys S, Bernardi F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. Molecular Physics 1970;19:553-566.

105. Verlet L. Computer "Experiments" on Classical Fluids. II. Equilibrium Correlation Functions. Physical Review Online Archive (Prola) 1968;165:201-214.

106. Ewald P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. Ann. Phys. 1921;369:253-287.

107. Berendsen H, Postma J, van Gunsteren W, Dinola A, Haak J. Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics 1984;81:3684-3690.

108. Andrea T, Swope W, Andersen H. The role of long ranged forces in determining the structure and properties of liquid water ff. The Journal of Chemical Physics 1983;79:4576-4584.

109. Andersen H. Molecular dynamics simulations at constant pressure and/or temperature. The Journal of Chemical Physics 1980;72:2384-2393.

110. Pastor R, Brooks B, Szabo A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. Molecular Physics 1988;65:1409-1419.

111. Loncharich R, Brooks B, Pastor R. Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. Biopolymers 1992;32:523-535.

112. Izaguirre J, Catarello D, Wozniak J, Skeel R. Langevin stabilization of molecular dynamics. J. Chem. Phys. 2001;114:2090-2098.

113. Ryckaert J, Ciccotti G, Berendsen H. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. Journal of Computational Physics 1977;23:327.

114. Miyamoto S, Kollman P. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. Journal of Computational Chemistry 1992;13:952-962.

115. Drabik P, Liwo A, Czaplewski C, Ciarkowski J. The investigation of the effects of counterions in protein dynamics simulations. Protein Eng. 2001;14:747-752.

116. Kollman P, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case D, Cheatham T. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. Accounts of Chemical Research 2000;33:889-897.

117. Shrake A, Rupley J. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. Journal of molecular biology 1973;79:351-371.

118. Hubbard SJ. NACCESS Computer Program. 1993;

119. Richards F. Areas, Volumes, Packing, and Protein Structure. Annual Review of Biophysics and Bioengineering 1977;6:151-176.

120. Connolly M. Analytical molecular surface calculation. Journal of Applied Crystallography 1983;16:548-558.

121. Richmond T. Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. Journal of molecular biology 1984;178:63-89.

122. Connolly M. The molecular surface package. Journal of molecular graphics 1993;11:139-141.

123. Simonson T. Electrostatics and dynamics of proteins. Reports of Progress in Physics 2003;66:737-787.

124. Still C, Tempczyk A, Hawley R, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. Journal of the American Chemical Society 1990;112:6127-6129.

125. Onufriev A, Case D, Bashford D. Effective Born radii in the generalized Born approximation: the importance of being perfect. J Comput Chem 2002;23:1297-1304.

126. Lafont V, Schaefer M, Stote R, Altschuh D, Dejaegere A. Protein-protein recognition and interaction hot spots in an antigen-antibody complex: free energy decomposition identifies "efficient amino acids". Proteins 2007;67:418-434.

127. Zwanzig R. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. The Journal of Chemical Physics 1954;22:1420-1426.

128. Cieplak P, Cornell W, Bayly C, Kollman P. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. Journal of Computational Chemistry 1995;16:1357-1377.

129. Lengauer T, Rarey M. Computational methods for biomolecular docking. Curr Opin Struct Biol 1996;6:402-406.

130. Kitchen D, Decornez H, Furr J, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and appliations. Nat Rev Drug Discov 2004;3:935-949.

131. Meng E, Shoichet B, Kuntz I. Automated docking with grid-based energy evaluation. Journal of Computational Chemistry 1992;13:505-524.

132. Morris G, Goodsell D, Halliday R, Huey R, Hart W, Belew R, Olson A. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 1999;19:1639-1662.

133. Feig M, Onufriev A, Lee M, Im W, Case D, Brooks C. Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. Journal of computational chemistry 2004;25:265-284.

134. Hamelberg D, Shen T, McCammon A. Insight into the role of hydration on protein dynamics. The Journal of Chemical Physics 2006;125:094905.

135. Amadasi A, Spyrakis F, Cozzini P, Abraham DJ, Kellogg GE, Mozzarelli A. Mapping the energetics of water-protein and water-ligand interactions with the natural HINT forcefield: predictive tools for characterizing the roles of water in biomolecules. J Mol Biol 2006;358:289-309.

136. Li Z, Lazaridis T. Thermodynamics of buried water clusters at a protein-ligand binding interface. J Phys Chem B 2006;110:1464-1475.

137. Levy Y, Onuchic J. Water mediation in protein folding and molecular recognition. Annu Rev Biophys Biomol Struct 2006;35:389-415.

138. Mihalek I, Res I, Lichtarge O. On itinerant water molecules and detectability of protein-protein interfaces through comparative analysis of homologues. J Mol Biol 2007;369:584-595.

139. Rodier F, Bahadur R, Chakrabarti P, Janin J. Hydration of protein-protein interfaces. Proteins 2005;60:36-45.

140. Pisabarro M, Serrano L, Wilmanns M. Crystal structure of the abl-SH3 domain complexed with a designed high-affinity peptide ligand: implications for SH3-ligand interactions. J Mol Biol 1998;281:513-521.

141. Bork P, Holm L, Sander C. The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol 1994;242:309-320.

142. Lupyan D, Leo-Macias A, Ortiz A. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 2005;21:3255-3263.

143. Craft J, Legge G. An AMBER/DYANA/MOLMOL Phosphorylated Amino Acid Library Set and Incorporation into NMR Structure Calculations. Journal of Biomolecular NMR 2005;33:15-24.

144. Team R. R: a language and enviroment for statistical computing. 2006;Vienna, Austria-Vienna, Austria.

145. Levy Y, Onuchic J. Water and proteins: a love-hate relationship. Proc Natl Acad Sci U S A 2004;101:3325-3326.

146. Thomas DJ, Casari G, Sander C. The prediction of protein contacts from multiple sequence alignments. Protein Eng 1996;9:941-948.

147. Nagl S. Can correlated mutations in protein domain families be used for protein design? Brief Bioinform 2001;2:279-288.

148. Smith GR, Sternberg MJ, Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol 2005;347:1077-1101.

149. Vitkup D, Ringe D, Petsko GA, Karplus M. Solvent mobility and the protein 'glass' transition. Nat Struct Biol 2000;7:34-38.

150. Koehl P. Electrostatics calculations: latest methodological advances. Curr Opin Struct Biol 2006;16:142-151.

151. Collins KD, Neilson GW, Enderby JE. Ions in water: characterizing the forces that control chemical processes and biological structure. Biophys Chem 2007;128:95-104.

152. Imai T, Hiraoka R, Kovalenko A, Hirata F. Water molecules in a protein cavity detected by a statistical-mechanical theory. J Am Chem Soc 2005;127:15334-15335.

153. Gregoret L, Sauer R. Additivity of Mutant Effects Assessed by Binomial Mutagenesis. PNAS 1993;90:4246-4250.

154. Lee C, Levitt M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature 1991;352:448-451.

155. Wells J. Additivity of mutational effects in proteins. Biochemistry 1990;29:8509-8517.

156. Fodor A, Aldrich R. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 2004;56:211-221.

157. Pokarowski P, Kloczkowski A, Nowakowski S, Pokarowska M, Jernigan R, Kolinski A. Ideal amino acid exchange forms for approximating substitution matrices. Proteins: Structure, Function, and Bioinformatics 2007;69:379-393.

158. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511-523.

159. Perez-Jimenez R, Godoy-Ruiz R, Parody-Morreale A, Ibarra-Molero B, Sanchez-Ruiz J. A simple tool to explore the distance distribution of correlated mutations in proteins. Biophys Chem 2006;119:240-246.

160. Fuchs A, Martin-Galiano A, Kalman M, Fleishman S, Ben-Tal N, Frishman D. Co-evolving residues in membrane proteins. Bioinformatics 2007;23:3312-3319.

161. Fodor A, Aldrich R. On evolutionary conservation of thermodynamic coupling in proteins. J Biol Chem 2004;279:19046-19050.

162. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins: Structure, Function, and Genetics 2001;45:157-162.

163. Xue B F. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins: Structure, Function, and Bioinformatics 2008;Ahead of print:

164. O'Shea E, Klemm J, Kim P, Alber T. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science 1991;254:539-544.

165. Samsonov S, Teyra J, Pisabarro M. A molecular dynamics approach to study the importance of solvent in protein interactions. Proteins 2008;73:515-525.

166. Schueler-Furman O, Baker D. Conserved residue clustering and protein structure prediction. Proteins 2003;52:225-235.

167. McLachlan A. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. Journal of molecular biology 1971;61:409-424.

168. Teyra J, Paszkowski-Rogacz M, Anders G, Pisabarro T. SCOWLP classification: Structural comparison and analysis of protein binding regions. BMC Bioinformatics 2008;9:

169. Finn R, Mistry J, Schuster-B”ckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy S, Sonnhammer E, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34:D247-D251.

170. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56-68.

171. Yan C, Wu F, Jernigan R, Dobbs D, Honavar V. Characterization of protein-protein interfaces. The protein journal 2008;27:59-70.

172. Cohen M, Reichmann D, Neuvirth H, Schreiber G. Similar chemistry, but different bond preferences in inter versus intra-protein interactions. Proteins 2008;72:741-753.

173. Link A, Mock M, Tirrell D. Non-canonical amino acids in protein engineering. Current Opinion in Biotechnology 2003;14:603-609.

174. Smart B. Fluorine substituent effects (on bioactivity). Journal of Fluorine Chemistry 2001;109:3-11.

175. Dunitz J. Organic Fluorine: Odd Man Out. ChemBioChem 2004;5:614-621.

176. Mikami K, Itoh Y, Yamanaka M. Fluorinated carbonyl and olefinic compounds: basic character and asymmetric catalytic reactions. Chemical reviews 2004;104:1-16.

177. Zanda M. Trifluoromethyl Group: An Effective Xenobiotic Function for Peptide Backbone Modification. ChemInform 2005;36:1401-1411.

178. Jaeckel C, Salwiczek M, Koksch B. Fluorine in a Native Protein Environment - How the Spatial Demand and Polarity of Fluoroalkyl Groups Affect Protein Folding. Angewandte Chemie International Edition 2006;45:4198-4203.

179. Ma B, Nussinov R. Molecular dynamics simulations of a beta-hairpin fragment of protein G: balance between side-chain and backbone forces. Journal of molecular biology 2000;296:1091-1104.

180. Horng J, Raleigh D. Φ-Values beyond the Ribosomally Encoded Amino Acids: Kinetic and Thermodynamic Consequences of Incorporating Trifluoromethyl Amino Acids in a Globular Protein. Journal of the American Chemical Society 2003;125:9286-9287.

181. Niemz A, Tirrell D. Self-Association and Membrane-Binding Behavior of Melittins Containing Trifluoroleucine. Journal of the American Chemical Society 2001;123:7407-7413.

182. Zheng H, Comeforo K, Gao J. Expanding the Fluorous Arsenal: Tetrafluorinated Phenylalanines for Protein Design. Journal of the American Chemical Society 2009;131:18-19.

183. Yoder N, Kumar K. Fluorinated amino acids in protein design and engineering. Chemical Society reviews 2002;31:335-341.

184. Gregory D, Gerig J. Structural effects of fluorine substitution in proteins. Journal of Computational Chemistry 1991;12:180-185.

185. O'Hagan D. Understanding organofluorine chemistry. An introduction to the C-F bond. Chemical Society reviews 2008;37:308-319.

186. Bondi A. Van der Waals volumes and radii. Journal of Physisal Chemistry 1964;441-451.

187. O'Hagan D, Rzepa H. Some influences of fluorine in bioorganic chemistry. Chem. Commun. 1997;645-652.

188. Hyla-Kryspin I, Haufe G, Grimme S. Weak Hydrogen Bridges: A Systematic Theoretical Study on the Nature and Strength of C-H...F-C Interactions. Chemistry - A European Journal 2004;10:3411-3422.

189. Fischer F, Schweizer W, Diederich F. Molecular torsion balances: evidence for favorable orthogonal dipolar interactions between organic fluorine and amide groups. Angew Chem Int Ed Engl 2007;46:8270-8273.

190. Karaminkov R, Chervenkov S, Neusser H. Fluorine substitution and nonconventional OH-intramolecular bond: high-resolution UV spectroscopy and ab initio calculations of 2-(p-fluorophenyl)ethanol. Phys. Chem. Chem. Phys. 2008;10:2852-2859.

191. Headley A, Starnes S. Conformational analysis of alpha-trifluoroalanine: a theoretical study. Journal of Molecular Structure: THEOCHEM 2001;572:89-95.

192. Caminati W, Melandri S, Moreschini P, Favero P. The C-F...H-C 'Anti-Hydrogen Bond' in the Gas Phase: Microwave Structure of the Difluoromethane Dimer. Angewandte Chemie International Edition 1999;38:2924-2925.

193. Erickson J, McLoughlin J. Hydrogen Bond Donor Properties of the Difluoromethyl Group. The Journal of Organic Chemistry 1995;60:1626-1631.

194. Liu X, Borho N, Xu Y. Molecular Self-Recognition: Rotational Spectra of the Dimeric 2-Fluoroethanol Conformers. Chemistry - A European Journal 2009;15:270-277.

195. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry 1983;4:187-217.

196. Hess B, Kutzner C, Vanderspoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and

Scalable Molecular Simulation. J. Chem. Theory Comput. 2008.

197. Becke A. Density-functional exchange-energy approximation with correct asymptotic behavior. Physical review. A 1988;38:3098-3100.

198. Lee C, Yang W, Parr R. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys. Rev. B 1988;37:785-789.

199. Becke A. Density-functional thermochemistry. III. The role of exact exchange. The Journal of Chemical Physics 1993;98:5648-5652.

200. Yu W, Liang L, Lin Z, Ling S, Haranczyk M, Gutowski M. Comparison of some representative density functional theory and wave function theory methods for the studies of amino acids. Journal of Computational Chemistry 2008;30:589-600.

201. Schmidt M, Baldridge K, Boatz J, Elbert S, Gordon M, Jensen J, Koseki S, Matsunaga N, Nguyen K, Su S, Windus T, Dupuis M, Montgomery J. General atomic and molecular electronic structure system. J. Comput. Chem. 1993;14:1347-1363.

202. Grabowski S. Ab Initio Calculations on Conventional and Unconventional Hydrogen BondsStudy of the Hydrogen Bond Strength. The Journal of Physical Chemistry A 2001;105:10739-10746.

203. Chemical Computing Group Inc M. MOE v2005.06. 2006.

204. Duan Y, Wu C, Chowdhury S, Lee M, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 2003;24:1999-2012.

205. Gough C, Debolt S, Kollman P. Derivation of fluorine and hydrogen atom parameters using liquid simulations. J. Comput. Chem. 1992;13:963-970.

206. Discovery Studio ViewerPro A. Discovery Studio ViewerPro, Accelrys Inc, San Diego, CA. 2002.

207. Tsui V, Case D. Theory and applications of the generalized born solvation model in macromolecular simulations. Biopolymers 2001;56:291, 275-291, 275.

208. Gnuplot 4.2. Williams T, Kelley C. Gnuplot, Version 4.2. 2008.

209. Renfrew D, Butterfoss G, Kuhlman B. Using quantum mechanics to improve estimates of amino acid side chain rotamer energies. Proteins: Structure, Function, and Bioinformatics 2008;71:1637-1646.

210. Gomez-Catalan J, Perez J, Jimenez A, Cativiela C. Study of the conformational profile of selected unnatural amino acid residues derived from L-phenylalanine. Journal of Peptide Science 1999;5:251-262.

211. Pendley S, Yu Y, Thomas. Molecular dynamics guided study of salt bridge length dependence in both fluorinated and non-fluorinated parallel dimeric coiled-coils. Proteins: Structure, Function, and Bioinformatics 2008;74:612-629.

212. Hao M. Theoretical Calculation of Hydrogen-Bonding Strength for Drug Molecules. Journal of Chemical Theory and Computation 2006;2:863-872.

213. Joseph J, Jemmis E. Red-, Blue-, or No-Shift in Hydrogen Bonds: A Unified Explanation. J. Am. Chem. Soc. 2007;129:4620-4632.

214. Sarkhel S, Desiraju G. N-H...O, O-H...O, and C-H...O hydrogen bonds in protein-ligand complexes: Strong and weak interactions in molecular recognition. Proteins: Structure, Function, and Bioinformatics 2004;54:247-259.

215. Panigrahi S, Desiraju G. Strong and weak hydrogen bonds in the protein-ligand interface. Proteins 2007;67:128-141.

216. Jiang B, Guo T, Peng L, Sun Z. Folding type-specific secondary structure propensities of amino acids, derived from alpha-helical, beta-sheet, alpha/beta, and alpha+beta; proteins of known structures. Biopolymers 1998;45:35-49.

217. Zhao Y, Abraham M, Zissimos A. Fast calculation of van der waals volume as a sum of atomic and bond contributions and its application to drug compounds. J. Org. Chem. 2003;68:7368-7373.

218. Yin D, Mackerell A. Combined ab initio/empirical approach for optimization of Lennard-Jones parameters. Journal of Computational Chemistry 1998;19:334-348.

219. Sato A, Viswanathan M, Kent R, Wood C. Therapeutic peptides: technological advances driving peptides into development. Current Opinion in Biotechnology 2006;17:638-642.

220. Sanford K, Kumar M. New proteins in a materials world. Current Opinion in Biotechnology 2005;16:416-421.

221. Baldauf C, Pisabarro M. Stable Hairpins with b-Peptides: Route to Tackle Protein-Protein Interactions. Journal of Physical Chemistry B 2008;112:7581-7591.

222. Zhao H. Directed evolution of novel protein functions. Biotechnology and Bioengineering 2007;98:313-317.

223. Brannigan J, Wilkinson A. Protein engineering 20 years on. Nature Reviews Molecular Cell Biology 2002;3:964-970.

224. Budisa N. Prolegomena to future experimental efforts on genetic code engineering by expanding its amino acid repertoire. Angewandte Chemie International Edition 2004;43:6426-6463.

225. Wang L, Schultz P. Expanding the genetic code. Angewandte Chemie International Edition 2005;44:34-66.

226. Davis B. Chemical modification of biocatalysts. Current Opinion in Biotechnology 2003;14:379-386.

227. Chan W, White P. Fmoc solid phase peptide synthesis : a practical approach. 2000;xxiv, 346.

228. Tam J, Xu J, Eom K. Methods and strategies of peptide ligation. Peptide Science 2001;60:194-205.

229. David R, Richter M, Beck-Sickinger A. Expressed protein ligation. Method and applications. European Journal of Biochemistry 2004;271:663-677.

230. Purser S, Moore P, Swallow S, Gouverneur V. Fluorine in medicinal chemistry. Chemical Society Reviews 2008;37:320-330.

231. Gerebtzoff G, Li-Blatter X, Fischer H, Frentzel A, Seelig A. Halogenation of drugs enhances membrane binding and permeation. ChemBioChem 2004;5:676-684.

232. Hakelberg M, Koksch B. Coiled coil model systems as tools to evaluate the influence of fluorinated amino acids on structure and stability of peptides. Chemistry Today 2007;25:48-53.

233. Naarmann N, Bilgicer B, Meng H, Kumar K, Steinem C. Fluorinated interfaces drive self-association of transmembrane alpha helices in lipid bilayers. Angewandte Chemie International Edition 2006;45:2588-2591.

234. Meng H, Kumar K. Antimicrobial activity and protease stability of peptides containing fluorinated amino acids. Journal of the American Chemical Society 2007;129:15615-15622.

235. Gottler L, Lee H, Shelburne C, Ramamoorthy A, Marsh E. Using fluorous amino acids to modulate the biological activity of an antimicrobial peptide. ChemBioChem 2008;9:370-373.

236. Panchenko T, Zhu W, Montclare J. Influence of global fluorination on chloramphenicol acetyltransferase activity and

stability. Biotechnology and Bioengineering 2006;94:921-930.

237. Yoo T, Link A, Tirrell D. Evolution of a fluorinated green fluorescent protein. Proceedings of National Acadamy of Sciences of the United States of America 2007;104:13887-13890.

238. Voloshchuk N, Lee M, Zhu W, Tanrikulu I, Montclare J. Fluorinated chloramphenicol acetyltransferase thermostability and activity profile: improved thermostability by a single-isoleucine mutant. Bioorganic & Medicinal Chemistry Letters 2007;17:5907-5911.

239. Woll M, Hadley E, Mecozzi S, Gellman S. Stabilizing and destabilizing effects of phenylalanine --> F5-phenylalanine mutations on the folding of a small protein. Journal of the American Chemical Society 2006;128:15932-15933.

240. Dunitz J, Taylor R. Organic fluorine hardly ever accepts hydrogen bonds. Chemistry - A European Journal 1997;3:89-98.

241. Olsen J, Banner D, Seiler P, Wagner B, Tschopp T, Obst-Sander U, Kansy M, Muller K, Diederich F. Fluorine interactions at the thrombin active site: protein backbone fragments H-Ca-C=O comprise a favorable C-F environment and interactions of C-F with electrophiles. ChemBioChem 2004;5:666-675.

242. Jaeckel C, Koksch B. Fluorine in peptide design and protein engineering. European Journal of Organic Chemistry 2005;4483-4503.

243. Tsushima T, Kawada K, Ishihara S, Uchida N, Shiratori O, Higaki J, Hirata M. Fluorine containing amino acids and their derivatives. 7. Synthesis and antitumor activity of [alpha]- and [gamma]-substituted methotrexate analogs. Tetrahedron 1988;44:5375-5387.

244. Winkler D, Burger K. Synthesis of enantiomerically pure D- and L-armentomycin and its difluoroanalogues from aspartic acid. Synthesis 1996;11:1419-1421.

245. Osipov S, Lange T, Tsouker P, Spengler J, Hennig L, Koksch B, Berger S, El-Kousy S, Burger K. Hexafluoroacetone as a Protecting and Activating Reagent: Synthesis of New Types of Fluoro-Substituted ?-Amino, ?-Hydroxy and ?-Mercapto Acids. Synthesis 2004;12:1821-1829.

246. Fields G, Noble R. Solid phase peptide synthesis utilizing 9-fluorenylmethoxycarbonyl amino acids. International Journal of Peptide and Protein Research 1990;35:161-214.

247. Carpino L, El-Faham A. The diisopropylcarbodiimide/ 1-hydroxy-7-azabenzotriazole system: Segment coupling and stepwise peptide assembly. Tetrahedron 1999;55:6813-6830.

248. Kaiser E, Colescott R, Bossinger C, Cook P. Color test for detection of free terminal amino groups in the solid-phase synthesis of peptides. Analytical Biochemistry 1970;34:595-598.

249. Garcia-Echeverria C. Probing the formation of a parallel heterodimeric coiled coil by fluorescence quenching. Bioorganic & Medicinal Chemistry Letters 1997;7:1695-1698.

250. Pagel K, Seeger K, Seiwert B, Villa A, Mark A, Berger S, Koksch B. Advanced approaches for the characterization of a de novo designed antiparallel coiled coil peptide. Organic & Biomolecular Chemistry 2005;3:1189-1194.

251. Krylov D, Mikhailenko I, Vinson C. A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. EMBO Journal 1994;13:2849-2861.

252. Billo E. Excel for Scientists and Engineers - Numerical Methods. 2007;

253. Marky L, Breslauer K. Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. Biopolymers 1987;26:1601-1620.

254. C”lfen H, Harding S. MSTARA and MSTARI: interactive PC algorithms for simple, model independent evaluation of sedimentation equilibrium data. European Biophysics Journal 1997;25:333-346.

255. Schuck P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. Biophysical Journal 2000;78:1606-1619.

256. Hobza P, Kabel c M, Sponer J, Mejzilik P, Vondrasek J. Performance of empirical potentials (AMBER, CFF95, CVFF, CHARMM, OPLS, POLTEV), semiempirical quantum chemical methods (AM1, MNDO/M, PM3), and ab initio Hartree-Fock method for interaction of DNA bases: Comparison with nonempirical beyond Hartree-Fock results. Journal of Computational Chemistry 1997;18:1136-1150.

257. Showalter S, Bruschweiler R. Quantitative Molecular Ensemble Interpretation of NMR Dipolar Couplings without Restraints. Journal of the American Chemical Society 2007;129:4158-4159.

258. Mobley D, Dumont E, Chodera J, Dill K. Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent. Journal of Physical Chemistry B 2007;111:2242-2254.

259. Nicholls A, Mobley D, Guthrie J, Chodera J, Bayly C, Cooper M, Pande V. Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. Journal of Medicinal Chemistry 2008;51:769-779.

260. Pigache A, Cieplak P, Dupradeau F. Automatic and highly reproducible RESP and ESP charge derivation: Application to the developement of programs RED and X RED. 227th ACS National Meeting 2004;

261. Mason J, Arndt K. Coiled coil domains: stability, specificity, and biological implications. Chembiochem 2004;5:170-176.

262. Rose A, Meier I. Scaffolds, levers, rods and springs: diverse cellular functions of long coiled-coil proteins. Cellular and Molecular Life Sciences 2004;61:1996-2009.

263. Woolfson D. The design of coiled-coil structures and assemblies. Advances in Protein Chemistry 2005;70:79-112.

264. Monera O, Zhou N, Kay C, Hodges R. Comparison of antiparallel and parallel two-stranded alpha-helical coiled-coils. Design, synthesis, and characterization. Journal of Biological Chemistry 1993;268:19218-19227.

265. Wang J, Morin P, Wang W, Kollman P. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. Journal of the American Chemical Society 2001;123:5221-5230.

266. Chiu H, Suzuki Y, Gullickson D, Ahmad R, Kokona B, Fairman R, Cheng R. Helix propensity of highly fluorinated amino acids. Journal of the American Chemical Society 2006;128:15556-15557.

267. Regan L. Helix is a helix is a helix? Proceedings of National Acadamy of Sciences of the United States of America 1997;94:2796-2797.

268. Kwok S, Mant C, Hodges R. Effects of a-helical and b-sheet propensities of amino acids on protein stability. 25th European Peptide Symposium 1998;34-35.

269. Wagschal K, Tripet B, Lavigne P, Mant C, Hodges R. The role of position a in determining the stability and oligomerization state of alpha-helical coiled coils: 20 amino acid stability coefficients in the hydrophobic core of proteins. Protein Science 1999;8:2312-2329.

270. Tripet B, Wagschal K, Lavigne P, Mant C, Hodges R. Effects of side-chain characteristics on stability and oligomerization state of a de novo-designed model coiled-coil: 20 amino acid substitutions in position "d". Journal of Molecular Biology 2000;300:377-402.

271. Parry DAD, Fraser RDB, Squire JM. Fifty years of coiled-coils and alpha-helical bundles: A close relationship between

sequence and structure. Journal of structural biology 2008;163:258-269.

272. Moutevelis E, Woolfson D. A Periodic Table of Coiled-Coil Protein Structures. Journal of Molecular Biology 2009;385:726-732.

273. Pagel K, Koksch B. Following polypeptide folding and assembly with conformational switches. Current opinion in chemical biology 2008;12:730-739.

274. Woolfson D, Alber T. Predicting oligomerization states of coiled coils. Protein science: a publication of the Protein Society 1995;4:1596-1607.

275. Lumb K, Kim P. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. Biochemistry 1995;34:8642-9648.

276. Gonzalez L, Woolfson D, Alber T. Buried polar residues and structural specificity in the GCN4 leucine zipper. Nat Struct Mol Biol 1996;3:1011-1018.

277. Bromley E, Sessions R, Thomson A, Woolfson D. Designed alpha-helical tectons for constructing multicomponent synthetic biological systems. Journal of the American Chemical Society 2009;131:928-930.

278. Schneider J, Lear J, Degrado W. A Designed Buried Salt Bridge in a Heterodimeric Coiled Coil. Journal of the American Chemical Society 1997;119:5742-5743.

279. Kretsinger J, Schneider J. Design and application of basic amino acids displaying enhanced hydrophobicity. Journal of the American Chemical Society 2003;125:7907-7913.

280. Diss M, Kennan A. Orthogonal recognition in dimeric coiled coils via buried polar-group modulation. Journal of the American Chemical Society 2008;130:1321-1327.

281. Kehoe J, Kay B. Filamentous phage display in the new millennium. Chemical reviews 2005;105:4056-4072.

282. Segal D, Dreier B, Beerli R, Barbas C. Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. Proceedings of the National Academy of Sciences of the United States of America 1999;96:2758-2763.

283. Dreier B, Beerli R, Segal D, Flippin J, Barbas C. Development of Zinc Finger Domains for Recognition of the 5'-ANN-3' Family of DNA Sequences and Their Use in the Construction of Artificial Transcription Factors. J. Biol. Chem. 2001;276:29466-29478.

284. Nathan S, Rader C, Barbas C. Neutralization of Burkholderia pseudomallei protease by Fabs generated through phage display. Bioscience, biotechnology, and biochemistry 2005;69:2302-2311.

285. Welch B, Vandemark A, Heroux A, Hill C, Kay M. Potent D-peptide inhibitors of HIV-1 entry. Proceedings of the National Academy of Sciences 2007;104:16828-16833.

286. Lai J, Fisk J, Weisblum B, Gellman S. Hydrophobic core repacking in a coiled-coil dimer via phage display: insights into plasticity and specificity at a protein-protein interface. Journal of the American Chemical Society 2004;126:10514-10515.

287. Hagemann UB, Mason JM, Mueller KM, Arndt KM. Selectional and mutational scope of peptides sequestering the jun-fos coiled-coil domain. Journal of molecular biology 2008;73-88.

288. Barbas C, Wagner, J. Synthetic Human Antibodies: Selecting and Evolving Functional Proteins. Methods, A Companion to Methods in Enzymology 1995;8:94-103.

289. Reidhaar-Olson J, Sauer R. Combinatorial cassette mutagenesis as a probe of the informational content of protein

sequences. Science (New York, N.Y.) 1988;241:53-57.

290. Batzer M. Phage Display: A Laboratory Manual. Edited by C. F. Barbas III, D. R. Burton, J. K. Scott, and G. J. Silverman. Analytical Biochemistry 2001;294:194.

291. Salwiczek M, Samsonov S, Vagt T, Nyakatura E, Fleige E, Numata J, Coelfen H, Pisabarro M, Koksch B. Position Dependent Effects of Fluorinated Amino Acids on Hydrophobic Core Formation of a Heterodimeric Coiled Coil. Chem. Eur. J. 2009; accepted for publication.

List of Tables

List of Figures

Figure 2.2.5. Predictions for interdomain dataset. Relative harmonic weighted difference statistic (X_d) dependence on α.

Figure 2.2.6. Proportion of residue pairs at distance bins for the interaction SH2-SH3.All residue pairs are shown in black, correlated pairs with α=0 in white, and correlated pairs with α=0.2 in grey. Reference structure used is PDB ID 2src.

Figure 3.1.1. Charges and bond lengths in ethane and ethane derivatives. All methods were applied with the same basis set (6-311G**++).

A) Charge of C1, a not-fluorinated carbon. B) Charge of C2, a fluorinated carbon. C) Charge of H(C1), hydrogen bound to C1. D) When 1 or 2 fluorine atoms are bonded to C2, the charge distribution on different H(C1) is not symmetric around the C1-C2 bond. This figure represents the highest charges on the H(C1) atoms. E) Charge of H(C2), hydrogen bound to C2. F) F charge. G) C1-C2 bond length. H) C1-H(C1) bond length. I) C2-H(C2) bond length. J) C2-F bond length.

Figure 3.1.2. Electric dipoles calculated in B3LYP (6-311G**++).

A) Monofluoroethane. B) Difluoroethane. C) Trifluoroethane. D) 2,2-difluoropropane. Fluorine atoms are green, hydrogens are white, carbons are dark green. Dipoles are represented as arrows.

Figure 3.1.3. Hydrogen bond acceptors properties.

The box plots contain data on all analyzed hydrogen bonds calculated in B3PLYP (6-311G**++). A) Hydrogen bond energy with BSSE correction. B) Relative shift in the covalent bond between hydrogen and heavy atom upon hydrogen bond formation. C) Hydrogen bond length. D) Hydrogen bond angle ∠(D-H-A). E) Hydrogen atom charge transfer. F) Acceptor atom charge transfer.

Figure 3.1.4. Volumes and solvent accessible surface areas (ASA).

A) Fluoromethyl groups after B3LYP (6-311G**++) geometry optimization. B) Amino acids side chains created for AMBER libraries.

Figure 3.1.5. Ramachandran plots of the fluorinated amino acids in comparison to canonical hydrophobic amino acids.

Figure 3.1.6. Covariance (propensity index) between probabilities obtained from calculated Ramachandran plots and PDB-derived secondary structure data. A) β-strand. B) α-helix. C) Left α-helix.

Figure 3.1.7. Side chain rotamers potential energy $E(\chi_1)$-$E_{min}(\chi_1)$ of the fluorinated ethylglycine derivatives in different backbone conformations. A) β-strand. B) α-helix. C) Left α-helix.

Figure 3.1.8. β-strand conformation for Ace-Xxx-Nme dipeptides, where Xxx= A) Abu. B) MfeGly. C)

List of publications in peer-reviewed journals:

1. **Samsonov S.**, Teyra J., Pisabarro M.T. A molecular dynamics approach to study the importance of solvent in protein interactions. Vol. 73. No. 2. *Proteins*. 2008. p. 515-525.

2. Vagt, T.; Jäckel, C.; **Samsonov, S.**; Pisabarro, M. T.; Koksch, B. "Selection of a buried salt bridge by phage display". *Bioorg. Med. Chem. Lett.* (2009) In Press.

3. Salwiczek, M.; **Samsonov, S.**; Vagt, T.; Nyakatura, E.; Fleige, E.; Numata, J.; Cölfen, H.; Pisabarro, M. T.; Koksch, B. "Position Dependent Effects of Fluorinated Amino Acids on Hydrophobic Core Formation of a Heterodimeric Coiled Coil". *Chem. Eur. J.* (2009) In Press.

4. **Samsonov, S.**; Teyra, J.; Anders, G.; Pisabarro, M. T. "Analysis of the impact of solvent on contacts prediction in proteins". *BMC Struct. Biol.* Vol. 9, No. 1, 2009. 22.

5. Samsonov S., Salwiczek M., Anders G., Koksch B., Pisabarro M.T. “Characterization of fluorinated amino acids by QM and MD approaches”. Proteins: Structure, Function, and Bioinformatics. Submitted.