

MEMBRANE PROTEIN STRUCTURE DETERMINATION USING NMR
SPECTROSCOPY AND COMPUTATIONAL TECHNIQUES

By

Julia Koehler Leman

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

August, 2012

Nashville, Tennessee

Approved:

Jens Meiler

Charles Sanders

Hassane Mchaourab

Brandt Eichman

In memory of Dr. Anne Karpay

A life without dreams is like a garden without flowers.

unknown

ACKNOWLEDGEMENTS

Most likely, the thanks that I will be expressing here will be insufficient to account for all that my family, colleagues and friends have done for me during the past six years.

I want to express my gratitude to my husband Sugani for all his love and support for me during the past years since we got to know each other. He always believed in me and has built me up in some of the most difficult times towards the end of my graduate school experience. Even though he had to endure living a long distance away from me during much of this time, he always supported me and cared for me in the sweetest ways possible.

Next, I would like to thank both of my advisors Jens Meiler and Chuck Sanders who are both absolutely wonderful mentors as well as great personalities. To come to Vanderbilt and join both of their labs was one of the best decisions I have ever made. From them I have not only learnt how to approach scientific questions, design and effectively carry out meaningful experiments and how to present my research to a broader community but they have also fostered my development as a person. Through their constant support, advice and encouragement I became the scientist and person that I am today and I could not be more grateful for that.

I am also very thankful to many former and current lab members who have helped me tremendously with computations and experiments, fruitful discussions and taught me a lot of little details of how to get things done. Not only have they helped me with technical questions but they made life in the lab a lot more enjoyable and it was fun to work with and be around such wonderful people: René Staritzbichler, Nils Wötzel, Mert Karakaş, Nathan Alexander, Ralf Müller, Mariusz Butkiewicz, Brian Weiner, Jeff Mendenhall, Wade van Horn, Congbao Kang, Masayoshi Sakakura, Arina Hadziselimovic, Danielle Kelley, Chuck Mobley. I also want to thank everybody else of

the Meiler and Sanders lab members who have not impacted the research on my projects on a major level but have helped me at one point or another. I am extremely grateful to have been able to work in such a great environment with so many friendly and terrific people. Along the same lines I want to thank the excellent support staff in the Center for Structural Biology that have made my research so much more productive: Markus Voehler, Don Stec, Sabuj Pattanayek, Roy Hoffman, and the staff at ACCRE.

Of course I want to thank all the wonderful friends that I have made during my time at Vanderbilt and I hope that at least some of these friendships will last a lifetime. Elizabeth Dong is an extraordinary colleague and friend whom I have met about a month after I joined the Meiler lab. She was doing a summer research project at Vanderbilt while still being an undergraduate at Washington University in Seattle. I sometimes have to smile thinking back to that first summer that we spent together in the lab trying to figure out Perl – none of us had any programming experience and we were both very happy about every single command that we could get to work. It was fortunate that she was later accepted to the MD/PhD program at Vanderbilt and she could not elude from Jens talking her into joining his lab. We spent a lot of time together in and out of the lab – talking about research projects, life in general, enjoying fun things like cooking, playing piano, and watching silly movies. ;o)

There are many more girls who made my life in and around the lab much more pleasurable: Steph – it was very enjoyable seeing her progress during the years she has been in the lab and she is truly a very good friend – Veronica, Meryl, Claudia, Nicole, Thuy, and Brittany.

I also want to mention the friends I got to know outside the lab: Nino – one of the dearest friends I ever had, who guided me through some really difficult times and helped me become a more mature person. She always gave me a different perspective and is a fun person to be around. Many people whom I have met during the past couple of years

have taught me many different things and I know for sure that I will miss each and every one of them. I really enjoyed getting to know such a diverse set of people from all around the globe, whether it is China, India, Turkey, Italy, Spain, Belgium, Colombia or Peru: Doris, Margarita, Stijn, Will, Jason, Hugo, Mark, Alper, Silvia, Carlos, Juan, Angela P., Vikash, Domenico, Steve, Wei, Angela Z. and many others.

TABLE OF CONTENTS

| | |
|---|-----|
| ACKNOWLEDGEMENTS..... | i |
| TABLE OF CONTENTS..... | iv |
| LIST OF FIGURES..... | vii |
| LIST OF TABLES..... | ix |
| ABSTRACT..... | x |
| REFERENCES..... | xii |
| | |
| INTRODUCTION..... | 1 |
| Techniques for protein structure determination..... | 1 |
| Membrane protein structure determination using conventional techniques remains difficult..... | 3 |
| Advances in NMR spectroscopy to push the limits for protein structure elucidation .. | 5 |
| Paramagnetic NMR as an avenue in protein structure elucidation..... | 9 |
| Paramagnetic effects depend on magnetic susceptibility anisotropy..... | 10 |
| Newer Methods in protein structure elucidation: the power of computation..... | 12 |
| Hydrophobicity as the beginnings of protein secondary structure prediction..... | 13 |
| Artificial Intelligence and Machine Learning Techniques as powerful tools in sequence-based protein structure prediction..... | 14 |
| Protein structure in three dimensions: challenges and available methods..... | 17 |
| References..... | 22 |
| | |
| CHAPTER 1 | |
| Expanding the utility of NMR restraints with paramagnetic compounds: | |
| Background and practical aspects..... | 27 |
| Introduction..... | 27 |
| Magnetic susceptibility and its anisotropy..... | 28 |
| Residual Dipolar Couplings..... | 36 |
| Chemical Shift contributions..... | 41 |
| Structure calculations using PCSs and RDCs..... | 52 |

| | |
|--|-----|
| Relaxation | 65 |
| Lanthanides and other paramagnetic probes..... | 86 |
| Interfaces | 90 |
| Conclusions..... | 92 |
| References..... | 93 |
| | |
| CHAPTER 2 | 107 |
| Structural studies on KCNE3 using paramagnetic restraints obtained from lanthanide tagging experiments..... | 107 |
| Introduction..... | 107 |
| Methods | 112 |
| Results | 116 |
| Discussion | 128 |
| Conclusion and future directions..... | 133 |
| References..... | 135 |
| | |
| CHAPTER 3 | 139 |
| A Unified Hydrophobicity Scale for Multi-Span Membrane Proteins..... | 139 |
| Introduction..... | 139 |
| Methods | 147 |
| Results and Discussion | 152 |
| Conclusions..... | 179 |
| References..... | 181 |
| | |
| CHAPTER 4 | 186 |
| Improved prediction of trans-membrane spans in proteins using an Artificial Neural Network..... | 186 |
| Introduction..... | 186 |
| Methods | 189 |
| Results and Discussion | 193 |
| Conclusion..... | 201 |
| References..... | 203 |

| | |
|---|---------|
| CHAPTER 5 | 206 |
| Simultaneous prediction of protein secondary structure and trans-membrane spans... | 206 |
| Introduction..... | 206 |
| Methods | 210 |
| Results | 215 |
| Discussion | 220 |
| Conclusions..... | 224 |
| References..... | 225 |
| CONCLUSIONS | 229 |
| APPENDIX | |
| Appendix to Chapter 1 | 236 |
| Appendix to Chapter 2..... | 252 |
| Appendix to Chapter 3..... | 259 |
| Appendix to Chapter 5..... | 272 |
| Lyso-Phospholipid Micelles Sustain the Stability and Catalytic Activity of Diacylglycerol Kinase in the Absence of Lipids..... | 286 |

LIST OF FIGURES

| Figure: | page |
|--|------|
| 1 Backbone NOEs on α -helical proteins and β -barrels | 4 |
| 2 Schematic of HSQC and TROSY peaks in an NMR spectrum | 6 |
| 3 Spectral improvements by the TROSY effect..... | 7 |
| 4 Overall context of paramagnetic restraints..... | 11 |
| 5 Architecture of an Artificial Neural Network..... | 15 |
| 6 Methods in protein structure prediction | 18 |
| 7 <i>De novo</i> models predicted using RosettaMembrane..... | 20 |
| | |
| 1-1 Methods to introduce a paramagnetic center into a protein..... | 34 |
| 1-2 Largest measurable paramagnetic restraints | 37 |
| 1-3 Structure calculation protocol for RDCs and PCSs | 56 |
| 1-4 Graphical representation of solutions of Eq.22 in the tensor frame | 60 |
| 1-5 Interpretation of PRE data | 86 |
| | |
| 2-1 Whole-cell currents for KCNE proteins | 108 |
| 2-2 NMR structure of KCNE1 and models of KCNQ1-KCNE1 complex | 109 |
| 2-3 Snake plot of KCNE3..... | 110 |
| 2-4 Lanthanide-binding tags for tagging KCNE3..... | 111 |
| 2-5 Dimerization tendency of KCNE3 mutants..... | 116 |
| 2-6 Dimerization tendency of KCNE3 at different pH | 117 |
| 2-7 Dimerization of KCNE3 before and after protocol optimization..... | 118 |
| 2-8 Mass-spectrometry data of KCNE3 tagged with different tags | 120 |
| 2-9 NMR spectra of WT KCNE3 with KCNE3 tagged with MTS-EDTA | 122 |
| 2-10 NMR spectra of untagged vs. tagged KCNE3 | 122 |
| 2-11 NMR spectra of KCNE3 tagged with different tags..... | 124 |
| 2-12 NMR spectra of paramagnetic vs. diamagnetic tagged KCNE3..... | 125 |
| 2-13 PCSs and PREs on KCNE3..... | 126 |
| 2-14 Histograms of RDCs and PCSs | 127 |

| | | |
|------|--|-----|
| 2-15 | RDCs vs. residue number using different tags | 128 |
| 3-1 | Range of transfer free energy values for different hydrophobicity scales | 142 |
| 3-2 | Definition of membrane regions | 149 |
| 3-3 | Correlation plots for different hydrophobicity scales | 156 |
| 3-4 | Prediction accuracies vs. window length..... | 159 |
| 3-5 | Correlation plot between UHS and scale of Moon & Fleming..... | 170 |
| 3-6 | Correlation plot Universal Hydrophobicity Scale vs. Mammalian Hydrophobicity Scale..... | 171 |
| 3-7 | Examples for predicting TM spans using Universal Hydrophobicity Scale..... | 174 |
| 3-8 | Close-up of one of the examples for testing..... | 176 |
| 4-1 | Side-chain orientation in trans-membrane β -barrels | 187 |
| 4-2 | Training behavior..... | 194 |
| 4-3 | Weight distribution in the Artificial Neural Network..... | 195 |
| 4-4 | Examples for predicting TM spans..... | 200 |
| 5-1 | Definition of the membrane used for training and predicting | 211 |
| 5-2 | Cross-validation procedure..... | 212 |
| 5-3 | Inputs for the Artificial Neural Networks | 214 |
| 5-4 | Prediction accuracies over nine and three states..... | 216 |
| 5-5 | Three-state secondary structure prediction accuracies compared to other methods | 217 |
| 5-6 | Two-state TM span prediction accuracies compared to other methods..... | 218 |
| 5-7 | Examples used for testing | 219 |

LIST OF TABLES

| Table: | page |
|--|------|
| 3-1 Values of the water-membrane transfer free energies | 153 |
| 3-2 Summary of the different hydrophobicity scales..... | 154 |
| 3-3 Free energy values in the Universal Hydrophobicity Scale and in the Mammalian Hydrophobicity Scale..... | 157 |
| 3-4 Per amino acid agreements for the two-state scenario | 161 |
| 3-5 Per amino acid agreements for the three state scenario | 165 |
| 4-1 Prediction accuracies in the three-state scenario..... | 199 |

ABSTRACT

The focus of this dissertation is the development of computational as well as experimental NMR methods to facilitate membrane protein structure determination. The introduction chapter first describes the techniques that are available for structural studies of proteins and what challenges have to be overcome when working with membrane proteins. Paramagnetic restraints are introduced that have the potential to replace conventional NOE restraints as structural restraints where they are unavailable. The computational section surveys hydrophobicity scales as the beginnings of sequence-based protein structure prediction and describes Artificial Neural Networks as a tool to produce high quality predictions. Computational techniques for protein fold determination are described.

Chapter 1 gives an overview of paramagnetic restraints that is largely a reproduction of a review published in *Progress in NMR Spectroscopy* [1]. The existence of magnetic susceptibility anisotropy dictates which of the paramagnetic effects – Residual Dipolar Couplings (RDCs), Paramagnetic Relaxation Enhancements (PREs), Pseudo-Contact Shifts (PCSs), and other Cross-Correlated Relaxation (CCR) effects – can be observed. Magnetic susceptibility anisotropy can be introduced into a protein by attachment of a small-molecule tag that chelates a metal ion, specifically a lanthanide ion. This chapter not only reviews the effects that lead to changes in chemical shifts or scalar couplings (such as RDCs and PCSs) but also examines relaxation effects that alter the line-width of the resonances (such as PREs and CCR). This work is the first to bring together these two specialized areas of NMR in the context of lanthanide-binding tags and evaluates the use of particular lanthanides for specific scenarios.

Paramagnetic restraints have been measured by tagging the protein KCNE3 with lanthanide-binding tags, as described in Chapter 2. The development of the protocols is

described together with control experiments that were carried out to assure that KCNE3 is fully tagged and that the lanthanide ion is bound to the tag. Paramagnetic restraints such as PREs, RDCs, and PCSs are measured on this system using three different types of lanthanide-binding tags. This work is currently being compiled into a manuscript.

The computational part begins with the development of a 'Unified Hydrophobicity Scale' described in Chapter 3 that is reproduced from reference [2]. The hydrophobicity scale was derived from a database of soluble proteins and membrane proteins and is the first knowledge-based scale that is equally valid for multi-span α -helical membrane proteins as well as β -barrels at the same time. To benchmark the hydrophobicity scale, a simple window averaging function is used. Results indicate that the developed hydrophobicity scale achieves higher accuracies at predicting trans-membrane spans than available hydrophobicity scales.

The developed hydrophobicity scale is a prerequisite for the prediction of trans-membrane spans in Chapter 4 where the simple window averaging function is replaced by an Artificial Neural Network (ANN). Prediction accuracies increase dramatically by this approach from an overall prediction accuracy of 57% using the simple averaging function to 79% using the ANN. This chapter is a reproduction of a manuscript that was published in the conference proceedings of the *IEEE symposium on Computational Intelligence in Bioinformatics and Computational Biology* [3] and for which I received the best overall paper award.

A continuation of this project is the combined prediction of secondary structure and trans-membrane spans for both α -helical proteins and β -barrels as described in Chapter 5 (reproduced from a manuscript submitted to *Proteins, Structure, Function, and Bioinformatics*). Again, the knowledge-based hydrophobicity scale is used as an input to the ANNs. This is the first prediction tool that is available for both secondary structure types in the membrane and supersedes the use of several prediction methods with the

requirement to build a consensus which can be especially difficult to obtain in case of contradicting predictions from multiple methods.

Appendices to Chapters 1, 2, 3, and 5 describe experiments with methods and protocols that are unpublished. Appendix 6 is a reproduction of reference [4] that probes interactions between Diacylglycerol kinase (DAGK) with its micelle environment.

REFERENCES

- [1] J. Koehler, J. Meiler, Expanding the utility of NMR restraints with paramagnetic compounds: Background and practical aspects, *Prog Nucl Mag Res Sp*, 59 (2011) 360-389.
- [2] J. Koehler, N. Woetzel, R. Staritzbichler, C.R. Sanders, J. Meiler, A unified hydrophobicity scale for multispan membrane proteins, *Proteins*, (2008).
- [3] J. Koehler, R. Mueller, J. Meiler, Improved prediction of trans-membrane spans in proteins using an Artificial Neural Network, *IEEE Comp. Intel. Bioinf. Comp. Biol.*, (2009).
- [4] J. Koehler, E.S. Sulistijo, M. Sakakura, H.J. Kim, C.D. Ellis, C.R. Sanders, Lysophospholipid Micelles Sustain the Stability and Catalytic Activity of Diacylglycerol Kinase in the Absence of Lipids, *Biochemistry*, (2010).

INTRODUCTION

Introduction

Currently, there are about 15 million non-redundant protein sequences known [1-3], about 71,000 of which have determined structures in the ProteinDataBank (PDB) [4]. Of these, only about 2%, which are about 1,500 structures, belong to membrane proteins [6].

Membrane proteins make up about 30% of the proteins in our body, however, over 50% of all drugs target them. This is not surprising, because as proteins sitting in the cell membrane that define the boundary of cells or cell organelles, they have crucial functions as signaling molecules, transport proteins, receptors, and are involved in cell adhesion, stabilization, and catalysis. The reason why so few membrane protein structures have been determined is the difficulty of elucidating their structure. The next section will briefly review some of the techniques used for protein structure determination and later discuss the challenges that arise when working with membrane proteins.

Techniques for protein structure determination

The most prominent method used to determine protein structures is X-ray crystallography. After over-expression of the protein of interest, the solution containing the protein is supplemented with salts and reagents that induce crystallization of the protein. The crystallized protein is subjected to an X-ray beam that is diffracted in different directions to yield a diffraction pattern. The diffraction pattern is a representation of the crystal structure in reciprocal space and the crystal structure in real space can be computed using a Fourier transform. The challenge is to overcome the phase problem

[7], that is to determine both the phase and the amplitude of the structure factor since the phase cannot be directly obtained from the experiment. Molecular replacement with a structure of a related protein or homologue [7] or isomorphous replacement [8] using heavy atoms are two possible solutions to this problem.

The second most used method is solution state nuclear magnetic resonance (NMR) spectroscopy. In contrast to X-ray crystallography, which yields static protein structures, NMR spectroscopy can provide dynamic information as well as information about binding interactions, secondary structure, the size of the protein complex(es) and even intrinsically unstructured proteins [9-10]. In NMR spectroscopy the protein sample is exposed to a static external magnetic field and the response of the nuclear spins to a pulse train of radio-frequency magnetic fields is measured. The conventional protocol for protein structure determination is over-expressing suitable amounts of protein, assignment of resonances using three-dimensional experiments, and measuring restraints that can be used in subsequent structure calculations. As restraints, typically Nuclear Overhauser Enhancements (NOEs [11]), J-couplings, and sometimes Residual Dipolar Couplings (RDCs [12]) are used for soluble proteins.

Another technique to obtain distance and accessibility restraints is electron paramagnetic resonance (EPR) Spectroscopy. Whereas being similar to NMR spectroscopy in subjecting the sample to an external magnetic field, EPR spectroscopy measures the response of an electron spin as compared to a nuclear spin in NMR to an external magnetic field. This requires the introduction of one or two free radicals (unpaired electrons) into the protein which is typically accomplished by attachment of a spin-label (methane-thio-sulfonate (MTSL) or other [13]) to a thiol-reactive cysteine at the surface of the protein. This requirement makes EPR spectroscopy less effective as NMR in terms of number of distance restraints that can be obtained since one cysteine double-mutant must be created for each EPR distance restraint to be measured. For

structure determination of membrane proteins, however, this limitation is sometimes outweighed by the ability to study membrane proteins in more native like membranes such as vesicles, which cannot be studied by NMR spectroscopy due to their size.

Electron microscopy is another important technique in structural biology. A flash-frozen sample of a large protein or membrane protein or virus particles is imaged using an electron beam. Thousands of these images are collected, analyzed and can be computationally reconstructed into a three-dimensional electron density map. This map is a low-resolution representation of the three-dimensional structure of the imaged molecule. The experimental resolution of the highest quality structures lies currently at less than 4 Å [14], however, a typical density map has a resolution of 10-15 Å. This may be sufficient to resolve helices in the protein, while strand and loop residues remain unresolved.

Membrane protein structure determination using conventional techniques remains difficult

The challenges for membrane protein structure determination are manifold. First, over-expression of suitable amounts of protein for NMR spectroscopy or X-ray crystallography is difficult because many membrane proteins are either toxic to the host cell or misfold. Expression in yeast, insect cells, eukaryotic strains or cell-free expression [15] may be able to alleviate this problem but these avenues are far less common and less well characterized than the typically used over-expression in *E.coli*.

When this challenge is overcome, some membrane proteins are found not to crystallize. In this regard, the use of NMR spectroscopy may be the only option to obtain a high-resolution protein structure. However, since membrane proteins have long stretches of hydrophobic residues, they require the presence of detergents or lipids to retain their native fold. This is a limitation for NMR spectroscopy because the

incorporation of the protein into detergent micelles or detergent/lipid bicelles increases the rotational correlation time of this complex. This leads to a whole cascade of difficulties that make membrane protein structure determination incomparably more challenging than structure determination of soluble proteins, even to the point where the

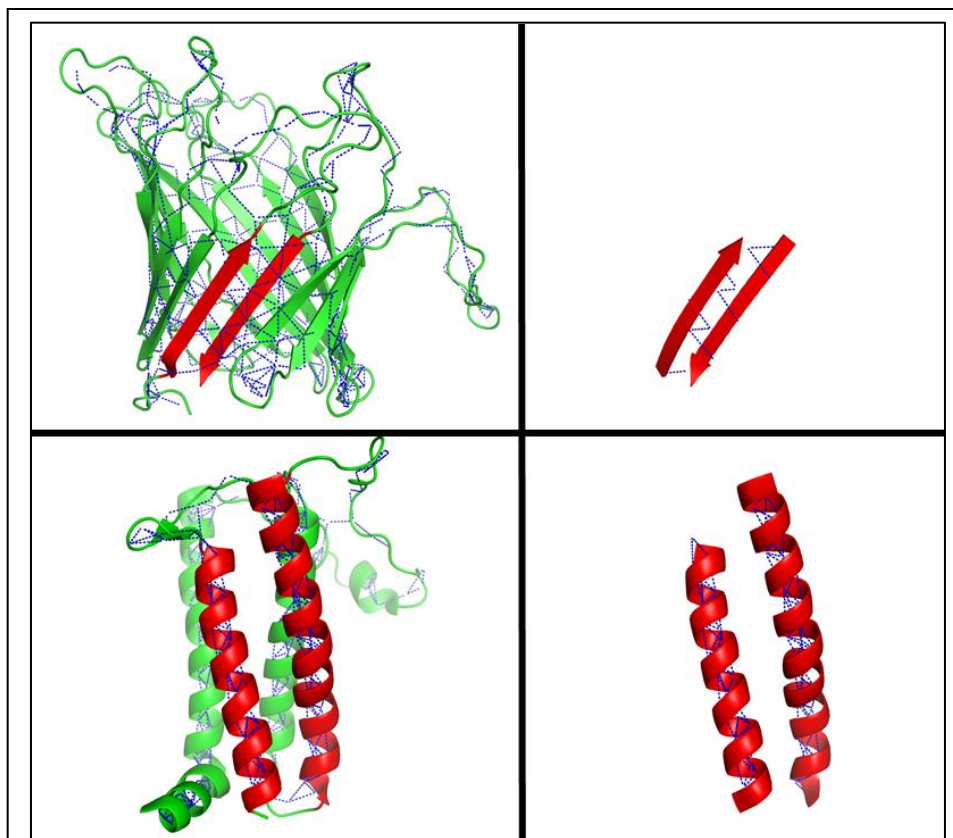


Figure 1: Measurable ^1H - ^1H backbone NOEs ($<5 \text{ \AA}$) for fold determination in membrane proteins. The upper panel shows that backbone NOEs in a β -barrel membrane protein (2JQY) allow for the connection of neighboring strands to determine the overall fold, whereas in the lower panel available NOEs are insufficient to connect neighboring helices in an α -helical membrane protein (2K73). Since side-chain information is difficult or impossible to acquire, the structure determination of large helical proteins remains challenging.

structure cannot be determined by NMR: an increase of the line-width of the resonances in the spectra leads to a smaller signal-to-noise ratio and therefore less detectable peaks. The signal-to-noise ratio of side-chain resonances is often so poor that these peaks remain undetectable. This means that side-chain resonances cannot be assigned and restraints involving these atoms cannot be obtained. This is particularly unfortunate

since the most useful distance restraints for structure calculations may be the NOEs measured between the side-chain atoms of neighboring secondary structure elements, especially for helical proteins. For structure determination of β -barrel membrane proteins NOE restraints from the side-chain atoms are fortunately not required to determine the three-dimensional fold of the protein since the fold can be established by the proton-proton distances involving the backbone atoms. For α -helical membrane proteins however, the most effective way to determine the fold of the protein involves NOEs between the side-chain atoms of neighboring helices as shown in Figure 1. Another complication for α -helical proteins is the smaller spectral dispersion in the hydrogen dimension of the spectrum resulting in overlapping peaks which complicates peak assignment and the measurement of restraints.

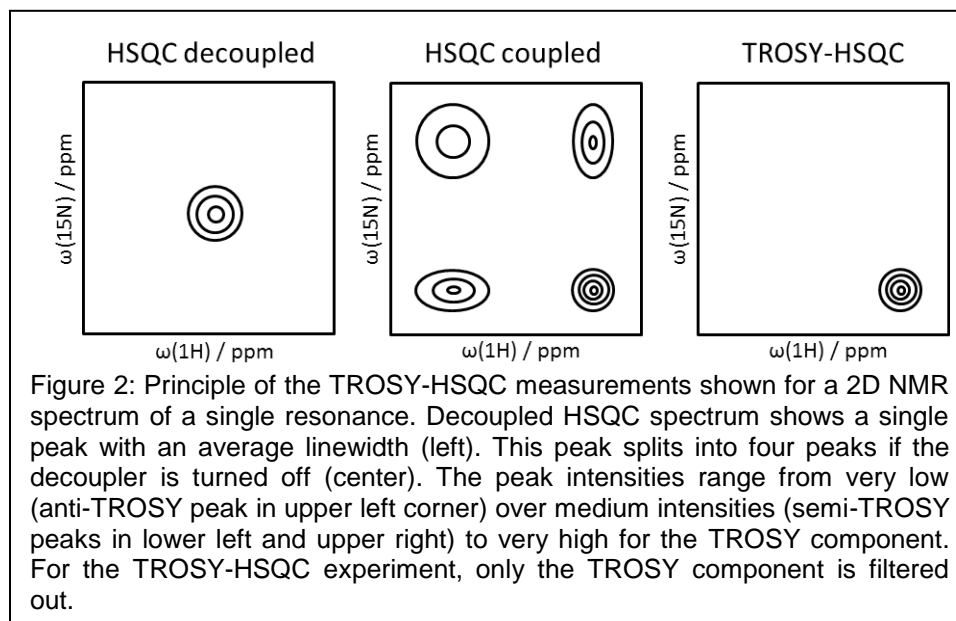
Advances in NMR spectroscopy to push the limits for protein structure elucidation

The first improvement in NMR for protein structure elucidation was the shift to higher magnetic field strengths. Field strengths have increased continually from ~30 MHz in the 1950's to ~400 MHz in the 1980's to 1 GHz in 2009. The increasing field strength led to a larger spectral dispersion resulting in less signal overlap, a smaller line-width and a larger signal-to-noise ratio. Peak assignment and the measurement of structural restraints were largely facilitated and the measurement time could be dramatically shortened.

Another major improvement in biomolecular NMR was the development of the Transverse Relaxation Optimized Spectroscopy (TROSY) in 1997 [16] enabled by the availability of high magnetic field strengths. The physical basis for the TROSY technique is the existence of interference effects between the transverse relaxation caused by the dipole-dipole (DD) interaction and the chemical shift anisotropy (CSA). These two relaxation effects can either interfere constructively – leading to an increased T_1

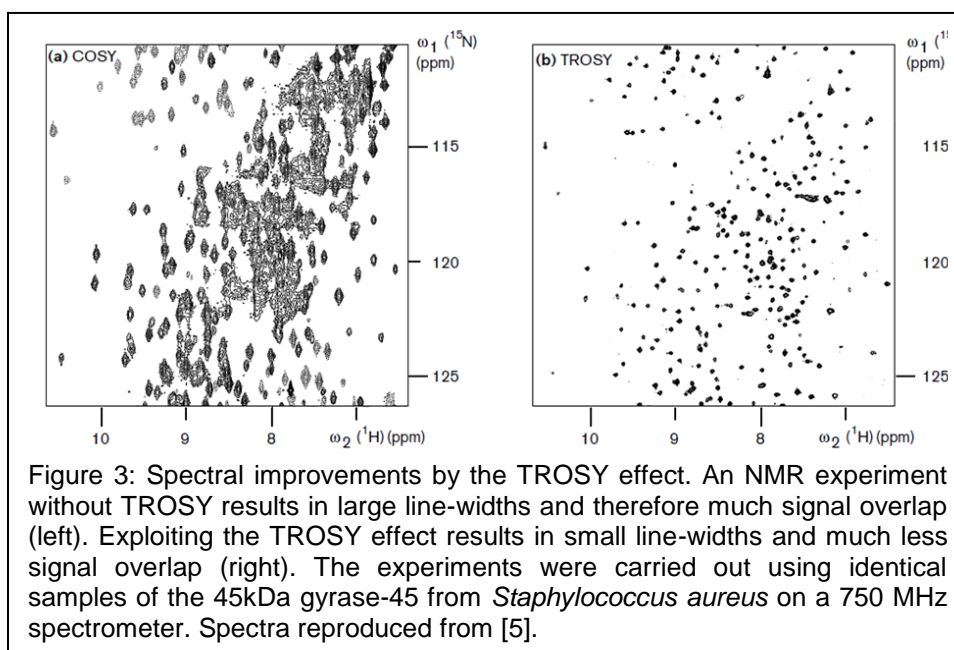
relaxation rate, an increased line-width, and a decreased signal-to-noise ratio – or the interference can be destructive where the components partially cancel each other leading to a decreased relaxation rate, a decreased line-width, and an increased signal-to-noise ratio. This results in less peak overlap, therefore higher spectral quality and a larger number of peak assignments and restraints measurable for the amide ^1H - ^{15}N correlations. The TROSY phenomenon is less effective at lower field strengths and it is most effective at ~ 1.1 GHz.

The typical pulse-sequence used for the measurement of the TROSY effect is analogous to a Hetero-nuclear Single Quantum Coherence (HSQC) experiment. There, the resonances are split by the J-coupling in both the ^1H and the ^{15}N -dimension such that for each residue four peaks appear in the spectrum (see Figure 2 and Figure 3). The pulse-sequence using the TROSY effect is modified to select only the resonance with the smallest line-width, i.e. select only the sharpest component of the four peaks.



Another major improvement in biomolecular NMR was the development of selective labeling techniques. These methods are based on the fact that only nuclei with a non-zero spin quantum number S are observable in a spectrum. There are many

naturally highly abundant isotopes with a nuclear spin of zero and these species are not observable with NMR spectroscopy. In these cases, isotopes with a non-zero spin quantum number have to be introduced into the protein which is accomplished by over-expressing the cells in ^{15}N and/or ^{13}C -enriched media. If only a small fraction of the residues is labeled with an observable isotope, less peaks will be seen in the spectrum leading to less peak overlap and simpler peak assignment. One way to achieve this is segmental labeling, where only a part of the sequence expressed in enriched media, the other parts are expressed in unlabeled media and the parts are ligated together using a series of chemical reactions. This technique was demonstrated on the 41 kDa maltose binding protein [17].



Labeling can also be achieved in a stereo-specific manner, as was shown by the SAIL technique (stereo-array isotope labeling) [18]. Amino acids are chemically and enzymatically synthesized such that one or more of the side-chain ^1H are replaced by ^2H . Additionally, Val, Leu and aromatic residues like Phe, Tyr, and Trp have some of their ^{12}C -atoms substituted by ^{13}C . Using these stereo-specific isotope labels in conjunction with cell-free expression reduces peak overlap and leads to a small line-

width since the deuterons contribute less to diamagnetic dipole-dipole (DD) relaxation. This technique was applied to the 17 kDa calmodulin and the 41 kDa maltodextrin binding protein [18]. The disadvantage of this method is the very high cost associated with obtaining the stereo-specific isotope labeled amino acids and the expertise required for cell-free expression. These may be the reasons why this technique is not commonly applied.

Sample perdeuteration is another approach that improves spectral quality for membrane proteins or large biomolecular complexes. The diamagnetic DD transverse relaxation rate T_2 depends on the square of the gyromagnetic ratio's of both the spin of interest as well as the surrounding spins. Since the gyromagnetic ratio of protons is about 6.5 as large as the one for deuterons sample perdeuteration can greatly facilitate the investigation of larger proteins by decreasing the line-broadening effects originating from nearby protons [19]. This technique is commonly applied but has the disadvantage that NH deuterons that are buried in the interior of the protein do not easily exchange with water protons and may consequently not be observed in the spectrum. Unfolding and refolding of the protein might alleviate this problem but may, in some cases, not completely overcome it [19].

The improvements described in the last sections have been absolutely essential to determine larger membrane protein structures. Nevertheless, such an effort can take many years to complete, as is exemplified by the structure determination of Diacylglycerol kinase (DAGK) [20]. Being a homo-trimeric protein of 121 residues per subunit and nine α -helical membrane spanning regions, it is to date one of the largest NMR structure of an α -helical membrane protein determined so far. Even with the advances described above structure determination of this enzyme took over 13 years to complete. This is the motivation for the development of new techniques or NMR restraints that can be used in structural studies.

Paramagnetic NMR as an avenue in protein structures elucidation

In the past two decades, paramagnetic restraints have become popular to facilitate protein structure determination, especially where NOE restraints were sparse or absent. A paramagnetic center in a protein leads to an interaction of the unpaired electron with the nuclear spins of the protein. This results in distance- and orientation-dependent effects that can be exploited as structural restraints. The three practically most often utilized phenomena are paramagnetic relaxation enhancements (PREs, i.e. contributions to the relaxation rate), Pseudo-Contact Shifts (PCSs, i.e. contributions to the chemical shift), and Residual Dipolar Couplings (RDCs), but also less prominent effects, such as cross-correlated relaxation (CCR) effects have been used.

Early work included Girvin & Fillingame's determination of local structure of the two TM spanning helical protein F_1F_0 ATP synthase [21]. The authors used chemical shifts, NOEs and PREs from PROXYL-labeled protein where the PREs were measured using spin label difference NMR spectroscopy. Later, Blackledge and co-workers have determined the structure of cytochrome c' only on the basis of paramagnetic restraints (PREs, RDCs, PCSs, and Curie-DD CCR), secondary structure, and without the use of NOEs [22]. They started from a random backbone structure and obtained a backbone RMSD of 0.7 Å for 82 of 129 residues. In another example, Gaponenko et al. have calculated the structure of the 110-residue protein barnase solely based on PREs from two different mutants to 2.9 Å compared to the crystal structure [23]. Paramagnetic restraints have also been used for the refinement of protein structures, as was shown for instance for calbindin D_{9k} [24-26], cytochrome c [27], the N-terminal domain of arginine repressor [28], and the 30 kDa N-terminal domain of STAT₄ [29].

The introduction of a paramagnetic group into the protein matrix can be achieved by substitution of the metal ion in metalloproteins (which make up to about 25% of the proteins in living organisms [30]). This was achieved as early as 1980 by Lee & Sykes

who substituted calcium in carp parvalbumin by ytterbium and lutetium to determine the magnetic susceptibility tensor from 1D proton PCSs. Alternatively, attachment of metal-binding peptides or small molecule tags coordinating a paramagnetic metal ion allows incorporation of paramagnetic metal ions into the protein. Metal ions suitable for the measurement of paramagnetic restraints are those from the transition or lanthanide series where each of the metal ions offers different characteristics.

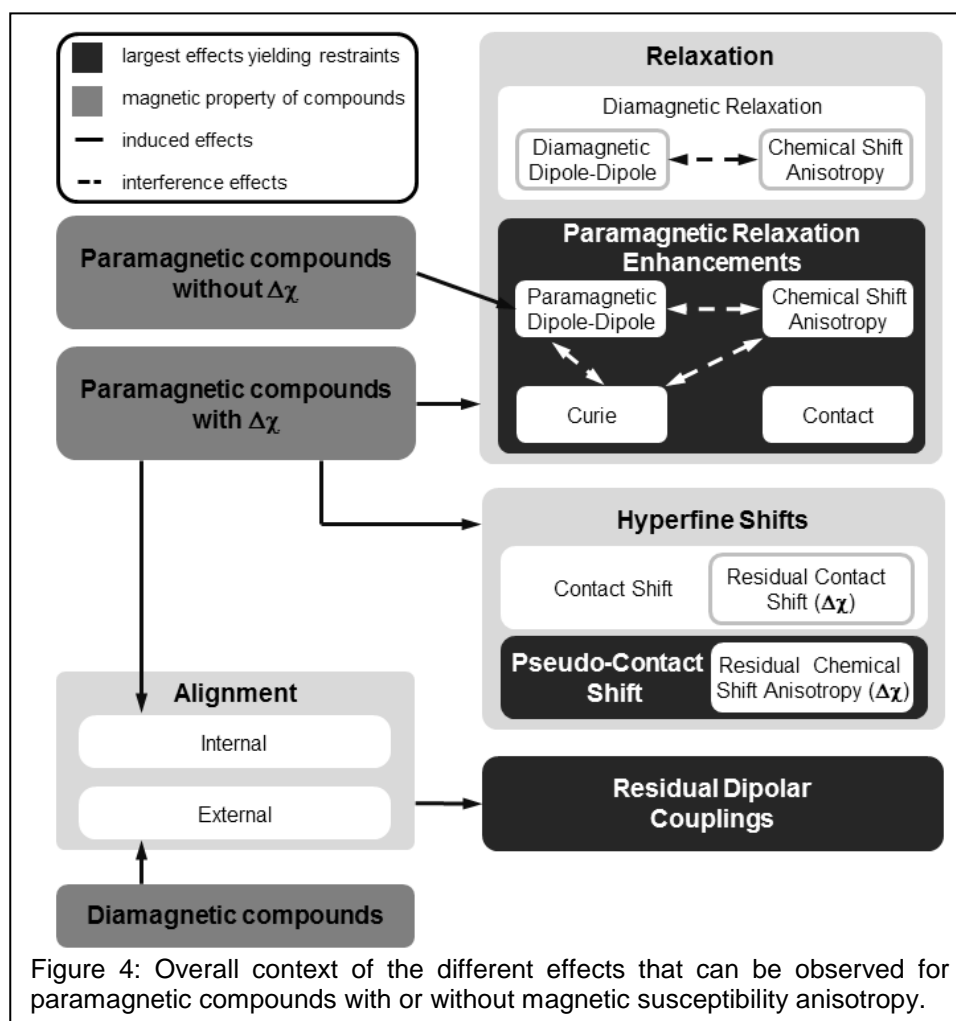
Paramagnetic effects depend on magnetic susceptibility anisotropy

Figure 4 shows a scheme that puts the paramagnetic restraints in context to relaxation and alignment. Which of these restraints are measurable depends primarily on the presence of magnetic susceptibility anisotropy (MSA), a deviation of the magnetic susceptibility tensor from isotropy. All paramagnetic species exhibit dipolar PREs: in compounds with (nearly) isotropic magnetic susceptibility, such as nitroxide spin-labels (methane-thio-sulfonate (MTSL) for instance), Gd, Mn, Cu, doxyl-stearic acid (DSA) the dipolar interactions of the unpaired electron with the nuclei of the protein result in distance-dependent line-broadening. The efficiency of the line-broadening depends on the magnetic properties of the metal ion.

If the paramagnetic center possesses MSA, as is the case for lanthanide ions (except for Gd, and the diamagnetic species Lu and La), other PRE contributions emerge that add to the dipolar PREs. The largest of these paramagnet-induced relaxation phenomena are the Curie and CSA relaxation that can interfere with each other in so called cross-correlation effects.

Additionally, MSA induces hyperfine shifts consisting of two contributions: (1) the contact shift that is only observed at very short distances around the metal ion along chemically bonded atoms, and (2) the PCS whose orientation- and distance-dependence can be exploited as structural restraints.

Furthermore, the presence of MSA will lead to partial alignment of the protein in the magnetic field – this is called internal alignment. While direct dipolar couplings between nuclei average out for isotropic tumbling of the molecule, partial alignment retains this spatial anisotropy and results in RDCs. These RDCs are a factor of ~1000 smaller than full dipolar couplings allowing their convenient determination. Experimentally, RDCs are observed as a perturbation of the J-couplings, if the nuclei are connected by a chemical bond. As RDCs depend on the mutual orientation of the internuclear vectors in the molecular frame they are useful restraints in structure determination. RDCs gained importance in conjunction with protein structure prediction in the last two decades, as they can also be measured if the protein is aligned by other



means than a paramagnetic center, for instance by using external alignment media such as bicelles [31], poly-acrylamide gels [32], or bacteriophage [33].

This concludes the introduction about NMR and paramagnetic restraints to study membrane protein structure. The following sections outline a completely different approach for membrane protein structural studies and are devoted to computational protein structure prediction.

Newer methods in protein structure elucidation: the power of computation

In 1965, the computer scientist and co-founder of Intel, Gordon E. Moore, stated that the number of transistors that can be inexpensively installed on an integrated circuit, doubles approximately every two years [34-35]. This doubling of computer power every two years has held true from 1965 to the present, 2012. The increase in computer power and the development of more efficient processors have dramatically impacted the field of bioinformatics and computational structural biology.

In the mid 1990's protein structure prediction produced more or less 'random' structures that were nowhere similar to a native protein structure and only succeeded in very rare cases for small proteins. Computer scientists that were developing novel protein structure prediction methods were completely disconnected from the experimentalists who were trying to determine protein structures. Since then, the increase in computational power and the development of advanced algorithms combined with the availability of a vast amount of experimental data (such as determined protein structures) for empirical energy potentials has led to more realistic protein models. These models can nowadays greatly enhance the understanding of particular protein structures for experimentalists, especially for proteins that are difficult to determine. This brought the work of computationalists and experimentalists closer together in a feedback loop: computer scientists build models where conventional methods are challenging,

these models inform the experimentalist to propose experiments, the experimental data can be used to refine the models, and so forth. The next paragraphs will give a brief overview of how the computational field has developed in the past decades and which methods for protein structure prediction are currently available.

Hydrophobicity scales as the beginnings of protein secondary structure prediction

Since the prediction of a protein's three-dimensional fold is a very complex problem that can hardly be tackled in its entirety, first attempts started with sequence-based methods to predict the protein secondary structure given its amino acid sequence. Early hydrophobicity scales were derived for instance from experimental measurements of partitioning energies of peptides between polar and non-polar solvents. Some of these early scales developed are the ones from Nozaki & Tanford in 1971 [36] who measured partition coefficients of amino acids, diglycine, and triglycine in water versus ethanol or dioxane solutions, Bull & Breese in 1974 [37], who computed a hydrophobicity scale from the surface tension measured on amino acids in a sodium chloride solution, Chothia in 1976 [38], whose energetic considerations were derived from solvent-accessible surface areas of a small number of proteins, and Chou & Fasman in 1978 [39] who deduced a hydrophobicity scale from amino acid propensities found in crystal structures. More of these scales are reviewed in chapter 3 which also describes the derivation of a knowledge-based hydrophobicity scale that is valid for both α -helical proteins as well as β -barrels [40]. This is the first 'unified hydrophobicity scale' described in the literature.

Hydrophobicity scales are useful in understanding which environments are preferred by particular amino acids and to describe the energetics of how proteins fold. They can also be used to predict trans-membrane spans from the protein sequence. If

hydrophobicity values are summed over a sequence window of 10-20 residues, stretches of positive hydrophobicity indicate a trans-membrane span. Furthermore, periodicities in hydrophobicity of 2 or 3.6 are a predictor for β -strands or α -helices. The knowledge of the hydrophobicity values is a first step towards the development of a trans-membrane span prediction tool, but the method has to be refined from using a simple window averaging function to a more sophisticated algorithm to achieve higher prediction accuracies, especially for β -barrel proteins. Artificial Neural Networks (or other machine learning techniques) have proven to be effective for this task and will be used here.

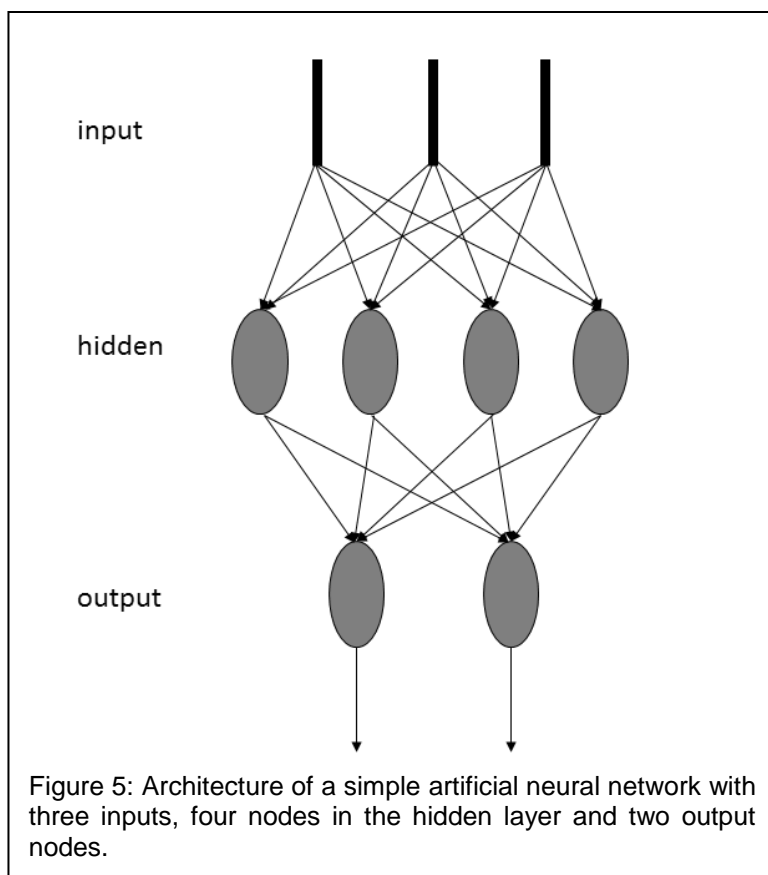
Artificial Intelligence and Machine Learning techniques as powerful tools in sequence-based protein structure prediction

Artificial Neural Networks (ANNs) were first described by the theoretical neuro-psychiatrists Warren McCulloch and Walter Pitts in 1943 who characterized the function of a net of biological neurons in a mathematical manner. The motivation for the development of ANNs was the design of a machine (or an early computer) that could make intelligent decisions based on the biological precursor, the brain. The brain, as well as an ANN consists of a number of neurons (or nodes) that are connected to each other.

In a very simple case, an ANN consists of an input layer and two layers of nodes: a hidden layer and an output layer (Figure 5). The signal is transmitted through the ANN starting from the inputs to the hidden layer and to the output. The connections between the nodes are associated with weights that store the information that the ANN contains. These weights are important for signal transmission: when one neuron receives signals from other neurons, these weighted signals are summed up to an overall signal. The neuron only transmits the overall signal to the next neuron if a certain threshold is

exceeded that is defined by the activation function, which is typically sigmoidal. This process is similar to the action potential in a biological neuron.

ANNs can be used for a wide variety of applications, where the inputs and outputs depend on the task that is to be accomplished by the ANN. The outputs are the states that should be predicted, for instance the activity of a small molecule towards a receptor. The inputs are descriptors that, in this example, characterize the state of the small molecule, namely its structure or energetic state. It is a non-trivial task to find suitable descriptors that are a useful characterization of the small molecule (or state to



be described) and that the ANN can abstract information from. The challenge lies in the fact that a large number of descriptors produce more noise in the output and make it more difficult for the ANN to decide which descriptors are important to characterize the particular problem. Therefore, it needs to be tested carefully which descriptors contain the most information that is useful for the ANN.

An ANN can be trained using unsupervised, supervised and reinforced learning. For the current application of trans-membrane span identification and secondary structure prediction, supervised learning is the method of choice. The weights in the ANN are initialized randomly and the ANN is provided with the inputs of the first dataset. The signals are passed through the network computing the value of each node moving onto the next node. At the end, the difference between the desired output and the predicted output is used to update the weights backwards using backpropagation of errors. When all the weights are updated, i.e. the ANN has 'learned' on this dataset, the next dataset will be read and processed.

Once the training of the ANN is completed, as can be judged from the performance on a monitoring dataset that is different from the training data, it can be used for prediction. The motivation for the use of ANNs is their ability to abstract or to recognize patterns in a database, meaning that it is able to make predictions for datasets that were not present in the training data. Such generalization is the goal of the network training whereas overtraining (or 'memorization') should be avoided. Overtraining can be minimized using a cross-validation procedure. In the present applications the database is divided into a training set, a monitoring set, and an independent set. The monitoring set is used for early termination to avoid overtraining, and the independent dataset is used to report the prediction accuracy. All of these sets should be completely independent of each other, which is the key to reporting an accurate predictive power.

Even though ANNs were used for the current applications, other machine learning techniques have been used for sequence-based predictions. Support Vector Machines (SVMs) are linear classifiers where the data points are projected into a hyperspace to allow for linear separation. The separation is accomplished by searching for a hyperplane that separates the two classes of data such that margins between the

hyperplane and the closest datapoints in each class (called support vectors) are maximized.

Hidden Markov Models describe a chain of unobservable (hidden) events the probability of which can only be inferred from known transition probabilities, output probabilities, and observable outputs. From these inferred event probabilities, the probabilities of future events can be computed. One example is secondary structure prediction where the prediction starts at the beginning of the sequence and moves on to the following residues one by one [41].

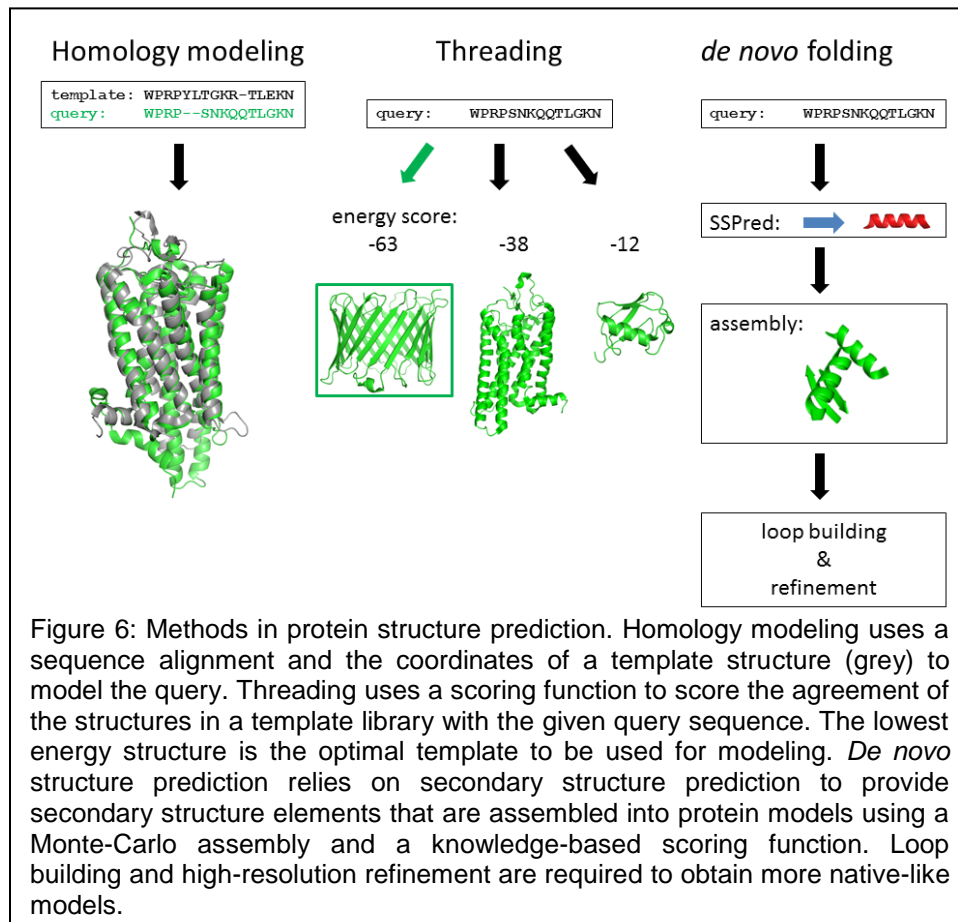
Protein structure prediction in three dimensions: challenges and available methods

In 1969, Cyrus Levinthal stated his famous paradox [42]: in a protein with 100 residues, where each residue has as little as 10 different conformations, the number of possible conformations of the protein is 10^{100} . If the protein would sample all possible conformations to find its native state and if it would only take picoseconds (10^{-12} s) to sample each of them, folding this protein would still take longer than the age of the universe. Consequently, this is clearly not how proteins fold. But the question is, how can we make protein folding *in silico* more efficient or even possible in the first place?

Homology modeling is one method with the highest 'success rate', where success is defined by building models that have the lowest root mean square deviation (RMSD) to the native protein structure. Homology modeling requires the structure of a homologue, a protein similar to the target protein, as a template. The first step in homology modeling is to create a sequence alignment of the target sequence to the template sequence. Once this alignment is optimized, the coordinates of the template structure are used to create the target structure. Remodeling of missing coordinates and loops as well as high-resolution refinement are carried out to obtain a final model.

Homology modeling works best when there is high sequence similarity between target and template sequence, as judged from the sequence alignment. Sequence similarities of ~70% can lead to models with an RMSD of 1-2 Å whereas sequence similarities of 25% can lead to highest-quality models with an RMSD of ~3-4 Å. The quality of the final model therefore depends not only on the sequence similarity but also on the choice of the template structure.

For a low sequence similarity (<25%), fold recognition [43] also known as threading [44], can be used to predict the fold of the protein. Threading uses a scoring function developed from an existing database of protein structures. The target sequence is aligned to all the template structures in the database and these models are scored using the knowledge-based scoring function. This way, the accuracy of the threading



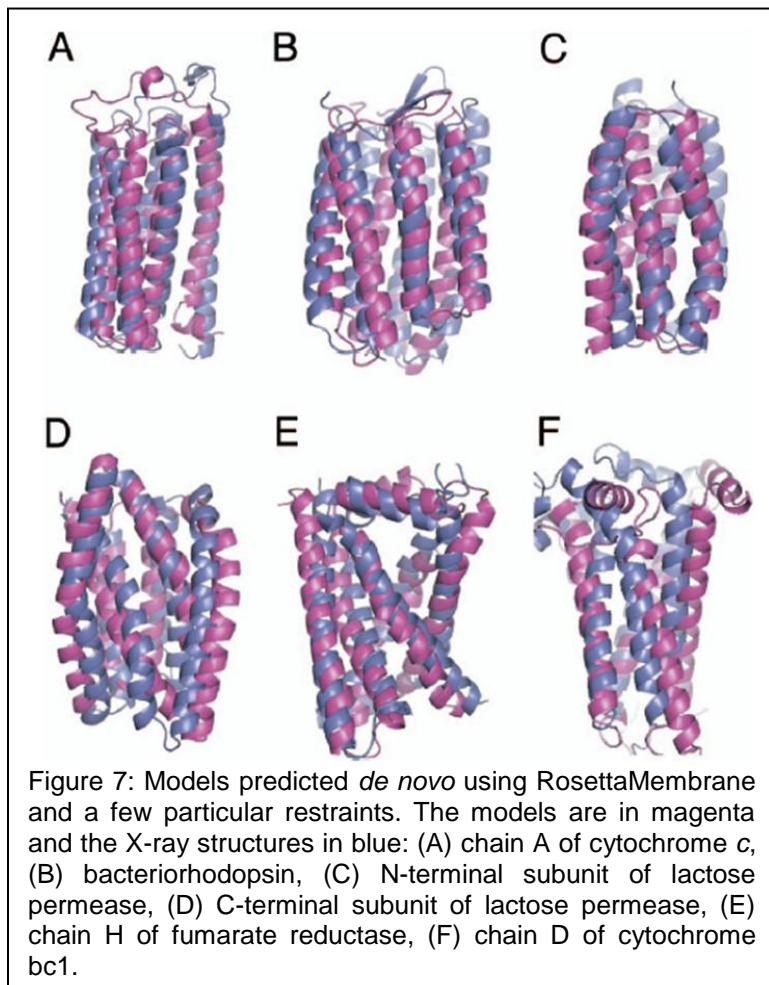
method relies on how well the scoring function is optimized. The most difficult step is the identification of a meaningful template structure, since the template cannot be identified from the sequence alignment.

Homology modeling and threading are both template-based methods (Figure 6). If no template can be identified or the protein is believed to exhibit a novel fold, *de novo* modeling can be applied where no initial assumptions about the protein structure are drawn. The gold standard method for *de novo* protein structure prediction is the software suite Rosetta [45], which is developed by the group of David Baker at the University of Washington in Seattle. Within Rosetta, the sequence of the protein is cut into 3 and 9 amino acid fragments which are subsequently assembled into three-dimensional space using a Monte-Carlo algorithm for model generation and knowledge-based energy functions for scoring. The knowledge-based scoring function utilizes statistical potentials seen in protein structures in the PDB but the scoring functions have been empirically optimized for many years [46].

It is very difficult to predict protein structures *de novo*. Rosetta may be successful at predicting small soluble proteins up to 120 residues, but encounters difficulties for larger proteins because using 3 or 9 residue fragments leads to a very large conformational search space that requires efficient sampling [47]. The incorporation of experimental data can reduce the sampling space and has been used in conjunction with Rosetta. Rosetta-CSI is able to use (even erroneous) chemical shift index data from NMR to build protein models of ~150 residues down to 2 Å RMSD to the native structure [48]. Rosetta-CSI has also been used in combination with unassigned NOESY data to build models in an automated fashion to 1 Å RMSD for similar-sized proteins [49]. Proteins up to 250 residues could be modeled *de novo* with backbone chemical shifts and NH RDCs as input to Rosetta-CSI to yield RMSDs of 1-2 Å to the native structure [50]. Membrane proteins are constrained by the membrane and therefore have a more

ordered fold compared their soluble counterparts. Consequently their structure prediction is somewhat easier to achieve. Rosetta-Membrane is able to predict structures of proteins up to 300 residues in seven trans-membrane spans using a single or very few particular restraints (see Figure 7). RMSDs to ~ 4 Å have been achieved [51].

Even though Rosetta has been successful in building protein models with the incorporation of experimental restraints, its ability for *de novo* prediction without restraints is somewhat limited due to the large conformational search space that needs to be sampled. Moreover, Rosetta is successful for modeling helical bundles, but has



difficulties for β -barrel proteins, as the long-range interactions between the strands in the sheet are difficult to capture. This is the motivation for the group around Jens Meiler, who has been trained in the laboratory of David Baker, to develop a protein folding

algorithm that overcomes the limitations of Rosetta and is able to successfully model even larger proteins *de novo* or with experimental restraints. This newly developed algorithm sets itself apart from Rosetta by reducing the conformational search space by using whole secondary structure elements (instead of 9 residue fragments) for assembly. Furthermore, the protein is simplified to initially use idealized (or 'straight') secondary structure elements represented only by the backbone and C β atoms that are assembled using a set of different 'moves'. At a refinement stage, the moves include bending the secondary structure elements which allows for the formation of different types of β -sheet topologies. The folding algorithm BCL::Fold is developed in the 'Bio-Chemical Library' (BCL) that entails a wide variety of applications for proteins and small molecules. A streamlined implementation of experimental restraints allows the usage of different types of restraints simultaneously. These restraints can be chemical shifts, NOEs, RDCs, and PREs from NMR, distance and accessibility restraints from EPR [52-53], cryo electron microscopy (Cryo-EM) density maps [54-55], and small-angle X-ray scattering data (SAXS).

Since BCL::Fold uses complete secondary structure elements for folding, very accurate secondary structure prediction is a basic requirement. Furthermore, for the identification of membrane spanning regions, a trans-membrane span prediction algorithm of high accuracy is needed. Since the apolar environment of the membrane is profoundly connected to the formation of secondary structure, our hypothesis was that a combined prediction of secondary structure and trans-membrane spans would enhance the prediction accuracy in both domains. This is the motivation for the development of a combined prediction tool described in chapter 5. Furthermore, many secondary structure prediction tools in the membrane are specialized for either α -helical proteins or β -barrels. To obtain a single prediction for a protein sequence, the output of several methods would have to be combined into one consensus prediction. This is particularly difficult to

achieve if some of the methods produce incorrect predictions that contradict each other. A combined prediction tool would alleviate this problem since it is trained on predicting both secondary structure elements simultaneously.

References

- [1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, *J Mol Biol*, 215 (1990) 403-410.
- [2] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res*, 35 (2007) D61-65.
- [3] K.D. Pruitt, T. Tatusova, W. Klimke, D.R. Maglott, NCBI Reference Sequences: current status, policy and new initiatives, *Nucleic Acids Res*, 37 (2009) D32-36.
- [4] P.W. Rose, B. Beran, C.X. Bi, W.F. Bluhm, D. Dimitropoulos, D.S. Goodsell, A. Prlic, M. Quesada, G.B. Quinn, J.D. Westbrook, J. Young, B. Yukich, C. Zardecki, H.M. Berman, P.E. Bourne, The RCSB Protein Data Bank: redesigned web site and web services, *Nucleic Acids Res*, 39 (2011) D392-D401.
- [5] G. Wider, K. Wuthrich, NMR spectroscopy of large molecules and multimolecular assemblies in solution, *Curr Opin Struct Biol*, 9 (1999) 594-601.
- [6] G.E. Tusnady, Z. Dosztanyi, I. Simon, PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank, *Nucleic Acids Res*, 33 (2005) D275-D278.
- [7] R. Shrestha, F. Berenger, K.Y.J. Zhang, Accelerating ab initio phasing with de novo models, *Acta Crystallogr D*, 67 (2011) 804-812.
- [8] L.A. Colip, A.T. Koppisch, R.D. Broene, J.A. Berger, S.M. Baldwin, M.N. Harris, L.J. Peterson, B.P. Warner, E.R. Birnbaum, A rapid method for quantifying heavy atom derivatives for multiple isomorphous replacement in protein crystallography, *J Appl Crystallogr*, 42 (2009) 329-332.
- [9] S.G. Sivakolundu, A. Nourse, S. Moshich, B. Bothner, C. Ashley, J. Satumba, J. Lahti, R.W. Kriwacki, Intrinsically Unstructured Domains of Arf and Hdm2 Form Bimolecular Oligomeric Structures In Vitro and In Vivo, *J Mol Biol*, 384 (2008) 240-254.
- [10] Y.F. Wang, I. Filippov, C. Richter, R.S. Luo, R.W. Kriwacki, Solution NMR studies of an intrinsically unstructured protein within a dilute, 75 kDa eukaryotic protein

assembly; Probing the practical limits for efficiently assigning polypeptide backbone resonances, *Chembiochem*, 6 (2005) 2242-2246.

- [11] A. Overhauser, Polarization of Nuclei in Metals, *Physical Review*, 92 (1953) 411-415.
- [12] J.R. Tolman, J.M. Flanagan, M.A. Kennedy, J.H. Prestegard, Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution, *Proc Natl Acad Sci U S A*, 92 (1995) 9279-9283.
- [13] X.C. Su, G. Otting, Paramagnetic labelling of proteins and oligonucleotides for NMR, *J Biomol NMR*, 46 (2010) 101-112.
- [14] H. Liu, L. Jin, S.B. Koh, I. Atanasov, S. Schein, L. Wu, Z.H. Zhou, Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks, *Science*, 329 (2010) 1038-1043.
- [15] S. Reckel, S. Sobhanifar, F. Durst, F. Lohr, V.A. Shirokov, V. Dotsch, F. Bernhard, Strategies for the cell-free expression of membrane proteins, *Methods Mol Biol*, 607 (2010) 187-212.
- [16] K. Pervushin, R. Riek, G. Wider, K. Wuthrich, Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution, *Proc Natl Acad Sci U S A*, 94 (1997) 12366-12371.
- [17] T. Otomo, K. Teruya, K. Uegaki, T. Yamazaki, Y. Kyogoku, Improved segmental isotope labeling of proteins and application to a larger protein, *J Biomol Nmr*, 14 (1999) 105-114.
- [18] M. Kainosho, T. Torizawa, Y. Iwashita, T. Terauchi, A.M. Ono, P. Guntert, Optimal isotope labelling for NMR protein structure determinations, *Nature*, 440 (2006) 52-57.
- [19] C.R. Sanders, F. Sonnichsen, Solution NMR of membrane proteins: practice and challenges, *Magn Reson Chem*, 44 Spec No (2006) S24-40.
- [20] W.D. Van Horn, H.J. Kim, C.D. Ellis, A. Hadziselimovic, E.S. Sulistijo, M.D. Karra, C. Tian, F.D. Sonnichsen, C.R. Sanders, Solution nuclear magnetic resonance structure of membrane-integral diacylglycerol kinase, *Science*, 324 (2009) 1726-1729.
- [21] M.E. Girvin, R.H. Fillingame, Determination of local protein structure by spin label difference 2D NMR: the region neighboring Asp61 of subunit c of the F1F0 ATP synthase, *Biochemistry*, 34 (1995) 1635-1645.

- [22] J.C. Hus, D. Marion, M. Blackledge, De novo determination of protein structure by NMR using orientational and long-range order restraints, *J Mol Biol*, 298 (2000) 927-936.
- [23] V. Gaponenko, J.W. Howarth, L. Columbus, G. Gasmi-Seabrook, J. Yuan, W.L. Hubbell, P.R. Rosevear, Protein global fold determination using site-directed spin and isotope labeling, *Protein Sci*, 9 (2000) 302-309.
- [24] M. Allegrozzi, I. Bertini, M.B.L. Janik, Y.M. Lee, G.H. Lin, C. Luchinat, Lanthanide-induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 angstrom from the metal ion, *Journal of the American Chemical Society*, 122 (2000) 4154-4161.
- [25] I. Bertini, G. Cavallaro, M. Cosenza, R. Kummerle, C. Luchinat, M. Piccioli, L. Poggi, Cross correlation rates between Curie spin and dipole-dipole relaxation in paramagnetic proteins: the case of cerium substituted calbindin D9k, *J Biomol NMR*, 23 (2002) 115-125.
- [26] I. Bertini, A. Donaire, B. Jimenez, C. Luchinat, G. Parigi, M. Piccioli, L. Poggi, Paramagnetism-based versus classical constraints: an analysis of the solution structure of Ca Ln calbindin D9k, *J Biomol NMR*, 21 (2001) 85-98.
- [27] M. Gochin, H. Roder, Protein structure refinement based on paramagnetic NMR shifts: applications to wild-type and mutant forms of cytochrome c, *Protein Sci*, 4 (1995) 296-305.
- [28] G. Pintacuda, A. Moshref, A. Leonchiks, A. Sharipo, G. Otting, Site-specific labelling with a metal chelator for protein-structure refinement, *J Biomol NMR*, 29 (2004) 351-361.
- [29] V. Gaponenko, S.P. Sarma, A.S. Altieri, D.A. Horita, J. Li, R.A. Byrd, Improving the accuracy of NMR structures of large proteins using pseudocontact shifts as long-range restraints, *J Biomol NMR*, 28 (2004) 205-212.
- [30] I. Bertini, C. Luchinat, G. Parigi, R. Pierattelli, NMR spectroscopy of paramagnetic metalloproteins, *Chembiochem*, 6 (2005) 1536-1549.
- [31] C.R. Sanders, 2nd, G.C. Landis, Reconstitution of membrane proteins into lipid-rich bilayered mixed micelles for NMR studies, *Biochemistry*, 34 (1995) 4030-4040.
- [32] H.J. Sass, G. Musco, S.J. Stahl, P.T. Wingfield, S. Grzesiek, Solution NMR of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes, *J Biomol NMR*, 18 (2000) 303-309.

- [33] M.R. Hansen, L. Mueller, A. Pardi, Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions, *Nat Struct Biol*, 5 (1998) 1065-1074.
- [34] G.E. Moore, Cramming more components onto integrated circuits, in: *Electronics Magazine*, 1965.
- [35] Excerpts from A Conversation with Gordon Moore: Moore's Law, in: Intel.
- [36] Y. Nozaki, C. Tanford, The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale, *J Biol Chem*, 246 (1971) 2211-2217.
- [37] H.B. Bull, K. Breese, Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues, *Arch Biochem Biophys*, 161 (1974) 665-670.
- [38] C. Chothia, The nature of the accessible and buried surfaces in proteins, *J Mol Biol*, 105 (1976) 1-12.
- [39] P.Y. Chou, G.D. Fasman, Prediction of the secondary structure of proteins from their amino acid sequence, *Adv Enzymol Relat Areas Mol Biol*, 47 (1978) 45-148.
- [40] J. Koehler, N. Woetzel, R. Staritzbichler, C.R. Sanders, J. Meiler, A unified hydrophobicity scale for multispan membrane proteins, *Proteins*, 76 (2009) 13-29.
- [41] K. Karplus, C. Barrett, R. Hughey, Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, 14 (1998) 846-856.
- [42] C. Levinthal, How to Fold Graciously, in: *Mossbauer Spectroscopy in Biological Systems: Proceedings 22-24*, Allerton House, Monticello, Illinois, 1969.
- [43] J.U. Bowie, R. Luthy, D. Eisenberg, A Method to Identify Protein Sequences That Fold into a Known 3-Dimensional Structure, *Science*, 253 (1991) 164-170.
- [44] D.T. Jones, W.R. Taylor, J.M. Thornton, A New Approach to Protein Fold Recognition, *Nature*, 358 (1992) 86-89.
- [45] K.T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, *J Mol Biol*, 268 (1997) 209-225.
- [46] J. Tsai, R. Bonneau, A.V. Morozov, B. Kuhlman, C.A. Rohl, D. Baker, An improved protein decoy set for testing energy functions for protein structure prediction, *Proteins*, 53 (2003) 76-87.

- [47] D.E. Kim, B. Blum, P. Bradley, D. Baker, Sampling bottlenecks in de novo protein structure prediction, *J Mol Biol*, 393 (2009) 249-260.
- [48] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J.M. Aramini, G. Liu, A. Eletsy, Y. Wu, K.K. Singarapu, A. Lemak, A. Ignatchenko, C.H. Arrowsmith, T. Szyperski, G.T. Montelione, D. Baker, A. Bax, Consistent blind protein structure generation from NMR chemical shift data, *Proc Natl Acad Sci U S A*, (2008).
- [49] S. Raman, Y.J. Huang, B. Mao, P. Rossi, J.M. Aramini, G. Liu, G.T. Montelione, D. Baker, Accurate Automated Protein NMR Structure Determination Using Unassigned NOESY Data, *Journal of the American Chemical Society*, 132 (2009) 202-207.
- [50] S. Raman, O.F. Lange, P. Rossi, M. Tyka, X. Wang, J. Aramini, G. Liu, T.A. Ramelot, A. Eletsy, T. Szyperski, M.A. Kennedy, J. Prestegard, G.T. Montelione, D. Baker, NMR structure determination for larger proteins using backbone-only data, *Science*, 327 (2010) 1014-1018.
- [51] P. Barth, B. Wallner, D. Baker, Prediction of membrane protein structures with complex topologies using limited constraints, *P Natl Acad Sci USA*, 106 (2009) 1409-1414.
- [52] K. Kazmier, N.S. Alexander, J. Meiler, H.S. McHaourab, Algorithm for selection of optimized EPR distance restraints for de novo protein structure determination, *J Struct Biol*, 173 (2011) 549-557.
- [53] N. Alexander, M. Bortolus, A. Al-Mestarihi, H. McHaourab, J. Meiler, De novo high-resolution protein structure determination from sparse spin-labeling EPR data, *Structure*, 16 (2008) 181-195.
- [54] S. Lindert, R. Staritzbichler, N. Wotzel, M. Karakas, P.L. Stewart, J. Meiler, EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps, *Structure*, 17 (2009) 990-1003.
- [55] S. Lindert, P.L. Stewart, J. Meiler, Hybrid approaches: applying computational methods in cryo-electron microscopy, *Curr Opin Struct Biol*, 19 (2009) 218-225.

CHAPTER 1

Expanding the utility of NMR restraints with paramagnetic compounds:

Background and practical aspects¹

Introduction

NMR spectroscopy is one of the most important methods for determining protein structures. The scientific community is constantly pushing the limits of NMR spectroscopy by investigating proteins of increasing sizes including membrane proteins, decreasing acquisition times by alternative sampling techniques, and automating signal assignment for high-throughput protein structure determination. Application of NMR spectroscopy to large or membrane proteins is one of the long-standing limitations as slow tumbling of the protein/membrane-mimetic complex results in line-broadening that complicates the acquisition of distance restraints based on the Nuclear Overhauser Effect (NOE) for structure elucidation. Furthermore, the spectral dispersion for alpha-helical membrane proteins is typically smaller than for beta-barrels resulting in peak overlap that complicates signal assignment. Therefore, other types of restraints are needed that complement or replace NOEs for structure elucidation. The present review focuses on a set of structural restraints that can be observed when a paramagnetic center is introduced into the protein.

This review provides a complete picture of the types of paramagnetic restraints and their origins. To maximize the practical use of this manuscript, it is emphasized which effects are usually negligible.

¹ This chapter has been published in: Koehler, J. and J. Meiler, *Expanding the utility of NMR restraints with paramagnetic compounds: Background and practical aspects*. Progress in Nuclear Magnetic Resonance Spectroscopy, 2011. **59**(4): p. 360-389.

While we attempt to review the theoretical background of the paramagnetic effects we will also outline the practical application, for instance how a paramagnetic center can be introduced into the protein. Spin-labeling methods using various nitroxide spin-labels are not discussed here as they have been reviewed elsewhere [13]. This review will also provide some practical insight on the selection of the metal ion from a structure determination standpoint.

Furthermore, we will describe a simple structure calculation protocol and review software packages available to complete particular tasks. The tensors and coordinate frames as the basis for comprehending the mathematical descriptions are explained in the Appendix.

Magnetic susceptibility and its anisotropy

To comprehend the theory behind RDCs and PCSs it is important to understand the concept of magnetic susceptibility anisotropy. Magnetic susceptibility χ is an inherent property of a substance that tells how much the substance becomes magnetized in a magnetic field or how much it interacts with a magnetic field

$$\chi = \frac{\mathbf{M}}{\mathbf{H}} \quad (1)$$

where \mathbf{M} is the magnetization and \mathbf{H} is the magnetic field strength. Magnetic Susceptibility Anisotropy (MSA) arises if the magnetization is orientation-dependent which can then be described by a second rank tensor

$$\chi = \begin{pmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{yx} & \chi_{yy} & \chi_{yz} \\ \chi_{zx} & \chi_{zy} & \chi_{zz} \end{pmatrix} \quad (2)$$

where (x, y, z) are the principal axes in a molecule-fixed coordinate system. Since the macroscopic magnetization of a sample is proportional to the sum of all microscopic electron magnetic moments μ_e the tensor elements are given by [14]

$$\chi_{aa} = \frac{\mu_0 \mu_B^2 J(J+1)}{3kT} g_{aa}^2 \quad (3)$$

where μ_0 is the permeability of vacuum, μ_B is the Bohr magneton, J is the total angular momentum quantum number, g_{aa} are the elements of the g-tensor ($a \in x, y, z$) which arises when the ratio of the electron magnetic moment and its spin quantum number becomes anisotropic (see Appendix A.1), k is Boltzmann's constant, and T is the temperature. MSA arises due to orbital contributions to the electron magnetic moment [15] where the rhombic and axial components

$$\Delta\chi_{rh} = \chi_{xx} - \chi_{yy} \quad (4a)$$

$$\Delta\chi_{ax} = \chi_{zz} - \frac{\chi_{xx} + \chi_{yy}}{2} \quad (4b)$$

are different from zero. Both equations hold true for both the principal axis frame of the tensor and the molecular frame.

The origin of Magnetic Susceptibility Anisotropy

The overall molecular susceptibility tensor is the sum of the diamagnetic and paramagnetic susceptibility tensors [16] where the diamagnetic component is usually neglected for molecules with unpaired electrons:

$$\chi_{aa}^{mol} = \chi_{aa}^{dia} + \chi_{aa}^{para}. \quad (5)$$

The paramagnetic contribution gives rise to PCSs whereas the total molecular MSA generates the overall partial alignment which is responsible for the RDCs. Note that Eq.5 refers to the overall tensors and not just the axial and rhombic parts that are responsible for the anisotropy.

As an example, these tensors have been determined from the reduced and oxidized form of cytochrome b5 using RDCs and PCSs [17].

Diamagnetic Susceptibility Anisotropy

The diamagnetic MSA is inherent in the protein through aromatic ring systems (side-chains of Phe, Tyr, Trp, and His) and peptide bonds [17]. When ring systems stack like in DNA or RNA, the diamagnetic parts of the individual MSAs are approximately additive and therefore large enough to lead to self-alignment in an applied magnetic field. In these cases the diamagnetic MSA needs to be taken into account [18], in all other cases it is very small compared to the paramagnetic contribution originating from the metal ion.

Paramagnetic Susceptibility Anisotropy

The paramagnetic MSA has two origins: low-lying excited energy states and zero-field-splitting. For low-lying excited energy states the spin-orbit coupling leads to an orbital contribution to the ground state which is orientation-dependent [14] and results in anisotropy of the g-tensor. G-anisotropy prevails for spins with $S = \frac{1}{2}$.

For spins with $S > \frac{1}{2}$ the zero-field-splitting comes into play which dominates the MSA over the g-tensor anisotropy [19]. Zero-field-splitting occurs when the electron spin density distribution can lift the degeneracy of the spin energy levels even in the absence of an external magnetic field [14].

Protein alignment and the introduction of paramagnetic metal ions

Protein alignment in the magnetic field of the spectrometer is a requirement for the measurement of RDCs and can be achieved in two different ways. The protein can be aligned externally by limiting the degrees of freedom through the confinement of the protein in its environment. In contrast, internal alignment can be achieved by exploiting the magnetic properties of the biomolecule itself or of the paramagnetic metal ion introduced into the protein. In the rare case that two different alignment media are used at the same time (external and internal – for instance a lanthanide substituted metalloprotein in a polyacrylamide gel) the magnetic susceptibility tensors are additive. Then the maximal measurable RDCs can be as large as the sum of the RDCs from the individual alignments [20]. For the present review we focus on internal alignment methods, i.e. the introduction of paramagnetic metal ions.

Advantages and disadvantages of external and internal alignment media

External alignment can be achieved by dissolving the protein in liquid-crystalline phases [21] such as rod-shaped viruses, bacteriophages [22], bicelles [2], cellulose crystallites [23], purple membrane fragments (using electrostatic interactions) [24], or by hydrated phospholipid bilayers on glass slides [25]. External alignment media are relatively robust, yield reproducible results and are tunable for instance by using compressed versus stretched gels. They are well established for measuring RDCs but they have several disadvantages: the alignment is difficult to estimate in advance [16] unless it is solely based on steric interactions where it is possible to predict from the molecular shape [26]. Furthermore, hydrophobic small ligands and membrane proteins are incompatible with many external alignment media [27].

Internal alignment produced by incorporating a paramagnetic center into the protein is not yet routinely used for structural studies. Disadvantages include that the protein of interest needs to be chemically modified to attach the paramagnetic center, which is usually a metal ion. Furthermore, the introduced metal ion induces additional line-broadening if it possesses large Curie-relaxation rates [16]. However, paramagnetic tagging has distinct advantages over external alignment media: (a) it is the only method to study protein ligand interactions with RDCs and PCSs (transferred to the ligand) because the ligand will only strongly align if bound to the partially aligned protein [27]; (b) it allows to break the symmetry degeneracy in homo-oligomeric proteins by tagging only one of the subunits [27] as was shown by Gaponenko et al. on the 28kDa dimeric protein STAT₄ [28]; (c) the alignment tensor can be tuned by using a different metal ion [29]; (d) the alignment tensor can be altered by introducing the metal ion at various positions within the protein [29] where four different placements should be sufficient to determine the structure entirely using PCSs [30]; (e) the magnetic susceptibility tensor can be cross-validated by the

measurement of both RDCs and PCSs with the knowledge of the diamagnetic tensor [16] (Eq.5); (f) inter-domain motion can be studied with paramagnetic tagging: a smaller alignment tensor of the untagged compared to the tagged domain can only originate from inter-domain motion. That means that identical alignment tensors indicate the absence of inter-domain motion for internal alignment. For external alignment media however, identical alignment tensors fixed to two separate domains of the protein do not necessarily indicate the absence of inter-domain motion [27].

When working with membrane proteins the situation becomes more difficult for both external as well as internal alignment media: the possible interaction of alignment medium with the protein and the compatibility of the alignment medium with lipids or detergents have to be tested [31].

Methods to introduce metal ions

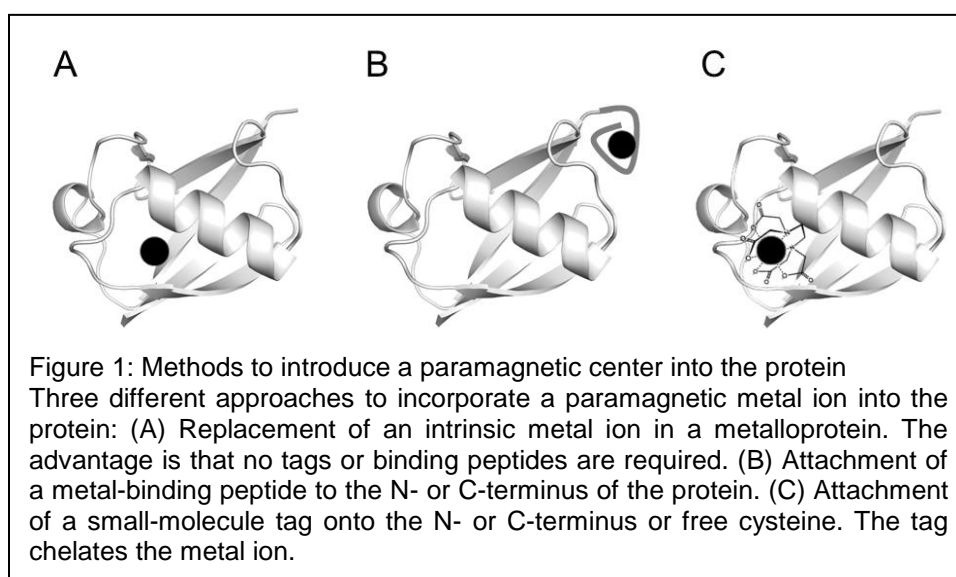
Figure 1 shows the three different options of introducing metal ions. For metalloproteins the substitution of the metal ion with a paramagnetic metal is a classical approach where the sidechains of Asp, Glu, Gln, Ser, Thr, Asn and the backbone carbonyl groups typically coordinate the metal ions [32-33].

For proteins not containing a metal-binding site the attachment of a lanthanide-binding peptide or a lanthanide tag is a viable option. Table 1 summarizes different lanthanide-binding peptides and lanthanide-binding tags used with their characteristics and measured restraints.

Lanthanide-binding peptides

Lanthanide-binding peptides can be attached at either the N- or C-terminus (which induces small PCSs because of flexibility) or at a thiol-reactive cysteine. Lanthanide-

binding peptides are designed to coordinate lanthanides [29] by interactions with the peptide side-chains. Some tags exhibit metal ion binding affinities in the μM range [33] and are in general very large in comparison to lanthanide tags: up to 17 residues [33] compared to a molecular weight of about three residues for a small molecule lanthanide tag. This is both an advantage as well as a disadvantage: the size of the lanthanide-binding peptide prevents large amplitude motions but also increases the tumbling time of the protein-tag complex.



Lanthanide-binding tags

Lanthanide-binding tags are small molecule chelating agents coordinating a metal ion. They are most commonly derived from EDTA, but DOTA or other frameworks have also been used. Ideally, the lanthanide or other paramagnetic metal ion should be rigidly attached to the protein, therefore, the length of the linker between the C_{α} atom in the protein backbone and the metal coordination site should be short. Longer linkers result in smaller RDCs and PCSs because flexibility of the tag with respect to the protein decreases the strength of the alignment and the amplitude of the alignment tensors. This also leads

to an imprecise definition of the metal position in structure calculations [11]. The effect of motion of the tag can be minimized by using bulky tags [30] such as DOTA-M8 [34].

A potential difficulty in using lanthanide-binding tags is the formation of enantiomers upon metal ion binding which leads to diastereomers when attached to the chiral protein. As a result, two slightly shifted sets of spectra are observed [11]. Using a chiral tag [35] can circumvent this problem because of their preference for a defined chirality when complexed with the metal ion [13].

Application to membrane proteins and two-point attachment

Both lanthanide-binding peptides as well as lanthanide-binding tags have been used to study membrane proteins, such as the EF-hand attached to the viral protein Vpu [31, 36] or the pyridylthio-cysteaminy-EDTA tag to study a subunit of F_1F_0 ATP synthase [31] containing two trans-membrane helices. To limit motional averaging of the peptides or the tags, a two-point attachment has been tested for both lanthanide-binding peptides and lanthanide-binding tags: Inagaki and co-workers covalently attached a 16-residue lanthanide-binding peptide to the N-terminus and a cysteine of the immunoglobulin-binding domain GB1 and measured RDCs of up to 10 Hz for Thulium at 600 MHz [37]. A DOTA-derived “caged lanthanide complex” has been attached to the 125 residue protein pseudoazurin via two thiol-reactive cysteines which are three residues apart in the sequence [38]. The observed RDCs ranged up to 6 Hz using Ytterbium at 600 MHz resonance frequency [38]. Similar RDCs (up to 6.6 Hz for Ytterbium) were observed for single-point attachment of the pyridylthio-cysteaminy-EDTA tag at the higher field strength of 800 MHz [31, 39].

Residual Dipolar Couplings (RDCs)

RDCs have first been introduced to structure elucidation in liquid state NMR spectroscopy of biological samples in 1995 when Prestegard and co-workers measured them on paramagnetic cyanometmyoglobin [40]. Since then they have evolved to one of the most important methods for obtaining structural information besides NOEs [22, 41-43].

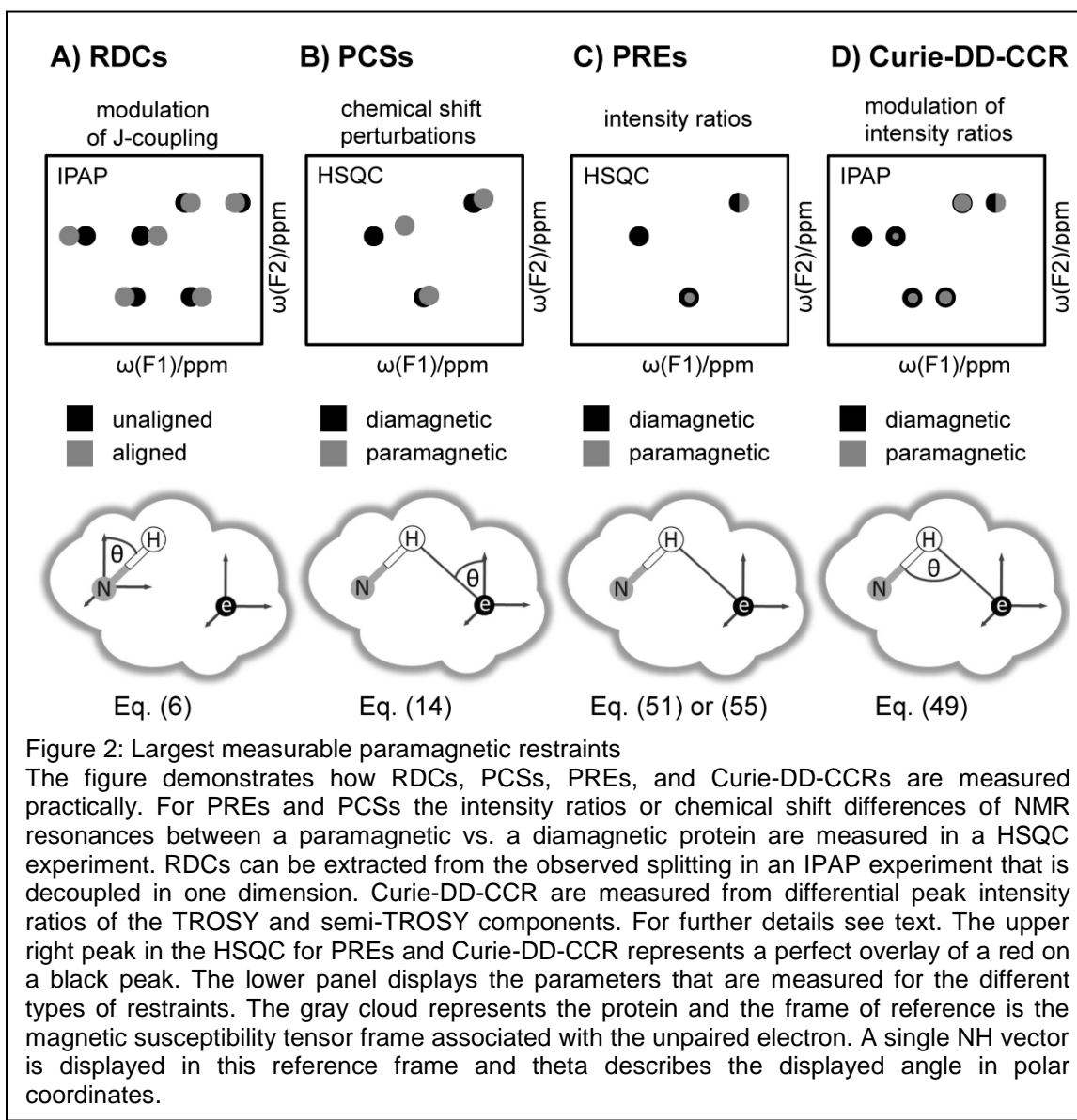
Dipolar interactions are through-space interactions between the magnetic moments of two (or more nuclear) spins. The dipolar coupling arises due to parallel or antiparallel orientation of these magnetic moments with respect to one another in an external magnetic field. If the components of the alignment tensor are zero, there is no partial alignment of the protein and therefore the protein reorients isotropically in solution. This renders the axial and rhombic components zero (see Eq.4, Eq.6 and Appendices A.1 and A.2) leading to RDCs of zero [22]. In contrast, if the proteins in a sample have a fixed orientation as in solid state NMR, these couplings are large and can be difficult to quantify, especially if numerous couplings are superimposed. In the intermediate case of a partially oriented protein, some RDCs can be determined.

The way this partial orientation or alignment is imposed is unimportant as long as the structure or dynamics of the protein are undisturbed. The measurement of RDCs does not require the introduction of a paramagnetic center into the protein since the alignment can be achieved in other ways such as external alignment. However, inversely, a paramagnetic center with anisotropic magnetic susceptibility will lead to partial alignment and will therefore yield RDCs.

RDCs for NH spins induced by MSA are described by [1]

$$D_{NH} = -\frac{B_0^2}{15kT} \cdot \frac{\gamma_H \gamma_N \hbar}{8\pi^2 r_{NH}^3} \left[\Delta\chi_{ax} (3 \cos^2 \theta_{NH} - 1) + \frac{3}{2} \Delta\chi_{rh} \sin^2 \theta_{NH} \cos 2\varphi_{NH} \right], \quad (6)$$

where B_0 is the magnetic field strength, γ_H and γ_N are the gyromagnetic ratios of the proton and nitrogen spin, $\hbar = \frac{h}{2\pi}$ is with h being Planck's constant, r_{NH} is the distance between the nitrogen and proton nuclei. As can be seen the amplitude of the RDCs



depends on the magnetic field strength, the anisotropy of the magnetic susceptibility, and the angles θ and φ that describe the polar coordinates of the NH vector in the principal frame of the molecular magnetic susceptibility tensor. RDCs are independent of the position of the metal ion. Expressing the RDCs as a function of the magnetic susceptibility tensor (and not as a function of the alignment tensor) reveals its dependence on the magnetic field strength that determines the strength of the alignment. Eq.6 is valid only when an external alignment medium is not used and if the molecular alignment originates solely from MSA. For external alignment the magnetic susceptibility components $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ should be represented by its corresponding alignment tensor components A_{ax} and A_{rh} that are related by Eq.A7. An excellent review about the derivation of Eq.6 is reference [44]. RDCs refer all internuclear vectors to the same molecule-fixed frame (Figure 2) and can therefore be considered long-range restraints [45] complementing local structural restraints such as short-range NOEs or chemical shifts.

Terms contributing to the observed splitting

Experimentally RDCs are measured in combination with J-couplings (usually -94 Hz for $^1J_{NH}$ for instance [22]) and this makes the observed splitting dependent on the magnetic field strength. The observed splitting $^1J_{NH}^{obs}(B_0)$ has the following contributions for paramagnetic ions where the largest contributions are the J-coupling and the RDCs produced by the alignment using the paramagnetic ion [15]:

$$^1J_{NH}^{obs}(B_0) = ^1J_{NH} + \Delta\nu_{RDC,dia}(B_0) + \Delta\nu_{RDC,para}(B_0) + \Delta\nu_{DFS,dia}^{CSA-DD} + \Delta\nu_{DFS,para}^{Curie-DD} \quad (7)$$

The first component on the right is the field-independent J-coupling representing the largest contribution. The terms Δv_{RDC} are the field-dependent diamagnetic and paramagnetic contributions to the RDCs, and Δv_{DFS} are the diamagnetic and paramagnetic contributions to the dynamic frequency shift which is the imaginary part of the spectral density function.

Dynamic frequency shifts are generally small

Both dynamic frequency shift contributions are perturbations of the splitting originating from cross-correlations that have corresponding relaxation effects (see below). The diamagnetic dynamic frequency shift arises due to cross-correlation between the CSA and DD interaction [15] and can be described by [46]

$$\Delta v_{DFS,dia}^{CSA-DD} \approx \frac{1}{10\pi} \left(\frac{\mu_0}{4\pi} \right) \gamma_H \gamma_N \hbar \Delta\sigma (3 \cos^2 \theta - 1) \left[\frac{1}{1 + 1/(\omega_H^2 \tau_C^2)} \right] \quad (8)$$

where θ is the angle between the symmetry axis of the assumed axially symmetric CSA tensor and the DD-interaction vector. Its corresponding relaxation contribution is responsible for the TROSY effect (see below). The paramagnetic dynamic frequency shift is due to the cross-correlation between the Curie and the DD interaction [15, 47]

$$\Delta v_{DFS,para}^{Curie-DD} = \frac{2}{10\pi} \left(\frac{\mu_0}{4\pi} \right)^2 \frac{\gamma_H^2 \gamma_N g_f^2 \mu_B^2 B_0 \hbar J(J+1)}{kT r_{MH}^3 r_{NH}^3} (3 \cos^2 \theta_{MHN} - 1) \left[\frac{\omega_H \tau_C^2}{1 + \omega_H^2 \tau_C^2} \right], \quad (9)$$

with g_J being the Landé-g-factor (see Eq.A8), r_{MH} is the distance between the metal and the proton nuclei, θ is the angle between the MH and HN vectors, ω_H is the proton Larmor frequency, and τ_C is the overall correlation time (Eq.A15). For large correlation times and high magnetic fields the approximation [15]

$$\frac{B_0 \gamma_H^2 \gamma_N \omega_H \tau_C^2}{1 + \omega_H^2 \tau_C^2} \approx -\gamma_H \gamma_N \quad (10)$$

makes the dynamic frequency shift independent of the magnetic field. Therefore, from the measurement of the observed coupling at two different magnetic fields the sum of the RDCs at these two fields is obtained. In contrast, subtracting the observed diamagnetic coupling from the observed paramagnetic coupling at the same magnetic field will yield the paramagnetic RDC and dynamic frequency shift contributions.

The dynamic frequency shift only has a measureable amplitude for correlation times close to the T_1 minimum [48]. It arises from cross-correlations between two competing relaxation pathways with similar parity [48] and has the largest influence if one of the pathways is quadrupolar relaxation. For paramagnetic molecules this effect is small [1, 17]. Dynamic frequency shifts could theoretically be exploited as restraints, however, they are too small to yield accurate information [15].

Pulse sequences for the measurement of RDCs

The most common experiment to measure RDCs is the IPAP (In-Phase-Anti-Phase) experiment [49] or, for larger complexes, the TROSY experiment [50], where the splitting is measured between the TROSY and semi-TROSY component. J-modulation

experiments have emerged which measure the RDCs based on the peak intensity ratios depending on the evolution time in the transverse plane [51]. Tugarinov and co-workers have recently introduced an experiment to measure one-bond methyl ^{13}C - ^1H and ^{13}C - ^{13}C interactions [52]. The monomeric 82 kDa enzyme malate synthase G was selectively ILV-methyl-protonated and RDCs up to 6 Hz were measured even for the ^{13}C - ^{13}C interactions. Pierattelli and co-workers introduced a ^{13}C -detected experiment to measure $^{13}\text{C}\alpha$ - $^{13}\text{C}'$, $^{13}\text{C}'$ - ^{15}N , and $^{13}\text{C}\alpha$ - $^1\text{H}\alpha$ RDCs [53].

RDCs and the influence of motion

There are two types of motion that need to be distinguished: (a) flexibility of a tag, if the paramagnetic metal ion is introduced using a peptide tag or small-molecule chelating agent; and (b) internal motion, which is the change in orientation of internuclear vectors with respect to each other. The effect of internal motion within the protein can be described by an order parameter S (not to be confused with the order tensor \mathbf{S}), which scales the observed RDCs relative to the RDCs of a rigid protein. Motion of the tag through flexible linkers reduces the amplitude of the measured RDCs because the effective order tensor is the probability weighted sum of the order tensors of the different motional states. The description of dynamics using RDCs is not the subject of this review. The reader is referred to [43, 54-56].

Chemical shift contributions

There are four contributions to the observed chemical shift when a paramagnetic center is introduced into the protein. The diamagnetic contribution δ^{dia} is always present and is the chemical shift of the nucleus in the diamagnetic protein. The binding term δ^{bind}

results from conformational changes and is a redistribution of electron density upon binding of the paramagnetic ion, inductive effects like ring-currents or direct field effects [10]. When the magnetic susceptibility of the paramagnetic ion is anisotropic, the so-called hyperfine shift or paramagnet-induced shift arises, which is the sum of two contributions, the contact shift δ^{con} and pseudo-contact shifts (PCS) δ^{PCS} [15]:

$$\delta^{obs} = \delta^{dia} + \delta^{bind} + \delta^{con} + \delta^{PCS} + \delta^{RCSA}. \quad (11)$$

The largest contributions in this equation are δ^{dia} , δ^{bind} , and δ^{PCS} if the nucleus of interest is more than 4 Å away from the paramagnetic metal ion. Contact shifts are only observed in close proximity to the paramagnetic center, their interpretation is not straightforward, and they are rarely used as restraints in structure calculations [57]. PCSs, however, are much more commonly used. To evaluate the PCSs it is necessary to separate the diamagnetic as well as the contact shifts from the observed chemical shift.

There are various ways used to determine the diamagnetic contribution: removing the metal ion, converting the metal ion into its diamagnetic form (for instance reduction of the free radical of nitroxide spin labels by ascorbic acid or other reducing agents), or coordinating a diamagnetic analog such as Ca, Zn, Lu, or La [57]. It is also possible to exploit the temperature dependence of the contact and pseudo-contact contributions, since the diamagnetic shift is ideally independent of the temperature (see below) [15].

If there are several metal binding sites in the protein and a residue is influenced by all the metals, the chemical shift contributions are additive but can have different signs

[58]. This is in contrast to the contributions to the relaxation rates which are additive but are always positive.

Contact Shifts

The contact or Fermi-contact shift arises from a through-bond interaction that connects the metal ion with the protein. Similarly to J-couplings it can provide reliable dihedral angle restraints [45] and information about the metal-ligand interaction can be inferred [59].

The contact shift arises when the spin density of the unpaired electron is distributed over the atomic orbitals of the metal ions and onto the donor atoms [15]. The spin density can be transmitted either through spin delocalization, which dominates for straight carbon chains, or through polarization, which dominates for cyclic compounds [57]. The contact shift is a very local interaction that affects only atoms closer than 4 Å from the metal for 4f electrons and 7 Å for 3d electrons in the absence of π -conjugated ligands [60]. Therefore the effect is negligible for the residues except the one that binds the metal ion [16]. When paramagnetic metal ions are present in the protein the line-broadening originating from the PREs generally masks the contact interaction for this first coordination shell. For a comprehensive discussion of all existing effects in paramagnetic NMR we include a brief discussion here.

General case

Assuming a single unpaired electron the equation for the contact shift includes the zero-field-splitting and anisotropy of the g-tensor (for definition see Appendix A.1) but requires that the spin- $\frac{1}{2}$ electron has no orbital degeneracy in the ground state [14]:

$$\delta^{con} = \frac{A \langle S_{z,lab} \rangle}{\hbar \gamma_H B_0} = \frac{A}{\hbar} \frac{1}{3 \gamma_H \mu_B \mu_0} \left(\frac{\chi_{xx}}{g_{xx}} + \frac{\chi_{yy}}{g_{yy}} + \frac{\chi_{zz}}{g_{zz}} \right). \quad (12)$$

Here, A is the hyperfine coupling constant and $\langle S_{z,lab} \rangle$ is the expectation value of the projection of the spin angular momentum onto the z-axis in the laboratory frame, which is defined as the direction of the external magnetic field. This equation assumes that the principal coordinate frames of the magnetic susceptibility tensor and the g-tensor are identical, which holds in case of paramagnetic tagging. This general and exact description makes the analysis and computation of contact shifts difficult. However, it is possible to estimate the contact shift using Karplus-type relationships [1], density-functional theory calculations, ligand field analyses, and *ab initio* procedures [16].

Simplified form

Under the assumptions of an isotropic g-tensor, high magnetic fields ($g_e \mu_B B_0 \gg A$), no zero-field-splitting and for a single unpaired electron with a large gap between the ground and the first excited state so that the spin-orbit coupling does not mix the d-orbitals [15] the McConnell equations [61-62] hold for metals except the lanthanides [14]

$$\delta^{con} = \frac{A g_e \mu_B S(S+1)}{\hbar 3kT \gamma_H} \quad (13a)$$

and for the lanthanides [14, 63]

$$\delta^{con} = \frac{A g_J (g_J - 1) \mu_B J (J + 1)}{\hbar 3kT \gamma_H}. \quad (13b)$$

The hyperfine coupling constant A is isotropic and can be calculated when the electron spin density distribution over the different nuclei is known [14, 57]. The assumptions imply that the hyperfine coupling constant A is represented by that for the ground state. According to the theory of Kurland and McGarvey [61, 64] each of the different energy levels has a different hyperfine coupling constant and in the limit of a large energy gap between ground state and the first excited state the theories of McConnell and Kurland and McGarvey coincide. Low-spin Ru(III) or Fe(III) for instance have low-lying excited states that prohibit the use of Eq.13 [61]. The contact shift is assumed to be isotropic, however, this is not generally the case, because the spin-orbit coupling causes anisotropy in S_z that only averages to zero for isotropic tumbling [61]. For anisotropic tumbling an anisotropic part of the contact shift arises which is called the residual contact shift.

Pseudo-Contact Shifts (PCS)

PCSs, also called dipolar shifts [14], arise from a through-space interaction of the unpaired electron with the nucleus (Figure 2). The dipolar magnetic field sensed by the nucleus is positive for a parallel orientation of the metal-proton vector with respect to the external magnetic field and negative if they are perpendicular [14]. In the case of no spin-orbit coupling the electron magnetic moment and therefore the magnetic susceptibility are isotropic, as is the case for a nitroxide spin-label (see below). Isotropic tumbling will then result in complete averaging over the positive and negative contributions. If, however, the spin-orbit coupling mixes the orbitals of the ground state with those from the excited

states, the magnetic moment and therefore the magnetic susceptibility become anisotropic [61]. Even under isotropic tumbling this average will not become zero [1, 15] and an additional magnetic field is induced that adds to the external one. It is assumed that the nucleus is sufficiently far away from the metal ion so that the point-dipole approximation is valid and that there is no delocalization of electron density onto the atom of interest [15].

Simplified case of isotropic reorientation

Under the assumption of isotropic tumbling of a molecule [15], the MSA is integrated over all orientations and the PCS in the principal frame of the susceptibility tensor is described by [30]

$$\delta^{PCS} = \frac{1}{12\pi r_{MH}^3} \left[\Delta\chi_{ax} (3 \cos^2 \theta_{MH} - 1) + \frac{3}{2} \Delta\chi_{rh} \sin^2 \theta_{MH} \cos 2\varphi_{MH} \right]. \quad (14)$$

If there is no MSA, both axial and rhombic anisotropy vanish which renders the PCSs zero. Even though Eq.14 is an approximation, it is typically used to extract restraints from the measured PCSs because the correction terms are small (see below). The angles θ_{MH} and φ_{MH} describe the polar coordinates of the metal-nucleus vector in the tensor frame. The PCSs depend on the distance between the nucleus of interest and the paramagnetic metal ion as $1/r^3$ and therefore have a longer range than relaxation derived parameters (such as PREs) that depend on the distance in $1/r^6$. PCSs are therefore distance- and orientational restraints that make it possible to position the metal ion into the protein frame. In the case of an axially symmetric magnetic susceptibility tensor the second term in brackets vanishes.

As seen from Eq.14 the PCSs are magnetic field independent and large PCSs are expected for metals with large MSA. In other words, different metals can be used to probe different distance ranges from the paramagnetic center. As an example, Allegrozzi et al. used calbindin with various lanthanides to measure effective distances of 5-15 Å for Ce, 9-25 Å for Yb, and 13-40Å for Dy [7].

Residual Dipolar Shift: Correction for a partially aligned protein is generally small

In the general case a term correcting for partial alignment of the protein is added to the PCSs. This correction is called residual dipolar shift and is described for axial symmetry (which cannot be assumed *a priori* [65]) in references [61, 66]. The correction term holds true under the assumption that the Zeeman energy is negligible with respect to kT because then the difference in the energy levels increases linearly with the magnetic field and makes the magnetic susceptibility field-independent. The residual dipolar shift is generally small but is expected to be measurable at magnetic fields larger than 10 T [15]. As an example, for Tb, that has the largest MSA of the metals in the lanthanide series, the correction at 800 MHz is expected to be ~0.8%.

Saturation effects are generally small

If kT is large compared to the Zeeman splitting of the electron energy levels, the population of these energy levels, although always following the Boltzmann distribution, can be approximated to be linear. When the Zeeman splitting becomes significant with respect to kT this linear approximation is not valid any longer. Therefore, the overall magnetization does not linearly increase with the magnetic field anymore (Eq.1) because the spins require a higher energy to “jump” to the excited states. This leads to a saturation effect resulting in a decrease of the magnetic susceptibility at high fields [66-67]. This

saturation term is larger and of opposite sign than the correction for anisotropic tumbling at high fields. For Tb at 800 MHz the saturation term is about 8% of the total magnetic susceptibility (for values for the lanthanides refer to [30]) and can be up to 2% of the total PCS. In case of saturation the magnetic susceptibility is described by the Brillouin equation [66-67]

$$\chi = \frac{g_J \mu_0 \mu_B}{2B_0} \left[(2J + 1) \coth \left((2J + 1) \frac{g_J \mu_B B_0}{2kT} \right) - \coth \left(\frac{g_J \mu_B B_0}{2kT} \right) \right]. \quad (15)$$

g_J is the Landé g-factor (Eq.A8). The saturation effect leads to a field dependence of the PCSs [68]. It may be stronger in the case of zero-field-splitting (as is the case for lanthanides) and it is also present but small for the contact shift [66].

Influence of motion on PCSs

PCSs are influenced by internal motion as well as flexibility of the metal ion within the protein frame as is the case for a lanthanide-binding peptide or lanthanide-binding tag. This results in a downscaling of the tensor values by the order parameter S that depends on the amplitude of the motion. For large amplitude motions also the distance dependence of the PCSs is affected. A mathematical description for structural averaging is just emerging in the literature [69].

Experimental measurement of PCSs

PCSs can be directly obtained from many different experiments because only the changes in chemical shifts need to be measured. In ^1H - ^{15}N -HSQC experiments PCSs are diagonal shifts in the spectrum, i.e. similar shifts in ppm are expected for both dimensions. This is because of the spatial proximity of the proton and the nitrogen in the backbone [70].

Whether the peaks shift upfield or downfield depends on the angle of the MH vector with respect to the MSA tensor frame, the existence of the contact shift (which is mostly neglected), changes in the g-tensor, or sign changes of the crystal field coefficients [58]. If a protein contains two or more paramagnetic centers, the PCSs are additive but can have opposite sign [71]. As structural restraints PCSs can be used to position the metal ion in the protein frame and define distance and orientation of parts of the protein with respect to the metal ion [71].

For the measurement of PCSs it is important to consider the exchange dynamics of the metal ion with the protein (or the tag, if a tag is to be used – see below). If the metal ion is “on” it causes paramagnetic effects, in the “off”-state the protein is diamagnetic. For a rapid exchange the paramagnetic and diamagnetic contributions average and the peaks can be tracked by titrations, however non-specific binding can influence the results [72]. For intermediate exchange both diamagnetic and paramagnetic contributions give peaks in the spectrum [30] which facilitates the accurate determination of the shifts and relaxation times but complicates the assignment of those peaks [72]. The temperature dependence of the PCS can then be exploited (see below) because for high temperatures the paramagnetic chemical shifts approach the diamagnetic ones [63].

Since PCSs are measurable at ranges even larger than relaxation derived restraints they are suitable for studying large proteins [30]. This was demonstrated on the 30kDa homo-dimeric STAT4_{NT} protein that was tagged with an EDTA-chelating agent with Co as shifting agent and subsequent refinement of its structure using PCSs [12].

Residual Chemical Shift Anisotropy

If PCSs are induced by a paramagnetic center that causes alignment of the protein, residual chemical shift anisotropy (RCSA) has to be taken into account. If the TROSY

sequence is used the PCSs should be measured as the difference of the midpoints between the TROSY and the semi-TROSY component because the chemical shift of the TROSY component is also perturbed by the RDCs. The difference measured is the sum of PCS and the RCSA (Supplementary Figure A3).

RCSA arises from anisotropic sampling of the chemical shifts [30] due to partial alignment of the protein. It is only significant at high magnetic fields and for nuclei with large CSA tensors. RCSA can affect the measurement of PCSs up to 0.2 ppm for ^{15}N at 800 MHz [73] which means that RCSAs can get larger than PCSs [74]. The RCSA are calculated from [74]

$$\delta^{RCSA} = \frac{B_0^2}{15\mu_0 kT} \sum_{i,j \in \{1,2,3\}} \sigma_{ii}^{CSA} \cos^2 \theta_{ij} \Delta\chi_{jj}, \quad (16)$$

where θ_{ij} are the angles of the principal axes of the MSA tensor $\Delta\chi_{jj}$ with respect to the principal axes of the CSA tensor σ_{ii}^{CSA} [73]. To account for the RCSA in the measurements of PCSs the CSA tensor has to be known [75]. The CSA tensor can be determined by solid-state NMR, *ab initio* quantum-chemical calculations, or from the cross-correlated relaxation of CSA and DD interaction [76].

RCSA are more pronounced for carbonyl/aromatic ^{13}C and amide ^{15}N spins and are negligible for protons, therefore protons are most suitable for the determination of the MSA tensor using PCSs [74]. Both PCSs and RCSA are temperature dependent [77] but in a first approximation only the RCSA depends on the magnetic field strength. RCSAs can be exploited as loose structural restraints but they possess large errors (10-20%) [77]. Since the RCSA are measured from the chemical shifts they define the relative orientation of

rigid secondary structure elements but are less effective for flexible regions of the protein [78]. Inclusion of RCSAs in structure calculations accelerates convergence [74].

Separation of contact and PCS

For correct interpretation of the hyperfine shift it is necessary to separate the contact from the PCS. One way is to consider only atoms further than 5 Å away from the metal ion where the contact shift contribution is negligible. Another way is to consider the temperature dependence of the two contributions. The temperature dependence originates from the magnetic susceptibility (see Eq.A7): for increasing temperature higher energy levels are more highly populated which leads to a more isotropic electronic distribution and therefore to smaller shifts [72]. Sm and Eu from the lanthanide series should therefore be hardly temperature dependent because they have low-lying excited states [79].

It is assumed that the diamagnetic contribution is independent of temperature, which should be fulfilled if there are no structural changes in the protein. If the logarithm of the shift is plotted vs. the logarithm of the temperature the absence of kinks in the slope indicate temperature independence [72].

The temperature dependence of the hyperfine shifts can be described as

$$\delta^{obs} - \delta^{dia} = \frac{A}{T} + \frac{B}{T^2} + \frac{C}{T^3} + \dots \quad (17)$$

where the first term describes the temperature dependence of the contact shift and the higher order terms (with the leading $1/T^2$ term) are attributed to the PCS [14]. Therefore for higher temperatures the PCSs decrease which is known as Curie-like behavior [7] and the chemical shifts approach the diamagnetic shifts.

As an example 53 of 56 resonances have been assigned within 7.5 Å of the iron in the heme group in cyanometmyoglobin. This is the region where the hyperfine shifts and the line-broadening is the strongest [80].

Structure calculations using PCS and RDCs

Eq.6 and Eq.14 show that the term in square brackets is identical for RDCs and PCSs where for RDCs the axial and rhombic MSA tensor components belong to the overall molecular MSA tensor whereas for PCSs only the MSA tensor of the metal ion is considered (see Eq.5). Both RDCs and PCSs are restraints defining the orientation of structural features in the protein with respect to one another therefore defining the fold of the protein. This interpretation is particularly powerful for protein fold determination if the structural features are relatively rigid such as the backbone of secondary structure elements [81]. Since the angular dependence and therefore the mathematical description is the same for both RDCs and PCSs, we restrict our description to the treatment of RDCs in the following paragraphs.

There are three differences, however: (a) even though the angular dependence is the same, the definition of the angles is not (see Figure 2); (b) PCSs arise from the MSA of the metal ion whereas RDCs arise from the MSA of the whole protein including the diamagnetic part (Eq.5). As discussed, if the alignment is caused exclusively by the paramagnetic metal ion, both can often be assumed identical; (c) whereas both RDCs and PCSs depend on $1/r^3$ the definition of the distance r is different. For RDCs r is the bond-length between the nuclei of interest (vibrationally averaged bonds lengths: $r(\text{NH}) = 1.041$ Å, $r(\text{C}_\alpha\text{H}_\alpha) = 1.117$ Å, $r(\text{C}'\text{N}) = 1.329$ Å, $r(\text{C}_\alpha\text{C}')=1.526$ Å [82]). Since these bond lengths are constants RDCs can be assumed to be distance-independent. For PCSs r describes the distance between the proton and the metal ion turning PCSs into distance restraints.

As a result, PCSs are richer in information but more difficult to interpret as the distance and orientation need to be determined simultaneously.

Mathematical treatment

In the molecular frame each vector ij (NH vector for RDCs for instance, and MH for PCSs) can be represented by its projections angles ε'_x , ε'_y , and ε'_z onto the coordinate axes [81] so that the RDCs $D^{ij}(\varepsilon'_x, \varepsilon'_y, \varepsilon'_z)$ can be represented as

$$D^{ij}(\varepsilon'_x, \varepsilon'_y, \varepsilon'_z) = F_{ij} \begin{pmatrix} \cos \varepsilon'_x \\ \cos \varepsilon'_y \\ \cos \varepsilon'_z \end{pmatrix}^T \begin{pmatrix} -\chi'_{yy} - \chi'_{zz} & \chi'_{xy} & \chi'_{xz} \\ \chi'_{xy} & \chi'_{yy} & \chi'_{yz} \\ \chi'_{xz} & \chi'_{yz} & \chi'_{zz} \end{pmatrix} \begin{pmatrix} \cos \varepsilon'_x \\ \cos \varepsilon'_y \\ \cos \varepsilon'_z \end{pmatrix}. \quad (18)$$

For RDCs F_{ij} is the pre-factor in Eq.6 with i and j representing the nuclei of interest

$$F_{ij} = -\frac{B_0^2}{15kT} \cdot \frac{\gamma_i \gamma_j \hbar}{8\pi^2 r_{ij}^3} \quad (19)$$

and for PCSs the pre-factor in Eq.14 being

$$F_{ij} = \frac{1}{12\pi r_{MH}^3}. \quad (20)$$

Eq.18 can be written in terms of the Saupe order matrix or the alignment tensor which are related to the MSA tensor as described in Appendix A.1. The MSA tensor in

Eq.18 is represented in the molecular frame where it has five unknown components due to its traceless property. Using a set of Euler angles α , β and γ the tensor can be rotated from the molecular frame into the principal frame. This converts the tensor into its diagonal form separating the five unknowns into an orientation of the tensor with respect to the molecule (α , β , γ) and the tensor size (χ_{ax}, χ_{rh}):

$$\begin{pmatrix} -\chi'_{yy} - \chi'_{zz} & \chi'_{xy} & \chi'_{xz} \\ \chi'_{xy} & \chi'_{yy} & \chi'_{yz} \\ \chi'_{xz} & \chi'_{yz} & \chi'_{zz} \end{pmatrix} = (R^z(\alpha)R^y(\beta)R^z(\gamma))^T \times \begin{pmatrix} -\chi_{ax} + \chi_{rh} & 0 & 0 \\ 0 & -\chi_{ax} - \chi_{rh} & 0 \\ 0 & 0 & 2\chi_{ax} \end{pmatrix} \times (R^z(\alpha)R^y(\beta)R^z(\gamma)) \quad (21)$$

The position of the metal ion (x_M, y_M, z_M) represents three additional unknowns.

For a set of RDCs (or PCSs) Eq.18 can be rewritten as a linear system of equations:

$$\begin{pmatrix} D_{exp}^{ij,1}/F_{ij} \\ \vdots \\ D_{exp}^{ij,n}/F_{ij} \end{pmatrix} = \begin{pmatrix} \cos^2 \varepsilon_y^1 - \cos^2 \varepsilon_x^1 & \cos^2 \varepsilon_z^1 - \cos^2 \varepsilon_x^1 & 2\cos \varepsilon_x^1 \cos \varepsilon_y^1 & 2\cos \varepsilon_x^1 \cos \varepsilon_z^1 & 2\cos \varepsilon_y^1 \cos \varepsilon_z^1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cos^2 \varepsilon_y^n - \cos^2 \varepsilon_x^n & \cos^2 \varepsilon_z^n - \cos^2 \varepsilon_x^n & 2\cos \varepsilon_x^n \cos \varepsilon_y^n & 2\cos \varepsilon_x^n \cos \varepsilon_z^n & 2\cos \varepsilon_y^n \cos \varepsilon_z^n \end{pmatrix} \begin{pmatrix} \chi'_{yy} \\ \chi'_{zz} \\ \chi'_{xy} \\ \chi'_{xz} \\ \chi'_{yz} \end{pmatrix} \quad (22)$$

where the left hand side are the experimentally measured RDCs between spins i and j for all datapoints 1 to n , the matrix describes the structure of the protein in the molecular frame, and the vector on the right hand side contains the five unknown elements of the MSA tensor. CSA values can be treated in a similar fashion [83].

Structure calculation protocol

An outline of a structure calculation protocol is given in Figure 3. The initial structure can either be a crystal structure, homology model, or other initial model if the restraints are used for refinement. If such a model is unavailable a random starting structure can be used and the tensor values can be approximated by an iterative procedure. Under such circumstances it can be advantageous to convert RDCs into projection angle restraints [84].

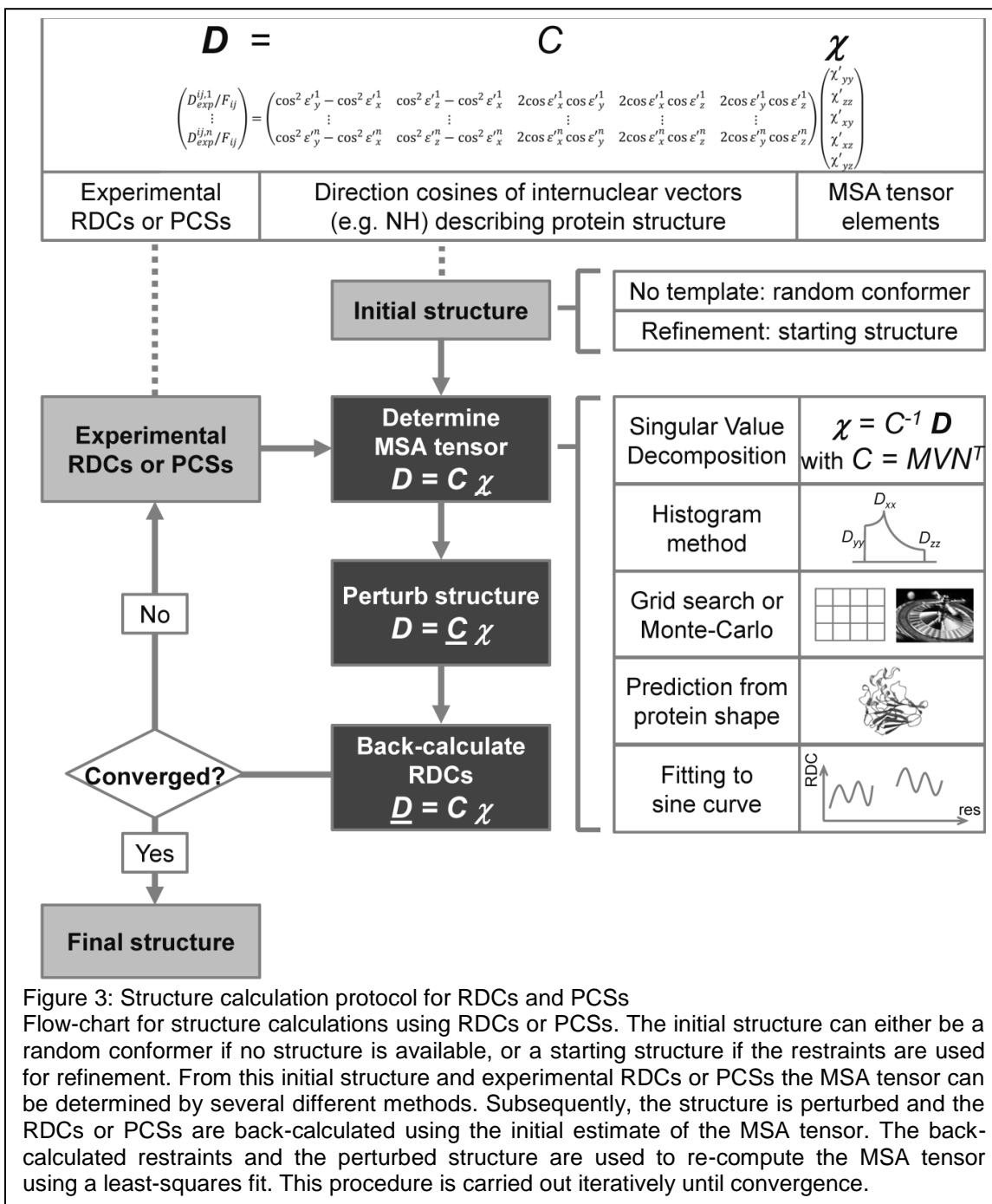


Figure 3: Structure calculation protocol for RDCs and PCSs

Flow-chart for structure calculations using RDCs or PCSs. The initial structure can either be a random conformer if no structure is available, or a starting structure if the restraints are used for refinement. From this initial structure and experimental RDCs or PCSs the MSA tensor can be determined by several different methods. Subsequently, the structure is perturbed and the RDCs or PCSs are back-calculated using the initial estimate of the MSA tensor. The back-calculated restraints and the perturbed structure are used to re-compute the MSA tensor using a least-squares fit. This procedure is carried out iteratively until convergence.

Refinement of protein structures

Earlier, RDCs and PCSs were only used for validation or refinement of protein structures [10, 85]. As an example, the inclusion of paramagnetic restraints in structure

calculations for calbindin D_{9k} led to a considerable improvement in the overall RMSD [9] from 0.69 Å to 0.25 Å. The first step in a refinement protocol is the determination of the MSA tensor from the measured RDCs and the known structure. This can be done in several ways:

(1) Eq.22 has the form $\mathbf{D} = \mathbf{C}\boldsymbol{\chi}$ where the MSA tensor $\boldsymbol{\chi}$ (in Eq.22 represented as a vector) can be determined by finding the pseudo-inverse (Moore-Penrose-Inverse) of the matrix \mathbf{C} (representing the protein structure) by Singular Value Decomposition (SVD) [83]. This approach requires a known protein structure and is very robust if the number of restraints is substantially larger than five.

(2) Given a three-dimensional structure the tensor elements can also be determined by a grid search, random search, or Monte-Carlo algorithms, which are very computation intensive and are most useful for the refinement of protein structures.

(3) In the absence of a structural model the principal values (eigenvalues) of the MSA tensor can be approximated from a histogram of the RDCs. For a uniform and isotropic distribution of internuclear vectors the shape of the histogram approximates a powder pattern where the lowest measured value depends only on χ_{yy} , the highest measured value on χ_{zz} , and the most populated value on χ_{xx} [86]. This approach requires a large number of measured values of RDCs because otherwise the estimates for the matrix elements are inaccurate. RDCs from different nuclei can be included [86] since the scaling factor F_{ij} in Eq.19 contains the nucleus-specific gyromagnetic ratios. This method will only provide the diagonal elements of the order tensor. The relative orientation to the molecular frame (Euler angles) need to be refined using an iterative least-squares optimization as described below.

(4) If the alignment mechanism is assumed to be completely steric, the alignment tensor can be predicted on the basis of the molecular shape [26]. Later, electrostatic interactions were included in the algorithms [87-89] (see *Available software*).

(5) The alignment tensor can be estimated from PISEMA spectra using an approach similar to PISA wheels [90]: plotting the RDCs over the residue number and fitting a sine curve. The tensor parameters are related to these fitting parameters. This procedure was successful for individual secondary structure elements. For helices this approach is well known since the NH vectors are almost parallel to the helix vector. For strands it is more difficult since the NH vectors are almost perpendicular to the strand vector. Then the C_αC' RDCs can be used which form an angle of ~35 degrees with the strand vector [91].

After the MSA tensor is determined the structure is changed by altering the angles ε'_x , ε'_y , and ε'_z . Then, both the new structure as well as the MSA tensor are used to recalculate the RDCs using Eq.22. Since the system of equations is over-determined there is no exact solution. The best solution can be found by method (1) using the equality

$$\begin{pmatrix} D_{exp}^{ij,1}/F_{ij} \\ \vdots \\ D_{exp}^{ij,n}/F_{ij} \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} D_{calc}^{ij,1}/F_{ij} \\ \vdots \\ D_{calc}^{ij,n}/F_{ij} \end{pmatrix} \quad (23)$$

and minimizing the square deviations. The initially estimated MSA tensor as well as the structure are iteratively refined until convergence [92].

Q-value as indicator of model quality

The difference between the experimental and the back-calculated data, i.e. the quality of a structural model, is expressed as the Q-value. It is defined as [93]

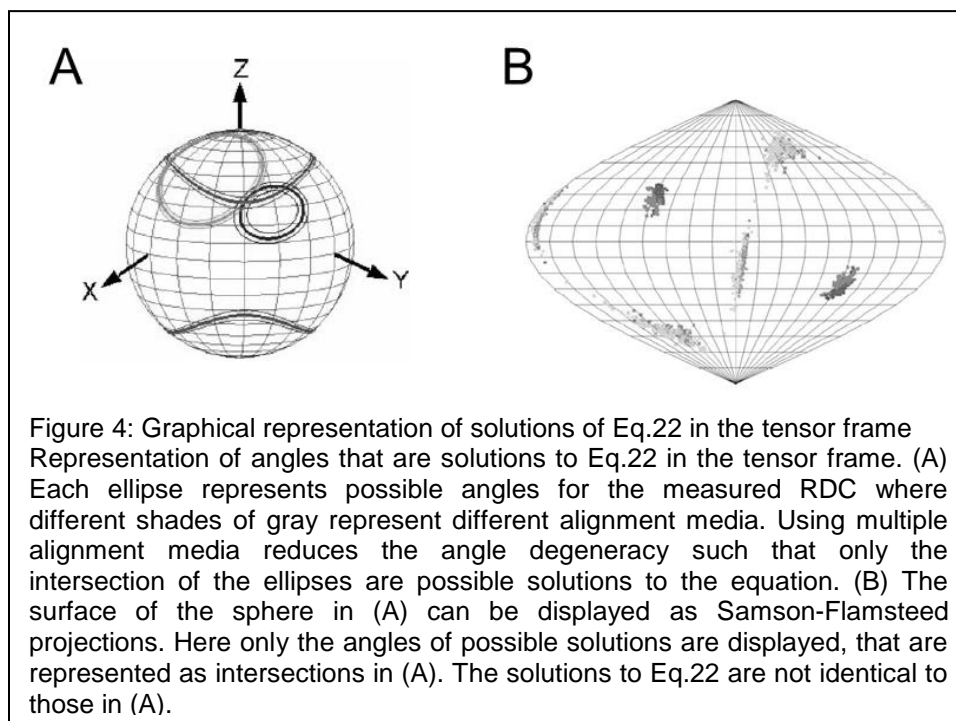
$$Q = \sqrt{\frac{\sum_n (D_{exp} - D_{pred})^2}{\sum_n (D_{exp})^2}} \quad (24)$$

where the sum is computed over the number of measured RDCs. The smaller the Q-value the better the agreement between the measured and back-calculated RDCs. Q-factors usually lie between 20% and 50% and can get as low as 10% for high-resolution crystal structures [22, 56]. Even structures refined by NMR restraints have Q-values between 10% to 15% [22]. The lower limit for the Q-value is about 10% because the ^{15}N chemical shift tensor is unknown and is variable among the residues [22]. The Q-value will not detect translational errors of structure elements since their relative orientations remain unchanged [22]. Moreover, a “bad” alignment tensor together with a “bad” set of vector orientations can still lead to a small Q-value because the distribution of internuclear vector orientations is not necessarily isotropic [94]. Therefore it is recommended to compare the principal components and the orientational components of the anisotropy (off-diagonal elements of the MSA tensor) in addition [95]. For smaller proteins the estimation of the order tensor is generally more difficult and leads to a larger error [95].

The problem of degeneracy

RDCs (and PCSs) were initially only used for refinement of protein structures only because each RDC is associated with a degeneracy of the NH vector angle in the tensor frame. An infinite number of angles satisfy Eq.22 for each coupling using a single

alignment medium. These angles can be illustrated in two graphical representations: as solutions on the surface of a sphere or as a Samson-Flamsteed projection that maps the surface area of this sphere onto a plane (see Figure 4). The degenerate angles form a



cone of solutions on the surface of a sphere in the tensor frame (Figure 4) where the inverted cone also represents possible solutions. Different alignment media (meaning that the eigenvectors in the two alignment frames are linearly independent of one another) result in different orientations of the tensor frame with respect to the molecular frame, therefore in different angles of the NH-vectors with respect to the tensor frame and consequently in different cones. As a result only the intersections of the two cones are possible solutions reducing the degeneracy to eight- or four-fold (depending on the number of intersections of the cones). RDCs in three independent alignment media yield two solutions with inverted chirality, i.e. mirror images of one another. In this case RDCs can be used from the beginning of a structure calculation protocol without the knowledge of a structure [94]. Three independent alignments could be produced by taking neutral,

positively, and negatively charged media. Using more alignment media does not further break the degeneracy but leads to higher resolution structures through the reduction of noise [96]. It should be noted that for a small diamagnetic contribution to the overall MSA tensor the RDCs and PCSs are not sufficiently complementary to break the degeneracy for the same alignment medium [45].

The relative orientation of two different domains in the protein can be calculated by determining the MSA tensors for each of the domains independently with subsequent superimposition [55]. The same approach can be applied for docking two molecules to one another [70].

Using RDCs/PCSs without the knowledge of a structure

Even though the angle degeneracy is a major obstacle in structure determination without a template, it is possible to start the structure calculation from a random initial conformer. If NOEs and J-couplings are available, RDCs can be used without any difficulties from the beginning of the structure calculation protocol. The principal components of the MSA tensor can be estimated using the histogram method, but the Euler angles are unknown. If they are guessed randomly convergence problems can occur in the iterative optimization procedure. This can be circumvented by translating the Euler angles into internuclear projection angles in the molecular frame and using allowed ranges as described by Meiler and co-workers. [84]. Alternatively, setting upper and lower limits of the tensor magnitude, aids in convergence. The best fit tensor can be filtered based on the average magnitude [94].

Habeck and co-workers used a probabilistic framework to estimate the structural coordinates, the tensor elements and the error of the RDCs [97] simultaneously. As a by-product the uncertainty of the coordinates and the alignment tensor were also computed.

RDCs can also be used in conjunction with molecular fragment replacement to determine the fold of proteins. Delaglio and co-workers have demonstrated the utility of this approach without further restraints [98].

Assignments using RDCs/PCSs

If RDCs or PCSs are used for assignment, the structure of the protein or of a homolog is required. The assignment is achieved iteratively until convergence: from some unambiguously assigned peaks (far away from the paramagnetic center where the peaks are unaffected) the tensor values are calculated by SVD [83], the structural coordinates and the order tensor are used to predict the shifts of the other peaks, with these a new order tensor is calculated, and so on [45]. Rabbit parvalbumin has been assigned using this procedure with the structure of the homologous carp protein as a starting point [99]. It should be noted that the tensor determination and the resonance assignment can only be achieved in conjunction with each other.

Using (unassigned) RDCs/PCSs for fold-recognition

Unassigned RDCs or PCSs from more than three alignment media can be used in the same way to identify the most likely fold of the protein or to calculate the fitness of a template structure with respect to the unknown protein structure, i.e. determine how well the RDCs fit to the model structure [95]. Meiler and Baker were able to quickly determine the correct fold of the fumarate sensor DcuS using un- or partially assigned RDC and NOE data [100-101]. For each of the homology models or *de novo* protein models the order tensor was calculated, the RDCs were back-calculated and the best model was identified by comparison of experimental with back-calculated RDCs. The final model had an RMSD

of 2.8 Å to the native structure. Bansal et al. have shown that this procedure is even viable using the automated protein structure prediction server ROBETTA [18].

RDCs (and also PCSs) can be used for fold recognition in the same way. The ProteinDataBank is searched for structures that fit the experimental data to identify homologous proteins that cannot be identified based on sequence similarity [81]. When a homologous protein is found, the target protein can be refined using the RDCs. Meiler et al. developed a program called DIPOCOUP for this purpose [81].

Positioning the metal-ion

In addition to contact shifts, PCSs are the only restraints that can position the metal ion in the protein frame. For an unknown metal position, the number of variables increases from five to eight. In structure calculations the metal ion with its magnetic susceptibility tensor can be represented by a pseudo-residue that is connected to the protein by linkers [92]. The linkers allow a flexible tensor position and orientation that get optimized under the influence of the restraints by minimizing the so-called target function. The target function is a potential energy term that introduces RDCs, PCSs and/or other restraints into the structure calculation procedure. When using paramagnetic restraints in structure calculations it is important that the restraints used for validation of the structure are not included in the structure calculation itself, i.e. a cross-validation is carried out. The best approach is to use an iterative process where a different subset of the paramagnetic restraints is excluded in each round [22]. It should also be noted that RDCs compete against each other in structure calculations, unlike NOEs [22].

PCSs and the order tensor can be iteratively refined for a family of conformers at the same time where the structures with the smallest target function are carried into the next round of refinement [45].

Bertini et al. studied the effect of different types of paramagnetic restraints on the structural quality of calbindin D_{9k}. The authors excluded classes of restraints from the structure calculation and reported the RMSD and the target function [9]. RDCs and PCSs turned out to be very important: when both were left out the RMSD increased considerably. In contrast, the removal of either RDCs or PCSs led to a minimal increase in RMSD. The inclusion of short-range PCSs (using ions from the first half of the lanthanide series) led to higher quality structures than structures calculated with long-range PCSs (using ions from the second half of the lanthanide series) [9]. It was also shown that it remains difficult to replace all NOEs by paramagnetic restraints.

Available software

The alignment tensor can be determined by the programs DIPOCOUP [81], FANTASIAN [71], or REDCAT [102]. From a structural model the axial and rhombic components and the three Euler angles are computed [103]. For an unknown structure the order tensor is calculated from a random initial conformer. REDCRAFT [104-105], as an extension of REDCAT, even goes one step further and computes the order tensor, the protein structure and identifies the location of internal motion *de novo*. It back-calculates RDCs from an initial two-residue fragment and compares them to experimental RDCs obtained using two different alignment media. In an iterative procedure the protein fragment is extended assuming planar peptide bond geometries and utilizing least-squares fitting of the back-calculated RDCs to the experimental RDCs until the whole protein structure is computed.

For purely steric interactions (i.e. for external alignment media and therefore only applicable for RDCs) the alignment tensor can be estimated from the molecular shape. The alignment is modeled as interactions between the molecule and flat obstacles (such

as bicelles for instance) eliminating impossible orientations caused by clashes. Appropriate software programs include PALES [26], PATI [89], and TRAMITE [106]. Recent improvements of these programs include the consideration of electrostatic interactions that are present in many alignment media [87-88].

Once the order tensor is known the structure calculation can be carried out with PSEUDYANA that is based on DYANA, or the PARArestraints module [107] of XPLOR-NIH [108]. The software is optimized for the use of PCSs [103] even from the beginning of the structure calculation process and not only for refinement. A protein structure is obtained by iterative refinement. PSEUDYANA works best with available NOEs but they are not required to achieve convergence [103].

For resonance assignments the programs ECHIDNA and PLATYPUS are available. ECHIDNA [109] is capable of automatically assigning most of the peaks in a paramagnetic HSQC from the given protein structure and the resonance assignments of the diamagnetic spectrum. It also determines the MSA tensor. PLATYPUS can be used to simultaneously compute the MSA tensor and to make automatic assignments on the basis of a known structure [110].

NUMBAT is an interactive software with a graphical user interface for the calculation of the MSA tensor from structural coordinates and PCSs [73]. The developers explicitly emphasize the improved user-friendliness compared to PSEUDYANA, GROMACS or PARArestraints within NIH-XPLOR. NUMBAT is linked to MOLMOL and PYMOL to visualize the protein structure and the order tensor, and to GNUPLOT to visualize the Samson-Flamsteed projections of the order tensor.

Relaxation

Nuclear spin relaxation leads to line-broadening in a distance-dependent manner that can be exploited as structural restraints. Relaxation can be classified into auto-

relaxation and cross-correlated relaxation (CCR) effects. Both relaxation mechanisms exist for longitudinal as well as transverse relaxation. Auto-relaxation is the relaxation of a spin under the influence of a single mechanism, whereas CCR describes the interaction of two different relaxation mechanisms that can either amplify or attenuate each other.

Generally speaking, the strength of the relaxation effect depends on the properties of the metal ion, the nuclear gyromagnetic ratio, and on the magnetic field strength [111]. Since the gyromagnetic ratio of ^{15}N is about $1/10^{\text{th}}$ of that of protons, the relaxation for nitrogens is 100 times less pronounced [14].

Origin of relaxation

Relaxation occurs when motional processes induce transitions between the $+1/2$ and $-1/2$ nuclear spin states such that thermal equilibrium of the nuclear spin states is achieved in the absence of external perturbation. The motional processes have different origins and can be divided into two parts: diamagnetic relaxation is always present and refers to the relaxation from the interaction of the nuclear spin with surrounding nuclear spins. Electron relaxation or paramagnetic relaxation contains several contributions and originates from the introduction of the unpaired electron into the protein. Electrons relax much faster than nuclei which sense the change of magnetization due to a population change of the M_s energy levels [14]. Mechanisms that contribute to electron relaxation in the solid state are interaction with phonons (lattice vibrations) and Orbach or Raman processes. For the solution state, mechanisms of relaxation include collisions with the solvent, anisotropy of the molecular susceptibility, and the spin-rotation interaction. The latter is usually very small and arises from induced magnetic moments when the electron density is misplaced after rotation of the molecule or solvent bombardment [14].

Contributions to relaxation

The relaxation rate is given by the sum of the different contributions

$$\begin{aligned} R_i &= R_i^{dia} + R_i^{para} \\ R_i^{dia} &= R_i^{dia,DD} + R_i^{dia,CSA} + R_i^{dia,CSA,DD} \\ R_i^{para} &= R_i^{contact} + R_i^{DD} + R_i^{Curie} + R_i^{Curie,DD} + R_i^{Curie,CSA} \end{aligned} \tag{25}$$

with $i \in \{\text{longitudinal} \equiv 1, \text{transverse} \equiv 2\}$

where $R_i^{contact}$ and $R_i^{Curie,CSA}$ are usually negligible for a nucleus more than 4 Å away from the paramagnetic metal ion. The index $i = 1$ represents contributions to longitudinal relaxation and $i = 2$ represents transverse relaxation contributions. In the fast motion limit $R_1 = R_2$, otherwise $R_1 < R_2$ [14]. In the literature relaxation equations are sometimes written in CGS units [19] (using centimeters, grams, and seconds as base units) where the presence of the factor $\left(\frac{\mu_0}{4\pi}\right)^2$ indicates SI units. We will use SI units throughout this review.

Diamagnetic relaxation

The diamagnetic relaxation contains three terms: the diamagnetic dipole-dipole relaxation, the relaxation originating from chemical shift anisotropy (CSA), and the cross-correlated relaxation between the DD and the CSA. The first term arises when surrounding

nuclear spins contribute to relaxation of the nuclear spin of interest. The dipolar relaxation rates [112-114]

$$R_1^{dia,DD} = \frac{2}{5} \left(\frac{\mu_0}{4\pi} \right)^2 \gamma_I^2 \gamma_S^2 \hbar^2 I(I+1) \sum \frac{1}{r_{IS}^6} \left[\frac{\tau_r}{1 + \omega_I^2 \tau_r^2} + \frac{4\tau_r}{1 + 4\omega_I^2 \tau_r^2} \right] \quad (26)$$

$$R_2^{dia,DD} = \frac{1}{5} \left(\frac{\mu_0}{4\pi} \right)^2 \gamma_I^2 \gamma_S^2 \hbar^2 I(I+1) \sum \frac{1}{r_{IS}^6} \left[3\tau_r + \frac{5\tau_r}{1 + \omega_I^2 \tau_r^2} + \frac{2\tau_r}{1 + 4\omega_I^2 \tau_r^2} \right] \quad (27)$$

depend on the weighted summed distances of the nucleus of interest (I) to the surrounding spins (S). It also depends on the nuclear spin quantum number I and the gyromagnetic ratio of the surrounding spins. The gyromagnetic ratio of protons is about 6.5 times as large as the one for deuterons. As a result perdeuteration facilitates the investigation of larger proteins by decreasing line-broadening effects from nearby protons. The diamagnetic DD relaxation is only modulated by the rotational motion of the molecule, described by a correlation time τ_r , in the absence of exchange processes.

CSA relaxation

Chemical shift anisotropy (CSA) originates from the orientation-dependence of the chemical shift, and hence changes under rotation of the molecule and induces minor variations in the magnetic field at the site of the nucleus [115]. Since the maximum measurable CSA is of the order of the isotropic chemical shift of a nucleus, the CSA of protons is negligible whereas ^{15}N , ^{13}C , and ^{31}P can have sizeable CSA.

The total chemical shielding tensor σ is a non-symmetric tensor that can be decomposed into three independent tensors: an isotropic component, a traceless symmetric component, and a traceless antisymmetric component [116-118]:

$$\sigma = \sigma^{iso} + \sigma^{sym} + \sigma^{anti} \quad (28)$$

Note the difference between a non-symmetric and an antisymmetric tensor where the antisymmetric tensor elements fulfill the condition $\sigma_{ij} = -\sigma_{ji}$ which is not a requirement for a non-symmetric tensor. The isotropic tensor can be represented by a scalar

$$\sigma_{iso} = \sigma_{avg} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ with } \sigma_{avg} = \frac{1}{3}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}). \quad (29)$$

σ_{avg} corresponds to the chemical shift seen in a spectrum (for instance from a liquid) and does not induce relaxation effects [117-119]. The symmetric component of the shielding tensor has tensor elements with $\sigma_{ij} = \sigma_{ji}$. This tensor is responsible for the CSA relaxation most often described in the literature and can be diagonalized by rotation into the shielding tensor principal coordinate system (which does not have the same orientation as the principal axes of the susceptibility tensor or related tensors described in this review). The antisymmetric tensor also induces CSA relaxation but this is almost impossible to measure because the induced effects are close to parallel to the external magnetic field. This tensor cannot be diagonalized.

The CSA relaxation rates depend on the anisotropy parameter

$$\Delta\sigma = \sigma_{zz} - \frac{\sigma_{xx} + \sigma_{yy}}{2}, \quad (30)$$

and an asymmetry parameter

$$\eta = \frac{\sigma_{yy} - \sigma_{xx}}{\sigma_{zz} - \sigma_{avg}}. \quad (31)$$

For an axially symmetric system $\sigma_{xx} = \sigma_{yy} = \sigma_{\perp}$ and $\sigma_{zz} = \sigma_{\parallel}$ such that the anisotropy parameter is

$$\Delta\sigma = \sigma_{\parallel} - \sigma_{\perp}, \quad (32)$$

and the asymmetry parameter $\eta = 0$. The terminology can be quite confusing; therefore it is important to understand the difference between the anisotropy of the symmetric tensor, axial symmetry of the symmetric tensor and the non-symmetry of the overall tensor. For a nice and comprehensibly written review refer to [117-118]. Finally, defining

$$\rho^2 = \left(\frac{\sigma_{xy} - \sigma_{yx}}{2}\right)^2 + \left(\frac{\sigma_{xz} - \sigma_{zx}}{2}\right)^2 + \left(\frac{\sigma_{yz} - \sigma_{zy}}{2}\right)^2. \quad (33)$$

for a non-zero antisymmetric tensor the relaxation rates are given by [117-118]

$$R_1^{dia,CSA} = \frac{2}{15} \gamma_S^2 B_0^2 \left[5\rho^2 \cdot \frac{\tau_{r,1}}{1 + \omega_S^2 \tau_{r,1}^2} + \Delta\sigma^2 \left(1 + \frac{\eta^2}{3} \right) \frac{\tau_{r,2}}{1 + \omega_S^2 \tau_{r,2}^2} \right] \quad (34)$$

$$R_2^{dia,CSA} = \frac{1}{45} \gamma_S^2 B_0^2 \left[15\rho^2 \cdot \frac{\tau_{r,1}}{1 + \omega_S^2 \tau_{r,1}^2} + \Delta\sigma^2 \left(1 + \frac{\eta^2}{3} \right) \left(4\tau_{r,2} + \frac{3\tau_{r,2}}{1 + \omega_S^2 \tau_{r,2}^2} \right) \right] \quad (35)$$

where $\tau_{r,1}$ and $\tau_{r,2}$ correspond to the correlation times for isotropic tumbling and small-step molecular rotation, respectively [117-118]. Eq.34 and Eq.35 simplify in the case of axial symmetry ($\eta = 0$) or for isotropic tumbling ($\tau_{r,1} = 3\tau_{r,2}$).

CSA-DD cross-correlated relaxation (CCR)

The cross-correlated relaxation between the CSA and the DD interaction results in interference effects between the two. This interference can either be constructive – where both terms add up to result in larger relaxation rates and therefore broader lines – or destructive where both terms partly cancel each other leading to smaller relaxation rates and sharper linewidths. The TROSY pulse sequence [50, 120] makes use of these interference effects by keeping only the sharpest component leading to enhanced spectral quality.

CSA-DD cross-correlated relaxation (CCR) as indicator of secondary structure

CSA-DD CCR can be used as long range restraints and are indicative of different types of secondary structure as was described by Griesinger and co-workers [121]. They used a ZQ/DQ-*ct*-HNCO (i.e. zero-quantum/double-quantum – constant-time) experiment to measure the double-quantum and single-quantum coherences of the NH and CH vectors to determine the angles between them. The relaxation interference is large in beta-

sheet structures and small in helices [76] and modulates the intensity ratios of the double-quantum coherences [121]. The relaxation rates of the four different components are described by [121]

$$\begin{aligned}
 R_{\alpha\beta}^{tot} &= R^{DD-auto} - R_i^{CSA/DD} + R_j^{CSA/DD} - R_{ij}^{CSA/DD} \\
 R_{\alpha\alpha}^{tot} &= R^{DD-auto} + R_i^{CSA/DD} + R_j^{CSA/DD} + R_{ij}^{CSA/DD} \\
 R_{\beta\beta}^{tot} &= R^{DD-auto} - R_i^{CSA/DD} - R_j^{CSA/DD} + R_{ij}^{CSA/DD} \\
 R_{\beta\alpha}^{tot} &= R^{DD-auto} + R_i^{CSA/DD} - R_j^{CSA/DD} - R_{ij}^{CSA/DD}
 \end{aligned} \tag{36}$$

where i and j denote the different internuclear vectors and α and β denote the $+1/2$ and $-1/2$ spin states. The last term represents the CSA-DD CCR between the two vectors whereas the other three components originate from auto-relaxation of a single internuclear vector [121]. The individual relaxation rates can be determined from the peak intensities by

$$R_i^{CSA/DD} = \frac{1}{4t} \cdot \ln \left(\frac{I_{\alpha\beta} \cdot I_{\beta\beta}}{I_{\alpha\alpha} \cdot I_{\beta\alpha}} \right) \tag{37a}$$

$$R_j^{CSA/DD} = \frac{1}{4t} \cdot \ln \left(\frac{I_{\beta\beta} \cdot I_{\beta\alpha}}{I_{\alpha\alpha} \cdot I_{\alpha\beta}} \right) \tag{37b}$$

$$R_{ij}^{CSA/DD} = \frac{1}{4t} \cdot \ln \left(\frac{I_{\alpha\beta} \cdot I_{\beta\alpha}}{I_{\alpha\alpha} \cdot I_{\beta\beta}} \right) \tag{37c}$$

where t is the evolution time of the double-quantum coherence. Angular restraints can be extracted by using

$$R_{ij}^{CSA/DD} = \frac{2}{5} \left(\frac{\mu_0}{4\pi} \right)^2 \frac{\gamma_H \gamma_N}{r_{NH}^3} \cdot \frac{\gamma_H \gamma_C \hbar^2}{r_{CH}^3} (3 \cos^2 \vartheta - 1) \cdot \tau_C \quad (38)$$

where ϑ is the torsion angle between the $C_\alpha H_\alpha$ bond vector of residue (i) and the NH bond vector of the following residue ($i+1$) and holds under the assumption of fast internal motion and isotropic reorientation [121]. The angle θ is related to the torsion angle ψ via a Karplus relationship as described in [121].

Contact relaxation

The contact contribution dominates for nuclei bound to the paramagnetic metal ion in a distance range up to 4 Å. The relaxation rates are given by the Bloembergen equations for contact relaxation [15, 63]

$$R_1^{contact} = \frac{2}{3} \left(\frac{A}{\hbar} \right)^2 J(J+1)(g_J - 1)^2 \left[\frac{\tau_C}{1 + \omega_S^2 \tau_C^2} \right] \quad (39)$$

$$R_2^{contact} = \frac{1}{3} \left(\frac{A}{\hbar} \right)^2 J(J+1)(g_J - 1)^2 \left[\tau_C + \frac{\tau_C}{1 + \omega_S^2 \tau_C^2} \right]. \quad (40)$$

The overall correlation time τ_C is given by [15]

$$\frac{1}{\tau_C} = \frac{1}{\tau_e} + \frac{1}{\tau_M} \quad (41)$$

where τ_e is the contribution from the electron spin and τ_M is the contribution from chemical exchange, if present.

As long as the electron spin density distribution around the metal ion is known the contact relaxation together with the contact shift can be used to determine the structure of the first coordination sphere around the metal ion [122].

Dipolar relaxation

As mentioned earlier dipole-dipole interactions are interactions of two (or more) magnetic moments through space. If the interacting dipole moments originate from two nuclear spins the Nuclear Overhauser Effect (NOE [123]) can be measured. If both spins are electron spins, then their dipolar interaction results in a Double-Electron-Electron-Resonance (DEER) signal in Electron Paramagnetic Resonance (EPR) spectroscopy. If the interaction occurs between a nuclear and an electron spin, then the resulting interaction is the one described in detail below. All of these interactions can be converted into distance restraints and the measurable distance is large for spins with large gyromagnetic ratios. NOE derived distances between two nuclear spins are typically smaller than 6 Å, electron-nucleus dipolar interactions range between 15 Å and 40 Å, and electron-electron dipolar interactions lead to distances up to 70 Å.

Electron-nucleus dipolar relaxation occurs when the electron spin density reaches further out in space and interacts with the magnetic moment of the nucleus. In this case the nucleus senses the change of the magnetic moment when the electron spin changes

between the $+\frac{1}{2}$ and $-\frac{1}{2}$ spin energy levels. Dipolar relaxation assumes that the point-dipole-approximation holds meaning that the unpaired electron is centered on the metal ion. Deviations from this approximation are assumed to be negligible further than 3-4 Å away from the metal ion [14]. The longitudinal and transverse relaxation rates are given by [14]

$$R_1^{DD} = \frac{2}{15} \left(\frac{\mu_0}{4\pi} \right)^2 \frac{\gamma_I^2 g_J^2 \mu_B^2 J(J+1)}{r_{MH}^6} \left[\frac{\tau_c}{1 + (\omega_I - \omega_S)^2 \tau_c^2} + \frac{3\tau_c}{1 + \omega_I^2 \tau_c^2} + \frac{6\tau_c}{1 + (\omega_I + \omega_S)^2 \tau_c^2} \right] \quad (42)$$

$$R_2^{DD} = \frac{1}{15} \left(\frac{\mu_0}{4\pi} \right)^2 \frac{\gamma_I^2 g_J^2 \mu_B^2 J(J+1)}{r_{MH}^6} \left[4\tau_c + \frac{\tau_c}{1 + (\omega_I - \omega_S)^2 \tau_c^2} + \frac{3\tau_c}{1 + \omega_I^2 \tau_c^2} + \frac{6\tau_c}{1 + (\omega_I + \omega_S)^2 \tau_c^2} + \frac{6\tau_c}{1 + \omega_S^2 \tau_c^2} \right]. \quad (43)$$

Since the gyromagnetic ratio of the electron (S-spin) is 658 times larger than the gyromagnetic ratio of the proton (I-spin), the terms in brackets are sometimes combined [92] using $(\omega_I - \omega_S)^2 \approx (\omega_I + \omega_S)^2 \approx \omega_S^2$. The above equations are only valid for an isotropic g-tensor (for definition see Appendix A.1.), which is not the case for Co and the lanthanides, although the g-anisotropy is generally small. For the more general case of an anisotropic g-tensor refer to [124]. The total correlation time is given by [92]

$$\frac{1}{\tau_c} = \frac{1}{\tau_e} + \frac{1}{\tau_r} + \frac{1}{\tau_M} \quad (44)$$

with τ_r being the rotational correlation time of the molecule. The electron spin correlation time τ_e has most likely the largest influence on the correlation time [15], and τ_M is the contribution from chemical exchange, if present. For isotropic magnetic susceptibility dipolar relaxation is the only mechanism contributing to PREs. If the magnetic susceptibility is anisotropic, the Curie spin relaxation is another major component.

Curie relaxation

Origin of Curie relaxation

The external magnetic field induces a magnetic moment in the electrons due to a difference in the $+1/2$ and $-1/2$ energy levels. A rotation of the molecule changes the electron's magnetic moment sensed by the nucleus and results in Curie relaxation [16] which is also called dipolar shielding anisotropy or dipolar shift anisotropy (DSA). Even though this interaction leads to negligible chemical shift changes, its contribution to the relaxation rate is significant [16]. The Curie interaction has a small effect on T_1 but a significant effect on T_2 [111]. Since the population difference of the energy levels increases with larger magnetic fields, Curie relaxation depends on the magnetic field strength [16, 63, 113]. Lower fields are more suitable for probing smaller distances while larger fields are more suited for longer distances. For instance Bertini et al. found that the best magnetic field strength for a six-coordinated Co(II) ion in a 100 kDa complex corresponds to a proton resonance frequency 60 MHz if proton signals from residues bound to Co are to be resolved [111].

Mathematical treatment

The relaxation rates are given by [16, 63]

$$R_1^{Curie} = \frac{2}{5} \left(\frac{\mu_0}{4\pi} \right)^2 \frac{\gamma_I^2 B_0^2 g_J^4 \mu_B^4 J^2 (J+1)^2}{(3kT)^2 r_{MH}^6} \left(\frac{3\tau_r}{1 + \omega_I^2 \tau_r^2} \right) \left[1 - \frac{1}{4\pi \text{Tr}(\chi)} \left(\Delta\chi_{ax} (3 \cos^2 \theta_{MH} - 1) + \frac{3}{2} \Delta\chi_{rh} \sin^2 \theta_{MH} \cos 2\varphi_{MH} \right) \right] \quad (45)$$

$$R_2^{Curie} = \frac{1}{5} \left(\frac{\mu_0}{4\pi} \right)^2 \frac{\gamma_I^2 B_0^2 g_J^4 \mu_B^4 J^2 (J+1)^2}{(3kT)^2 r_{MH}^6} \left(4\tau_r + \frac{3\tau_r}{1 + \omega_I^2 \tau_r^2} \right) \left[1 - \frac{1}{4\pi \text{Tr}(\chi)} \left(\Delta\chi_{ax} (3 \cos^2 \theta_{MH} - 1) + \frac{3}{2} \Delta\chi_{rh} \sin^2 \theta_{MH} \cos 2\varphi_{MH} \right) \right]. \quad (46)$$

For both equations the second term in square brackets describing the effect originating in anisotropic magnetic susceptibility is usually neglected. This term also contains the trace $\text{Tr}(\chi)$ of the isotropic or overall magnetic susceptibility as outlined in Appendix A1 (Eq.A4). The angles describe the polar angles of the metal-nucleus vector in the tensor principal coordinate frame.

The Curie relaxation is modulated by the rotational correlation time τ_r , and not by the overall correlation time which includes the electron spin correlation time, because it is already averaged over all electron spin states [14]. Since the relaxation rates depend on the rotational correlation time, the effect is most pronounced for large molecules or macromolecules. The advantage for large molecules is that the percentage of peaks affected by Curie relaxation is smaller [7] than for small molecules. Since the rotational correlation time is inversely proportional to the temperature, the Curie relaxation rates scale with $\sim 1/T^3$ [14].

The Curie relaxation has the same functional form as the CSA and the two terms can therefore be combined into an effective shielding anisotropy. In this approach the effective tensor is the sum of the Curie and the CSA tensor [125].

Curie-DD cross-correlated relaxation

When the Curie term is sufficiently large, a cross-correlated relaxation involving the Curie and the dipolar interaction is observed (Figure 2 and Table 2). For an isotropic tensor the transverse relaxation rate is given by [8]

$$R_2^{Curie,DD} = \frac{2}{5} \left(\frac{\mu_0}{4\pi} \right)^2 \frac{\gamma_H^2 \gamma_N g_f^2 \mu_B^2 B_0 \hbar J(J+1)}{(3kT) r_{MH}^3 r_{NH}^3} (3 \cos^2 \theta_{MHN} - 1) \left[4\tau_r + \frac{3\tau_r}{1 + \omega_I^2 \tau_r^2} \right]. \quad (47)$$

This equation holds true under the assumption of isotropic molecular motion (of the NH-vectors for example) where internal motion can be considered to a first approximation by multiplication of the CCRs with the order parameter S^2 [8]. θ is the angle between the MH and HN vectors [15]. Compared to DD autorelaxation rates, which depend on $1/r^6$, Curie-DD relaxation rates depend on $1/r^3$ making longer distances observable. Curie-DD CCRs are small and have large errors that have to be taken into account. It seems that MSA can have a noticeable effect on Curie-DD CCR if it is at least the same order of magnitude as the isotropic magnetic susceptibility [126]. This does not apply to the lanthanides but applies to cyano-metmyoglobin and might be seen on high-spin Co.

Curie-DD CCR influences the TROSY effect

The Curie-DD CCR is analogous to the CSA-DD CCR responsible for the TROSY effect as it results in differential line-broadening due to interference effects. It enhances or counteracts the TROSY effect – depending on the angle between the MH and HN vectors [126]. This can complicate the acquisition of TROSY spectra for large proteins tagged with a lanthanide ion [30].

Pulse sequences used to measure Curie-DD CCR

In principle, all pulse sequences that are used to measure the diamagnetic CSA-DD CCR can also be used to measure the paramagnetic Curie-DD CCR [8]. One complication in the measurement however, is the competing Curie relaxation. Mainly two pulse sequences have been employed: the relaxation allowed coherence transfer (RACT) experiment and the TROSY sequence. The RACT experiment requires a reference experiment to account for auto-relaxation effects (which are the dipole-dipole relaxation and CSA relaxation of a single spin) [8]. The TROSY sequence measures the CCR with a variable spin-echo delay t [127]. For large molecules with short relaxation times and quickly decaying magnetization, short pulse sequences are usually preferred. This makes the TROSY sequence more suitable than RACT because less signal is lost during the course of the pulse program until the FID can be acquired [8].

Extraction of restraints from peak intensities

The total cross-correlated relaxation rate

$$R^{CCR} = R^{CSA,DD} + R^{Curie,DD} \quad (48)$$

contains the diamagnetic CSA-DD interaction and the paramagnetic Curie-DD interaction, if present. The intensity ratios of the two doublet components α and β (TROSY and semi-TROSY components in the first dimension and TROSY component in the second dimension) are given by [126-127]

$$\frac{I_{\alpha}(t)}{I_{\alpha}(0)} = \exp[-(R^{DD} + R^{CCR})t] \text{ and } \frac{I_{\beta}(t)}{I_{\beta}(0)} = \exp[-(R^{DD} - R^{CCR})t] \quad (49)$$

where t is a variable relaxation delay. Measuring the intensity ratios for different delays t in both the diamagnetic and paramagnetic case and subtracting the diamagnetic relaxation rate from the paramagnetic one yields $R^{Curie,DD}$. Then, Eq.48 can be used to determine r_{MH} and the angle θ .

Curie-DD CCR as restraints in structure calculations

A protocol for the use of Curie-DD CCRs was implemented in DYANA [8] (PSEUDYANA module) and XPLOR-NIH [107] (PARArestraint package). It has been shown that Curie-DD CCRs are good for refining families of protein structures [8]. They improve the RMSD of the structures (especially of more disordered regions) but do not have much effect on the dihedral angles. In this respect they are complementary to RDCs which improve the dihedral angles [8].

Bertini et al. have measured Curie-DD CCR from -6.8 to 9.1 Hz on met-aquomyoglobin [128]. Using these they were able to elucidate distance ranges from 9.7 – 28.5 Å. Curie-DD CCR have also been used to refine the structure of calbindin D_{9k} where one of the two Ca ions was substituted with Ce [8].

Curie-CSA cross-correlated relaxation

Another CCR effect is the interaction between Curie and CSA relaxation. This effect has been recently described [125] and is not experimentally separable from the Curie relaxation. Similarly to the other CCR effect, the overall relaxation rate can be

increased or decreased depending on the relative orientation of the CSA and the Curie tensors. This effect is usually small but may be significant for spins with large Curie relaxation [125], i.e. for rapidly relaxing electron spins such as Ce, Fe, Yb, and Dy but not for slowly relaxing spins such as Mn, Gd, or nitroxide radicals. It is larger for T_2 with large rotational correlation times but also contributes to T_1 in rapidly tumbling molecules containing a metal ion with large magnetic susceptibilities [125].

Paramagnetic Relaxation Enhancements (PRE)

PREs, also called paramagnetic broadening effects, can be used to extract distance restraints from the peak intensity ratios when certain relaxation rate enhancements (as described above) are operative.

PREs (Figure 2) define distance spheres around the paramagnetic center. The radius of these shells depend on several parameters, such as the number of unpaired electrons, the electron spin correlation time τ_e , the rotational correlation time τ_r and the magnetic field strength [1]. PREs are determined by the size of the magnetic susceptibility tensor (not so much by its anisotropy) and are less pronounced for ^{15}N and ^{13}C spins (in contrast to ^1H) because of their lower gyromagnetic ratios. Computationally PREs can be handled similarly to NOEs because they have the same $1/r^6$ distance dependence [15].

Main contributions to PREs

Under the assumption of negligible contact relaxation (or for spins sufficiently far away from the paramagnetic center) there are basically three main contributions to PREs (see Table 2): the dipolar relaxation described in Eq.42 and Eq.43 which usually dominates for long electron spin correlation times (for example Gd, Mn, MTSL), the Curie relaxation (Eq.45 and Eq.46) which usually dominates for short electron spin correlation

times (lanthanides other than Gd) and the Curie-DD CCR (Eq.47). In contrast to paramagnet-induced chemical shift changes the relaxation rates are always positive and additive. When all paramagnetic effects are combined to R_2^{para} the total relaxation rate can be described by

$$R_2^{tot} = R_2^{dia} + R_2^{para} \quad (50)$$

where R_2^{dia} is the sum of all diamagnetic contributions.

For large complexes the Curie term dominates the PREs and contributes to T_2 approximately as much as the dipolar PREs contribute to T_1 [63]. The transverse relaxation rate is more affected by the paramagnetic center than the longitudinal relaxation rate. Therefore experiments where the magnetization is stored along the z-axis are better suited for measuring PREs [1].

PREs are derived from ratios of peak intensities or linewidths of the paramagnetic vs. the diamagnetic spectrum. If nitroxide spin labels such as MTSL are used as paramagnetic species the first measurement is taken with the oxidized spin label which is paramagnetic. Subsequently, the spin label is reduced using a reducing agent such as ascorbic acid yielding a diamagnetic species. If reduction of the spin-label is unfavorable because of interference with the protein, a parallel sample preparation of diamagnetic and paramagnetically labeled protein is an option.

Methods of converting PREs into distance restraints

Single-point measurements

The approach most widely used in conjunction with MTSL is the method described by Wagner and co-workers [129]. By considering the peak intensities the transverse PREs can be obtained by solving for R_2^{para} in

$$\frac{I^{para}}{I^{dia}} = \frac{R_2^{dia} \exp(-R_2^{para} \cdot t)}{R_2^{dia} + R_2^{para}} \quad (51)$$

where t is the total time the magnetization evolves in the transverse plane during the INEPT transfer. A value of R_2^{dia} is obtained for each residue from

$$R_2^{dia} = \pi \cdot \Delta\nu, \quad (52)$$

with $\Delta\nu$ being the linewidth of the peak at half maximum height. From R_2^{para} the distances can be obtained by adding the relaxation terms responsible for the PREs and computing r_{MH} . For more than one paramagnetic center in the protein the relaxation contributions are additive.

Two-point measurements

Another method is described by Clore and co-workers [130] where a flexible delay t is incorporated into the pulse sequence. This delay is varied and the peak intensities for

both diamagnetic and paramagnetic sample at different time-points are measured. The diamagnetic and paramagnetic peak intensities decay exponentially as [130]

$$\frac{I^{dia}(t)}{I^{dia}(0)} = \exp(-R_2^{dia}t) \quad (53)$$

$$\frac{I^{para}(t)}{I^{para}(0)} = \exp(-(R_2^{dia} + R_2^{para})t). \quad (54)$$

Taking the ratios for two time-points $t = 0$ and t and rearranging yields

$$R_2^{para} = \frac{1}{t} \cdot \ln \left[\frac{I^{dia}(t)}{I^{dia}(0)} \cdot \frac{I^{para}(0)}{I^{para}(t)} \right]. \quad (55)$$

Even though this approach is rarely used, it has several advantages. Since it uses two time-points to estimate the relaxation rate the time delay to start the subsequent experiment can be shorter than in the single-point measurement where a long time delay is important to achieve complete equilibrium. The two-point measurement does not use a Lorentzian lineshape that is assumed for use of Eq.51 and that can impede spectra analyses because Lorentzians are broad and can lead to a decrease in the number of analyzable peaks in case of partial overlap. Also it does not require scaling of the spectra to account for different sample concentrations. The errors can be estimated as described in [130]. The authors have shown that increasing the number of time-points to more than two does not increase accuracy of the estimate.

Practical considerations for the interpretation of PRE data

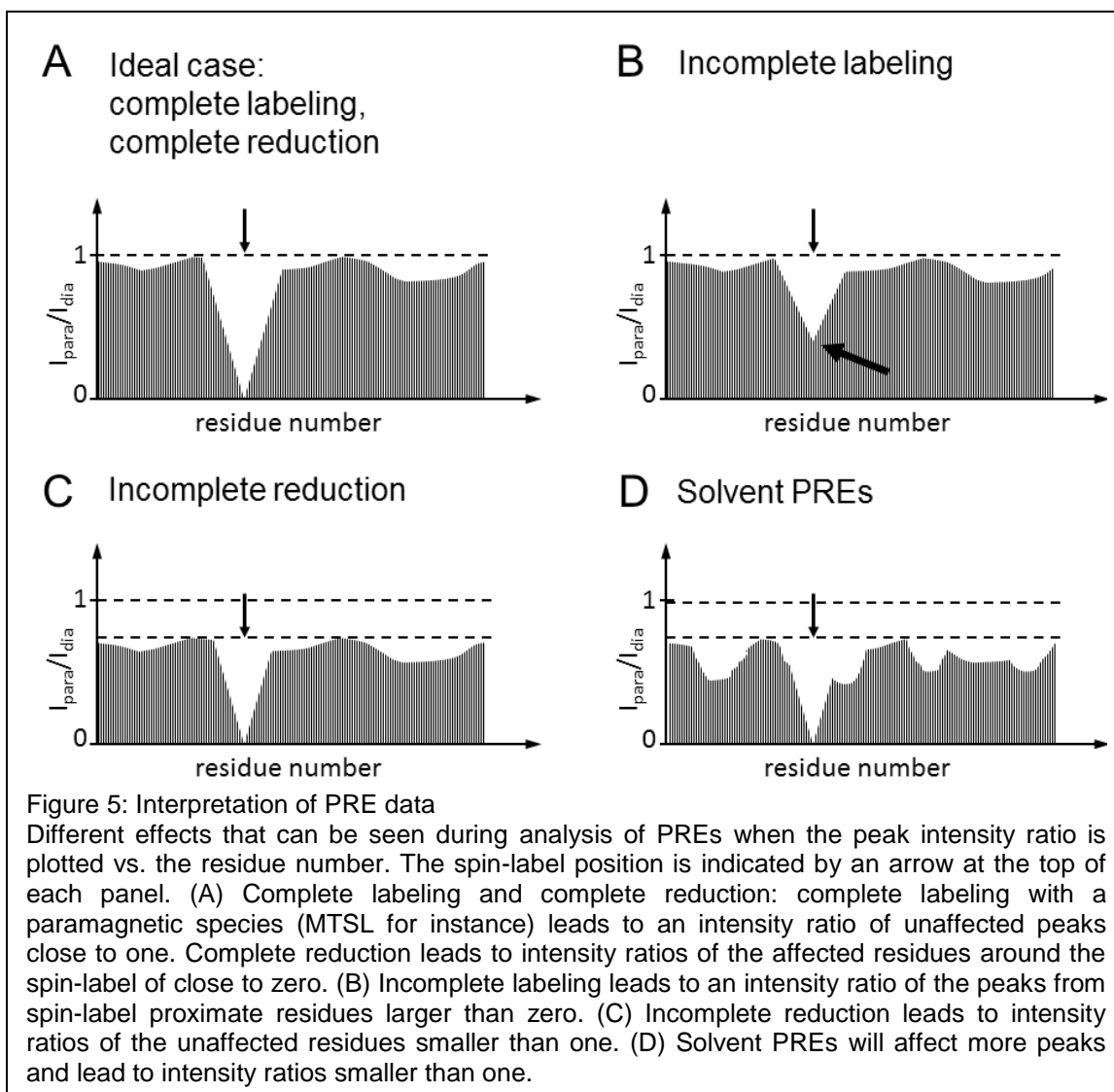
Several effects can influence the peak intensities or linewidths in the spectrum and therefore lead to incorrect distance estimates between the free electron and the nucleus of interest. Incomplete labeling of the protein with the paramagnetic species or contamination of the paramagnetic sample with the diamagnetic species lead to an additional diamagnetic contribution, therefore to underestimates of the relaxation rate yielding longer distances [130]. This can be detected as residues neighboring the spin label position will generate signals in the oxidized (paramagnetic) sample (Figure 5). For complete labeling or no contamination these peaks will be broadened beyond detection.

When nitroxide spin-labels are reduced using reducing agents such as ascorbic acid, one has to make sure that all spin-labels are reduced, since the reduction rate of nitroxides with ascorbic acid depends on the pH [131]. Full reduction is obtained when the intensity ratios of peaks from unaffected residues is close to one (Figure 5).

PREs can be dependent on the sample concentration since crowding of paramagnetic species can lead to a “solvent-PRE-effect”: the paramagnetic species of one molecule broadens the lines of residues of neighboring molecules. This can be detected as an offset of the peak intensity ratios from one (Figure 5).

PRE and the influence of motion

In case of external attachment of the paramagnetic center to the protein, for instance using a small-molecule tag, binding peptide or nitroxide spin-label, the paramagnetic ion will exhibit flexibility with respect to the protein. Clore and co-workers described the effect of fast motion of the tag for isotropic tumbling of the protein [132-133].



Examples

PREs are widely used for structure determination of proteins [6, 134]. Nitroxide spin labels (such as PROXYL [135] or MTSL [6]) have found widespread application and can elucidate distances of about 8 – 35 Å [6].

Lanthanides and other paramagnetic probes

Lanthanides (also called rare earth metals) have distinct properties that make them a desired target for use in protein structure determination [136]. Lanthanides are

chemically very similar [57] and can easily replace Ca^{2+} , Mg^{2+} , or Mn^{2+} in metalloproteins. The lanthanides La and Lu are good diamagnetic references [45].

Chemical properties of lanthanide series

Lanthanides have partially filled 4f shells that are shielded towards the exterior by the 5s and 5p orbitals [57]. This results in almost negligible contact shifts in comparison to other paramagnetic metals [99].

The lanthanides are paramagnetic except for the first and the last members in the series (La, Lu), which are diamagnetic metals. Dy, Tb, and Tm are highly paramagnetic; Er and Yb are moderately paramagnetic and Ce, Sm and Eu exhibit small paramagnetism [30].

Lanthanides and Magnetic Susceptibility Anisotropy (MSA)

Lanthanides exhibit the spin-orbit interaction leading to anisotropic magnetic susceptibility. For a non-negligible spin-orbit-interaction one has to consider the total angular momentum quantum number J (as the sum of the spin angular momentum quantum number S and the orbital angular momentum quantum number L) instead of the spin angular momentum quantum number S . For all other metals except the lanthanides, the latter is sufficient [15]. The MSA also affects the g-factor which is the Landé g-factor g_J (Eq.A8) for the lanthanides or the electron g-factor g_e for all other metals. The energy level of interest for all calculations is the ground state with the largest S , largest L , and smallest J for the first half of the series (Ce to Eu), and the largest S , largest L , and largest J for the second half of the series (Tb to Yb) [14].

Choosing lanthanides for structural studies

For structural studies lanthanides should be chosen based on their magnetic properties and their biological activity (if the metal in a metallo-protein is replaced). It was indicated though, that the substitution of Ca(II) ions with lanthanides in proteins rarely affects their biological activity [137].

The ionic radii decrease throughout the series from 1.17 Å for La to 1.00 Å for Lu – the so-called lanthanide contraction. As a result different lanthanides have different binding affinities when bound to identical sites in proteins [70]. For this reason the diamagnetic references used for measuring paramagnetic restraints should have a similar ionic radius to the paramagnetic ion. La, Lu, Y, or Sc are good candidates where La is better suited for lanthanides from the first half of the series and Lu for the second half [30].

Factors influencing the measurability of paramagnetic restraints

Several variables influence the magnetic properties of paramagnetic metal ions. The most important factors are: (a) the total angular momentum quantum number J for lanthanides or the spin angular momentum S otherwise: the higher this number the more line-broadening will be induced; (b) the magnetic susceptibility anisotropy $\Delta\chi$: larger anisotropy leads to more alignment and larger RDCs and PCSs; (c) the correlation time of the unpaired electron τ_e : larger electron spin correlation times lead to more line-broadening i.e. larger PREs; (d) for the lanthanides with anisotropic magnetic susceptibility the magnetic field strength B is important: for larger magnetic fields the Curie relaxation dominates. Therefore in some cases it is desirable to carry out the measurements at lower fields.

Values of the most important properties, and theoretical values for PCSs, RDCs, and relaxation times, can be found in Table 2. In general, the second half of the lanthanide

series has higher J -values and larger magnetic anisotropies, therefore the PCSs, RDCs, and PREs are generally larger than for the first half of the series.

Specific properties of individual lanthanides

Sm and Eu have low-lying excited states [138] which result in a small population difference between the ground and the first excited state. This leads to small relaxation and line-broadening effects [58].

Gd has a large electron spin correlation time resulting in extremely large line-broadening. This effect can be measured up to 20 Å but its accuracy decreases with increasing distance [57]. The magnetic susceptibility anisotropy of Gd is almost negligible leading to no alignment of the protein and therefore no measurable PCSs and RDCs. Due to its line-broadening capabilities Gd is often used as a surface probe to study protein interfaces (see below). The chemical shift changes induced by Gd are contact shifts that are much larger than for other lanthanides for which they can be neglected [57].

Tm induces moderate to large PCSs, RDCs, and PREs. ^1H - ^{15}N RDCs up to 20 Hz have been measured at 800 MHz [29]. Yb has the smallest contact shift among the lanthanides [72] and it has a similar ionic radius to Ca, making it a very good Ca analog [32].

Paramagnetic metals/compounds other than lanthanides

Mn and nitroxide spin labels such as MTSL have (similarly to Gd) large electron spin correlation times (~0.1 ns and 100 ns respectively) [139] and negligible magnetic anisotropy. The line-broadening effect is less pronounced for Mn and much less for MTSL. Still, MTSL is typically used for measuring PREs in proteins.

For the first-row transition metals the contact shift is very large making them useful shift reagents [58]. Co^+ induces moderate PCSs and binds tightly and specific to EDTA making it a good candidate for structural studies [12].

Interfaces

Paramagnetic centers can be used to map surfaces or binding interfaces of proteins in two different ways: (a) by transferred RDCs or PCSs or (b) by surface PRE effects. Both approaches require an excess of a free ligand and fast exchange between bound and free ligand [131] because otherwise two peaks would be observed.

Transferred RDCs and PCSs

For transferred RDCs and PCSs consider a paramagnetically labeled protein with anisotropic magnetic susceptibility and a ligand without a paramagnetic center. RDCs and PCSs can only be measured for the ligand if it binds to the aligned protein and will therefore also undergo alignment [140]. This works only for internal alignment so that the alignment originates from the MSA. For external alignment transferred RDCs or PCSs are not measurable because the external alignment medium will align the ligand even if it is not bound to the protein. The binding interface can be determined by exploiting the distance-dependence of the PCSs. Transferred PCSs can be used to elucidate the structure of a small molecule ligand bound to the protein. This was illustrated on the ligand thymidine bound to the lanthanide-labeled subunit θ domain of *E.coli* DNA polymerase III [141]. The methodology of transferred RDCs and PCSs also allows probing of conformational changes that might occur upon association.

Surface probes

Broadening reagents can be used to map interfaces by broadening only the peaks of surface residues. Depending on the reagent distances up to 20 Å can be elucidated [142]. The methodology is the same as was described for PREs where the relaxation rate is directly proportional to the concentration of the broadening agent [131]. Interfacial contacts are identified by taking the difference of the spectra in the absence and presence of the ligand since the binding interface is protected from the broadening reagents when the ligand is bound [143]. A limitation, however, is that conformational changes occurring upon binding could be interpreted as being located in the binding interface [143].

Nitroxide spin labels

Several broadening reagents are available [144]. TEMPO, TEMPONE, and TEMPOL (4-hydroxy-2,2,6,6-tetramethyl-piperidine-1-oxyl) are soluble nitroxyl radicals that are frequently used as surface probes [142, 145-146]. The H-bonding donor and acceptor characteristics of TEMPOL make it more similar to water whereas TEMPO or TEMPONE are more hydrophobic, requiring a lower concentration of TEMPOL to obtain identical PREs [145]. However, such nitroxide derivatives or salts of Mn^{2+} sometimes interact with negatively charged amino acid side-chains or detergent head groups of micelles [146].

Gadolinium reagents

Gd-compounds, such as Gd-EDTA [143], Gd-DTPA [146] [147-148] or Gd-DOTA have much more effective line-broadening capabilities and are less prone to interact with protein side-chains, making them widely applicable. Gd-DTPA-BMA for example has been used to elucidate helix orientations and tilt angles using paramagnetic relaxation waves [149].

Doxylstearic acid

16-DSA (16-doxylstearic acid) is a hydrophobic paramagnetic substance that can be used to probe membrane-exposed residues [150]. 5-DSA can be used to probe surface residues because it resides closer to the polar head-groups [151].

Conclusions

Residual Dipolar Couplings (RDCs) and Paramagnetic Relaxation Enhancements (PREs) have become widely applicable restraints for protein structure determination. Whereas RDCs are obtained by partial alignment of the protein in the magnetic field, PREs are usually measured by introducing a paramagnetic spin-label, for instance (1-oxyl-2,2,5,5-tetramethylpyrroline-3-methyl) methanethiosulfonate (MTSL), into the protein. Even though these two approaches seem very different at first, both effects can be observed by exploiting the magnetic properties of certain paramagnetic species that are introduced into the protein. This procedure provides additionally other structural restraints that are not as well-known, such as pseudo-contact shifts (PCSs) or cross-correlated relaxation (CCR) effects. For this reason, the present review gives a complete overview of the paramagnetic restraints available and how they are connected. To maintain practical applicability, small effects are pointed out.

The existence and amplitude of the restraints depends on the anisotropy of the magnetic susceptibility, the total angular momentum quantum number J (or the spin-quantum number S for metals other than lanthanides), the electron spin correlation time, the magnetic field strength, and the size of the molecule as outlined in Tables 1 and 2.

Lanthanide ions are a perfect choice for measuring paramagnetic restraints since they allow partial alignment of the protein in the magnetic field while yielding PREs, PCSs, and other effects. Paramagnetic restraints contain a wealth of structural information that

has, in most cases, only been applied when more easily accessible data was unavailable. They have, however, the potential to replace conventional NMR restraints for larger proteins or protein complexes where they are not available.

References

- [1] I. Bertini, C. Luchinat, G. Parigi, R. Pierattelli, NMR spectroscopy of paramagnetic metalloproteins, *Chembiochem*, 6 (2005) 1536-1549.
- [2] C.R. Sanders, 2nd, G.C. Landis, Reconstitution of membrane proteins into lipid-rich bilayered mixed micelles for NMR studies, *Biochemistry*, 34 (1995) 4030-4040.
- [3] H.J. Sass, G. Musco, S.J. Stahl, P.T. Wingfield, S. Grzesiek, Solution NMR of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes, *J Biomol NMR*, 18 (2000) 303-309.
- [4] M.R. Hansen, L. Mueller, A. Pardi, Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions, *Nat Struct Biol*, 5 (1998) 1065-1074.
- [5] J.C. Hus, D. Marion, M. Blackledge, De novo determination of protein structure by NMR using orientational and long-range order restraints, *J Mol Biol*, 298 (2000) 927-936.
- [6] V. Gaponenko, J.W. Howarth, L. Columbus, G. Gasmi-Seabrook, J. Yuan, W.L. Hubbell, P.R. Rosevear, Protein global fold determination using site-directed spin and isotope labeling, *Protein Sci*, 9 (2000) 302-309.
- [7] M. Allegrozzi, I. Bertini, M.B.L. Janik, Y.M. Lee, G.H. Lin, C. Luchinat, Lanthanide-induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 angstrom from the metal ion, *Journal of the American Chemical Society*, 122 (2000) 4154-4161.
- [8] I. Bertini, G. Cavallaro, M. Cosenza, R. Kummerle, C. Luchinat, M. Piccioli, L. Poggi, Cross correlation rates between Curie spin and dipole-dipole relaxation in paramagnetic proteins: the case of cerium substituted calbindin D9k, *J Biomol NMR*, 23 (2002) 115-125.

- [9] I. Bertini, A. Donaire, B. Jimenez, C. Luchinat, G. Parigi, M. Piccioli, L. Poggi, Paramagnetism-based versus classical constraints: an analysis of the solution structure of Ca Ln calbindin D9k, *J Biomol NMR*, 21 (2001) 85-98.
- [10] M. Gochin, H. Roder, Protein structure refinement based on paramagnetic NMR shifts: applications to wild-type and mutant forms of cytochrome c, *Protein Sci*, 4 (1995) 296-305.
- [11] G. Pintacuda, A. Moshref, A. Leonchiks, A. Sharipo, G. Otting, Site-specific labelling with a metal chelator for protein-structure refinement, *J Biomol NMR*, 29 (2004) 351-361.
- [12] V. Gaponenko, S.P. Sarma, A.S. Altieri, D.A. Horita, J. Li, R.A. Byrd, Improving the accuracy of NMR structures of large proteins using pseudocontact shifts as long-range restraints, *J Biomol NMR*, 28 (2004) 205-212.
- [13] X.C. Su, G. Otting, Paramagnetic labelling of proteins and oligonucleotides for NMR, *J Biomol NMR*, 46 (2010) 101-112.
- [14] I. Bertini, C. Luchinat, S. Aime, NMR of paramagnetic substances, *Coordination Chemistry Reviews*, 150 (1996) R7-&.
- [15] I. Bertini, C. Luchinat, G. Parigi, Paramagnetic constraints: An aid for quick solution structure determination of paramagnetic metalloproteins, *Concepts in Magnetic Resonance*, 14 (2002) 259-286.
- [16] I. Bertini, C. Luchinat, M. Piccioli, Paramagnetic probes in metalloproteins, *Methods Enzymol*, 339 (2001) 314-340.
- [17] L. Banci, I. Bertini, J.G. Huber, C. Luchinat, A. Rosato, Partial orientation of oxidized and reduced cytochrome b(5) at high magnetic fields: Magnetic susceptibility anisotropy contributions and consequences for protein solution structure determination, *Journal of the American Chemical Society*, 120 (1998) 12903-12909.
- [18] S. Bansal, X. Miao, M.W. Adams, J.H. Prestegard, H. Valafar, Rapid classification of protein structure models using unassigned backbone RDCs and probability density profile analysis (PDPA), *J Magn Reson*, 192 (2008) 60-68.
- [19] I. Bertini, C. Luchinat, K.V. Vasavada, The Effect of Magnetic-Anisotropy on the Longitudinal Nuclear-Relaxation Time in Paramagnetic Systems, *Journal of Magnetic Resonance*, 89 (1990) 243-254.

- [20] R. Barbieri, I. Bertini, Y.M. Lee, C. Luchinat, A.H. Velders, Structure-independent cross-validation between residual dipolar couplings originating from internal and external orienting media, *J Biomol NMR*, 22 (2002) 365-368.
- [21] M. Ruckert, G. Otting, Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments, *Journal of the American Chemical Society*, 122 (2000) 7793-7797.
- [22] A. Bax, Weak alignment offers new NMR opportunities to study protein structure and dynamics, *Protein Sci*, 12 (2003) 1-16.
- [23] K. Fleming, D.G. Gray, S. Matthews, Cellulose crystallites, *Chemistry*, 7 (2001) 1831-1835.
- [24] J. Sass, F. Cordier, A. Hoffmann, A. Cousin, J.G. Omichinski, H. Lowen, S. Grzesiek, Purple membrane induced alignment of biological macromolecules in the magnetic field, *Journal of the American Chemical Society*, 121 (1999) 2047-2055.
- [25] C. Li, P. Gao, H. Qin, R. Chase, P.L. Gor'kov, W.W. Brey, T.A. Cross, Uniformly aligned full-length membrane proteins in liquid crystalline bilayers for structural characterization, *J Am Chem Soc*, 129 (2007) 5304-5305.
- [26] M. Zweckstetter, A. Bax, Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR, *Journal of the American Chemical Society*, 122 (2000) 3791-3792.
- [27] F. Rodriguez-Castaneda, P. Haberz, A. Leonov, C. Griesinger, Paramagnetic tagging of diamagnetic proteins for solution NMR, *Magn Reson Chem*, 44 Spec No (2006) S10-16.
- [28] V. Gaponenko, A.S. Altieri, J. Li, R.A. Byrd, Breaking symmetry in the structure determination of (large) symmetric protein dimers, *J Biomol NMR*, 24 (2002) 143-148.
- [29] X.C. Su, K. McAndrew, T. Huber, G. Otting, Lanthanide-binding peptides for NMR measurements of residual dipolar couplings and paramagnetic effects from multiple angles, *J Am Chem Soc*, 130 (2008) 1681-1687.
- [30] G. Otting, Prospects for lanthanides in structural biology by NMR, *J Biomol NMR*, 42 (2008) 1-9.
- [31] D.E. Kamen, S.M. Cahill, M.E. Girvin, Multiple alignment of membrane proteins for measuring residual dipolar couplings using lanthanide ions bound to a small metal chelator, *J Am Chem Soc*, 129 (2007) 1846-1847.

- [32] J.G. Shelling, M.E. Bjornson, R.S. Hodges, A.K. Taneja, B.D. Sykes, Contact and Dipolar Contributions to Lanthanide-Induced Nmr Shifts of Amino-Acid and Peptide Models for Calcium-Binding Sites in Proteins, *Journal of Magnetic Resonance*, 57 (1984) 99-114.
- [33] K.J. Franz, M. Nitz, B. Imperiali, Lanthanide-binding tags as versatile protein coexpression probes, *ChemBiochem*, 4 (2003) 265-271.
- [34] D. Haussinger, J.R. Huang, S. Grzesiek, DOTA-M8: An Extremely Rigid, High-Affinity Lanthanide Chelating Tag for PCS NMR Spectroscopy, *J Am Chem Soc*, (2009).
- [35] A. Leonov, B. Voigt, F. Rodriguez-Castaneda, P. Sakhaii, C. Griesinger, Convenient synthesis of multifunctional EDTA-based chiral metal chelates substituted with an S-mesylcysteine, *Chemistry*, 11 (2005) 3342-3348.
- [36] C. Ma, S.J. Opella, Lanthanide ions bind specifically to an added "EF-hand" and orient a membrane protein in micelles for solution NMR spectroscopy, *J Magn Reson*, 146 (2000) 381-384.
- [37] T. Saio, K. Ogura, M. Yokochi, Y. Kobashigawa, F. Inagaki, Two-point anchoring of a lanthanide-binding peptide to a target protein enhances the paramagnetic anisotropic effect, *J Biomol NMR*, 44 (2009) 157-166.
- [38] P.H. Keizers, J.F. Desreux, M. Overhand, M. Ubbink, Increased paramagnetic effect of a lanthanide protein probe by two-point attachment, *J Am Chem Soc*, 129 (2007) 9292-9293.
- [39] X.C. Su, B. Man, S. Beeren, H. Liang, S. Simonsen, C. Schmitz, T. Huber, B.A. Messerle, G. Otting, A dipicolinic acid tag for rigid lanthanide tagging of proteins and paramagnetic NMR spectroscopy, *J Am Chem Soc*, 130 (2008) 10486-10487.
- [40] J.R. Tolman, J.M. Flanagan, M.A. Kennedy, J.H. Prestegard, Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution, *Proc Natl Acad Sci U S A*, 92 (1995) 9279-9283.
- [41] C.D. Schwieters, J.Y. Suh, A. Grishaev, R. Ghirlando, Y. Takayama, G.M. Clore, Solution Structure of the 128 kDa Enzyme I Dimer from *Escherichia coli* and Its 146 kDa Complex with HPr Using Residual Dipolar Couplings and Small- and Wide-Angle X-ray Scattering, *J Am Chem Soc*, (2010).
- [42] A. Bax, G. Kontaxis, N. Tjandra, Dipolar couplings in macromolecular structure determination, *Methods Enzymol*, 339 (2001) 127-174.

- [43] G. Bouvignies, P.R. Markwick, M. Blackledge, Simultaneous definition of high resolution protein structure and backbone conformational dynamics using NMR residual dipolar couplings, *Chemphyschem*, 8 (2007) 1901-1909.
- [44] F. Kramer, M.V. Deshmukh, H. Kessler, S.J. Glaser, Residual dipolar coupling constants: An elementary derivation of key equations, *Concepts in Magnetic Resonance Part A*, 21A (2004) 10-21.
- [45] I. Bertini, M.B. Janik, Y.M. Lee, C. Luchinat, A. Rosato, Magnetic susceptibility tensor anisotropies for a lanthanide ion series in a fixed protein matrix, *J Am Chem Soc*, 123 (2001) 4181-4188.
- [46] E. de Alba, N. Tjandra, On the accurate measurement of amide one-bond ^{15}N - ^1H couplings in proteins: effects of cross-correlated relaxation, selective pulses and dynamic frequency shifts, *J Magn Reson*, 183 (2006) 160-165.
- [47] R. Ghose, J.H. Prestegard, Electron spin-nuclear spin cross-correlation effects on multiplet splittings in paramagnetic proteins, *J Magn Reson*, 128 (1997) 138-143.
- [48] L. Werbelow, R.E. London, Dynamic frequency shift, *Concepts in Magnetic Resonance*, 8 (1996) 325-338.
- [49] K. Ding, A.M. Gronenborn, Sensitivity-enhanced 2D IPAP, TROSY-anti-TROSY, and E.COSY experiments: alternatives for measuring dipolar ^{15}N - ^1H couplings, *J Magn Reson*, 163 (2003) 208-214.
- [50] K. Pervushin, R. Riek, G. Wider, K. Wuthrich, Attenuated T_2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution, *Proc Natl Acad Sci U S A*, 94 (1997) 12366-12371.
- [51] Y. Liu, J.H. Prestegard, Measurement of one and two bond N-C couplings in large proteins by TROSY-based J-modulation experiments, *J Magn Reson*, 200 (2009) 109-118.
- [52] C. Guo, R. Godoy-Ruiz, V. Tugarinov, High Resolution Measurement of Methyl $(^{13}\text{C}(m))$ - (^{13}C) and $(^1\text{H}(m))$ - $(^{13}\text{C}(m))$ Residual Dipolar Couplings in Large Proteins, *J Am Chem Soc*, (2010).
- [53] W. Bermel, I. Bertini, I.C. Felli, R. Peruzzini, R. Pierattelli, Exclusively heteronuclear NMR experiments to obtain structural and dynamic information on proteins, *Chemphyschem*, 11 (2010) 689-695.

- [54] N.A. Lakomek, K.F. Walter, C. Fares, O.F. Lange, B.L. de Groot, H. Grubmuller, R. Bruschweiler, A. Munk, S. Becker, J. Meiler, C. Griesinger, Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics, *J Biomol NMR*, (2008).
- [55] J.R. Tolman, Dipolar couplings as a probe of molecular dynamics and structure in solution, *Curr Opin Struct Biol*, 11 (2001) 532-539.
- [56] J. Meiler, J.J. Prompers, W. Peti, C. Griesinger, R. Bruschweiler, Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins, *J Am Chem Soc*, 123 (2001) 6098-6107.
- [57] J.A. Peters, J. Huskens, D.J. Raber, Lanthanide induced shifts and relaxation rate enhancements, *Prog Nucl Mag Res Sp*, 28 (1996) 283-350.
- [58] B.C. Mayo, Lanthanide Shift Reagents in Nuclear Magnetic-Resonance Spectroscopy, *Chemical Society Reviews*, 2 (1973) 49-74.
- [59] C.D. Barry, J.A. Glasel, R.J. Williams, A.V. Xavier, Quantitative determination of conformations of flexible molecules in solution using lanthanide ions as nuclear magnetic resonance probes: application to adenosine-5'-monophosphate, *J Mol Biol*, 84 (1974) 471-409.
- [60] R.M. Golding, L.C. Stubbs, Nmr Shifts in Paramagnetic Systems - Nonmultipole Expansion Method, *Journal of Magnetic Resonance*, 33 (1979) 627-647.
- [61] I. Bertini, C. Luchinat, G. Parigi, Hyperfine shifts in low-spin iron(III) hemes: A ligand field analysis, *European Journal of Inorganic Chemistry*, (2000) 2473-2480.
- [62] H.M. McConnell, A Pseudovector Nuclear Hyperfine Interaction, *Proc Natl Acad Sci U S A*, 44 (1958) 766-767.
- [63] M.D. Kemple, B.D. Ray, K.B. Lipkowitz, F.G. Prendergast, B.D.N. Rao, The Use of Lanthanides for Solution Structure Determination of Biomolecules by Nmr - Evaluation of the Methodology with Edta Derivatives as Model Systems, *Journal of the American Chemical Society*, 110 (1988) 8275-8287.
- [64] B.M. McGarvey, R.J. Kurland, *NMR of Paramagnetic Molecules*, Academic Press, p.559, New York, 1973.
- [65] W.D. Horrocks, Jr., J.P. Sipe, 3rd, Lanthanide Complexes as Nuclear Magnetic Resonance Structural Probes: Paramagnetic Anisotropy of Shift Reagent Adducts, *Science*, 177 (1972) 994-996.

- [66] I.I. Bertini, I.C. Felli, C. Luchinat, High magnetic field consequences on the NMR hyperfine shifts in solution, *J Magn Reson*, 134 (1998) 360-364.
- [67] B.I. Bleaney, Bleaney, B., *Electricity and magnetism*, 3rd Edition ed., Oxford University Press, Oxford, 1976.
- [68] M. John, G. Otting, Strategies for measurements of pseudocontact shifts in protein NMR spectroscopy, *Chemphyschem*, 8 (2007) 2309-2313.
- [69] B. Shapira, J.H. Prestegard, Electron-nuclear interactions as probes of domain motion in proteins, *J Chem Phys*, 132 (2010) 115102.
- [70] G. Pintacuda, M. John, X.C. Su, G. Otting, NMR structure determination of protein-ligand complexes by lanthanide labeling, *Acc Chem Res*, 40 (2007) 206-212.
- [71] L. Banci, I. Bertini, G.G. Savellini, A. Romagnoli, P. Turano, M.A. Cremonini, C. Luchinat, H.B. Gray, Pseudocontact shifts as constraints for energy minimization and molecular dynamics calculations on solution structures of paramagnetic metalloproteins, *Proteins-Structure Function and Genetics*, 29 (1997) 68-76.
- [72] L. Lee, B.D. Sykes, Strategies for the uses of lanthanide NMR shift probes in the determination of protein structure in solution. Application to the EF calcium binding site of carp parvalbumin, *Biophys J*, 32 (1980) 193-210.
- [73] C. Schmitz, M.J. Stanton-Cook, X.C. Su, G. Otting, T. Huber, Numbat: an interactive software tool for fitting Delta chi-tensors to molecular coordinates using pseudocontact shifts, *J Biomol NMR*, 41 (2008) 179-189.
- [74] M. John, A.Y. Park, G. Pintacuda, N.E. Dixon, G. Otting, Weak alignment of paramagnetic proteins warrants correction for residual CSA effects in measurements of pseudocontact shifts, *J Am Chem Soc*, 127 (2005) 17190-17191.
- [75] S. Tate, H. Shimahara, N. Utsunomiya-Tate, Molecular-orientation analysis based on alignment-induced TROSY chemical shift changes, *J Magn Reson*, 171 (2004) 284-292.
- [76] N. Tjandra, A. Bax, Solution NMR measurement of amide proton chemical shift anisotropy in N-15-enriched proteins. Correlation with hydrogen bond length, *Journal of the American Chemical Society*, 119 (1997) 8076-8082.
- [77] A.L. Hansen, H.M. Al-Hashimi, Insight into the CSA tensors of nucleobase carbons in RNA polynucleotides from solution measurements of residual CSA: towards new long-range orientational constraints, *J Magn Reson*, 179 (2006) 299-307.

- [78] A. Grishaev, J. Ying, A. Bax, Pseudo-CSA restraints for NMR refinement of nucleic acid structure, *J Am Chem Soc*, 128 (2006) 10010-10011.
- [79] J. Reuben, Origin of Chemical-Shifts in Lanthanide Complexes and Some Implications Thereof, *Journal of Magnetic Resonance*, 11 (1973) 103-104.
- [80] S.D. Emerson, G. La Mar, Solution structural characteristics of cyanometmyoglobin: resonance assignment of heme cavity residues by two-dimensional NMR, *Biochemistry*, 29 (1990) 1545-1556.
- [81] J. Meiler, W. Peti, C. Griesinger, DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts, *J Biomol NMR*, 17 (2000) 283-294.
- [82] M. Ottiger, A. Bax, Determination of relative N-H-N N-C', C-alpha-C', and C(alpha)-H-alpha effective bond lengths in a protein by NMR in a dilute liquid crystalline phase, *J Am Chem Soc*, 120 (1998) 12334-12341.
- [83] J.A. Losonczi, M. Andrec, M.W. Fischer, J.H. Prestegard, Order matrix analysis of residual dipolar couplings using singular value decomposition, *J Magn Reson*, 138 (1999) 334-342.
- [84] J. Meiler, N. Blomberg, M. Nilges, C. Griesinger, A new approach for applying residual dipolar couplings as restraints in structure elucidation, *J Biomol NMR*, 16 (2000) 245-252.
- [85] A.C. Drohat, N. Tjandra, D.M. Baldisseri, D.J. Weber, The use of dipolar couplings for determining the solution structure of rat apo-S100B(beta-beta), *Protein Sci*, 8 (1999) 800-809.
- [86] G.M. Clore, A.M. Gronenborn, A. Bax, A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information, *J Magn Reson*, 133 (1998) 216-221.
- [87] M. Zweckstetter, NMR: prediction of molecular alignment from structure using the PALES software, *Nat Protoc*, 3 (2008) 679-690.
- [88] M. Zweckstetter, G. Hummer, A. Bax, Prediction of charge-induced molecular alignment of biomolecules dissolved in dilute liquid-crystalline phases, *Biophys J*, 86 (2004) 3444-3460.
- [89] K. Berlin, D.P. O'Leary, D. Fushman, Improvement and analysis of computational methods for prediction of residual dipolar couplings, *J Magn Reson*, (2009) 25-33.

- [90] M.F. Mesleh, G. Veglia, T.M. DeSilva, F.M. Marassi, S.J. Opella, Dipolar waves as NMR maps of protein structure, *J Am Chem Soc*, 124 (2002) 4206-4207.
- [91] K. Chen, N. Tjandra, Top-down approach in protein RDC data analysis: de novo estimation of the alignment tensor, *J Biomol NMR*, 38 (2007) 303-313.
- [92] R. Barbieri, C. Luchinat, G. Parigi, Backbone-only protein solution structures with a combination of classical and paramagnetism-based constraints: a method that can be scaled to large molecules, *Chemphyschem*, 5 (2004) 797-806.
- [93] G. Cornilescu, J.L. Marquardt, M. Ottiger, A. Bax, Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase, *Journal of the American Chemical Society*, 120 (1998) 6836-6837.
- [94] K. Ruan, K.B. Briggman, J.R. Tolman, De novo determination of internuclear vector orientations from residual dipolar couplings measured in three independent alignment media, *J Biomol NMR*, 41 (2008) 61-76.
- [95] X. Miao, R. Mukhopadhyay, H. Valafar, Estimation of relative order tensors, and reconstruction of vectors in space using unassigned RDC data and its application, *J Magn Reson*, 194 (2008) 202-211.
- [96] P. Haberz, F. Rodriguez-Castaneda, J. Junker, S. Becker, A. Leonov, C. Griesinger, Two new chiral EDTA-based metal chelates for weak alignment of proteins in solution, *Org Lett*, 8 (2006) 1275-1278.
- [97] M. Habeck, M. Nilges, W. Rieping, A unifying probabilistic framework for analyzing residual dipolar couplings, *J Biomol NMR*, 40 (2008) 135-144.
- [98] F. Delaglio, G. Kontaxis, A. Bax, Protein structure determination using molecular fragment replacement and NMR dipolar couplings, *Journal of the American Chemical Society*, 122 (2000) 2142-2143.
- [99] F. Capozzi, M.A. Cremonini, C. Luchinat, M. Sola, Assignment of Pseudo-Contact-Shifted H-1-Nmr Resonances in the Ef Site of Yb³⁺-Substituted Rabbit Parvalbumin through a Combination of 2d Techniques and Magnetic-Susceptibility Tensor Determination, *Magnetic Resonance in Chemistry*, 31 (1993) S118-S127.
- [100] J. Meiler, D. Baker, Rapid protein fold determination using unassigned NMR data, *Proc Natl Acad Sci U S A*, 100 (2003) 15404-15409.
- [101] J. Meiler, D. Baker, The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy, *J Magn Reson*, 173 (2005) 310-316.

- [102] H. Valafar, J.H. Prestegard, REDCAT: a residual dipolar coupling analysis tool, *J Magn Reson*, 167 (2004) 228-241.
- [103] L. Banci, I. Bertini, M.A. Cremonini, G. Gori-Savellini, C. Luchinat, K. Wuthrich, P. Guntert, PSEUDYANA for NMR structure calculation of paramagnetic metalloproteins using torsion angle molecular dynamics, *Journal of Biomolecular Nmr*, 12 (1998) 553-557.
- [104] P. Shealy, M. Simin, S.H. Park, S.J. Opella, H. Valafar, Simultaneous structure and dynamics of a membrane protein using REDCRAFT: membrane-bound form of Pf1 coat protein, *J Magn Reson*, 207 (2010) 8-16.
- [105] M. Bryson, F. Tian, J.H. Prestegard, H. Valafar, REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data, *J Magn Reson*, 191 (2008) 322-334.
- [106] H.F. Azurmendi, C.A. Bush, Tracking alignment from the moment of inertia tensor (TRAMITE) of biomolecules in neutral dilute liquid crystal solutions, *J Am Chem Soc*, 124 (2002) 2426-2427.
- [107] L. Banci, I. Bertini, G. Cavallaro, A. Giachetti, C. Luchinat, G. Parigi, Paramagnetism-based restraints for Xplor-NIH, *J Biomol NMR*, 28 (2004) 249-261.
- [108] C.D. Schwieters, J.J. Kuszewski, N. Tjandra, G.M. Clore, The Xplor-NIH NMR molecular structure determination package, *J Magn Reson*, 160 (2003) 65-73.
- [109] C. Schmitz, M. John, A.Y. Park, N.E. Dixon, G. Otting, G. Pintacuda, T. Huber, Efficient chi-tensor determination and NH assignment of paramagnetic proteins, *J Biomol NMR*, 35 (2006) 79-87.
- [110] G. Pintacuda, M.A. Keniry, T. Huber, A.Y. Park, N.E. Dixon, G. Otting, Fast structure-based assignment of ¹⁵N HSQC spectra of selectively ¹⁵N-labeled paramagnetic proteins, *J Am Chem Soc*, 126 (2004) 2963-2970.
- [111] I. Bertini, L. Banci, C. Luchinat, Proton magnetic resonance of paramagnetic metalloproteins, *Methods Enzymol*, 177 (1989) 246-263.
- [112] A. Abragam, *Principles of Nuclear Magnetism*, Clarendon Press, Oxford, 1994.
- [113] L. Lee, B.D. Sykes, Nuclear magnetic resonance determination of metal-proton distances in the EF site of carp parvalbumin using the susceptibility contribution to the line broadening of lanthanide-shifted resonances, *Biochemistry*, 19 (1980) 3208-3214.

- [114] I. Solomon, Relaxation Processes in a System of Two Spins, *Physical Review Letters*, 99 (1955) 559-565.
- [115] G.S.H. Rule, T. Kevin, *Fundamentals of Protein NMR Spectroscopy*, 1st edition ed., Springer, 2005.
- [116] R. Paquin, P. Pelupessy, L. Duma, C. Gervais, G. Bodenhausen, Determination of the antisymmetric part of the chemical shift anisotropy tensor via spin relaxation in nuclear magnetic resonance, *J Chem Phys*, 133 (2010) 034506.
- [117] F.A.L. Anet, D.J. O'Leary, The shielding tensor part II: Understanding its strange effects on relaxation, *Concepts in Magnetic Resonance*, 4 (1992) 35-52.
- [118] F.A.L. Anet, D.J. O'Leary, The shielding tensor. Part I: Understanding its symmetry properties, *Concepts in Magnetic Resonance*, 3 (1991) 193-214.
- [119] P. Luginbuhl, K. Wuthrich, Semi-classical nuclear spin relaxation theory revisited for use with biological macromolecules, *Prog Nucl Mag Res Sp*, 40 (2002) 199-247.
- [120] J. Weigelt, Single scan, sensitivity- and gradient-enhanced TROSY for multidimensional NMR experiments, *Journal of the American Chemical Society*, 120 (1998) 10778-10779.
- [121] B. Reif, M. Hennig, C. Griesinger, Direct measurement of angles between bond vectors in high-resolution NMR, *Science*, 276 (1997) 1230-1233.
- [122] D.F. Hansen, J.J. Led, Determination of the geometric structure of the metal site in a blue copper protein by paramagnetic NMR, *Proc Natl Acad Sci U S A*, 103 (2006) 1738-1743.
- [123] A. Overhauser, Polarization of Nuclei in Metals, *Physical Review*, 92 (1953) 411-415.
- [124] K.V. Vasavada, B.D.N. Rao, Nuclear-Spin Relaxation in Liquids Due to Interaction with Paramagnetic-Ions Having Anisotropic G-Tensors, *Journal of Magnetic Resonance*, 81 (1989) 275-283.
- [125] G. Pintacuda, A. Kaikkonen, G. Otting, Modulation of the distance dependence of paramagnetic relaxation enhancements by CSA x DSA cross-correlation, *J Magn Reson*, 171 (2004) 233-243.

- [126] G. Pintacuda, K. Hohenthanner, G. Otting, N. Muller, Angular dependence of dipole-dipole-Curie-spin cross-correlation effects in high-spin and low-spin paramagnetic myoglobin, *J Biomol NMR*, 27 (2003) 115-132.
- [127] J. Boisbouvier, P. Gans, M. Blackledge, B. Brutscher, D. Marion, Long-range structural information in NMR studies of paramagnetic molecules from electron spin-nuclear spin cross-correlated relaxation, *Journal of the American Chemical Society*, 121 (1999) 7700-7701.
- [128] I. Bertini, C. Luchinat, P. Turano, G. Battaini, L. Casella, The magnetic properties of myoglobin as studied by NMR spectroscopy, *Chemistry*, 9 (2003) 2316-2322.
- [129] J.L. Battiste, G. Wagner, Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data, *Biochemistry*, 39 (2000) 5355-5365.
- [130] J. Iwahara, C. Tang, G.M. Clore, Practical aspects of H-1 transverse paramagnetic relaxation enhancement measurements on macromolecules, *Journal of Magnetic Resonance*, 184 (2007) 185-195.
- [131] P.A. Kosen, Spin Labeling of Proteins, *Methods in Enzymology*, 177 (1989) 86-121.
- [132] G.M. Clore, J. Iwahara, Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population States of biological macromolecules and their complexes, *Chem Rev*, 109 (2009) 4108-4139.
- [133] J. Iwahara, C.D. Schwieters, G.M. Clore, Ensemble approach for NMR structure refinement against (1)H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule, *J Am Chem Soc*, 126 (2004) 5879-5896.
- [134] M.J. Sutcliffe, C.M. Dobson, Relaxation data in NMR structure determination: model calculations for the lysozyme-Gd³⁺ complex, *Proteins*, 10 (1991) 117-129.
- [135] M.E. Girvin, R.H. Fillingame, Determination of local protein structure by spin label difference 2D NMR: the region neighboring Asp61 of subunit c of the F1F0 ATP synthase, *Biochemistry*, 34 (1995) 1635-1645.
- [136] K.N. Allen, B. Imperiali, Lanthanide-tagged proteins--an illuminating partnership, *Curr Opin Chem Biol*, 14 (2010) 247-254.

- [137] R.B. Martin, *Calcium in Biology*, John Wiley, New York, 1983.
- [138] B. Bleaney, R.J. Williams, A.V. Xavier, R.B. Martin, B.A. Levine, C.M. Dobson, Origin of Lanthanide Nuclear Magnetic-Resonance Shifts and Their Uses, *Journal of the Chemical Society-Chemical Communications*, (1972) 791-793.
- [139] P.S. Nadaud, J.J. Helmus, S.L. Kall, C.P. Jaroniec, Paramagnetic ions enable tuning of nuclear relaxation rates and provide long-range structural restraints in solid-state NMR of proteins, *J Am Chem Soc*, 131 (2009) 8108-8120.
- [140] T. Ikegami, L. Verdier, P. Sakhaii, S. Grimme, B. Pescatore, K. Saxena, K.M. Fiebig, C. Griesinger, Novel techniques for weak alignment of proteins in solution using chemical tags coordinating lanthanide ions, *J Biomol NMR*, 29 (2004) 339-349.
- [141] M. John, G. Pintacuda, A.Y. Park, N.E. Dixon, G. Otting, Structure determination of protein-ligand complexes by transferred paramagnetic shifts, *J Am Chem Soc*, 128 (2006) 12910-12916.
- [142] N.U. Jain, A. Venot, K. Umemoto, H. Leffler, J.H. Prestegard, Distance mapping of protein-binding sites using spin-labeled oligosaccharide ligands, *Protein Sci*, 10 (2001) 2393-2400.
- [143] S. Arumugam, C.L. Hemme, N. Yoshida, K. Suzuki, H. Nagase, M. Berjanskii, B. Wu, S.R. Van Doren, TIMP-1 contact sites and perturbations of stromelysin 1 mapped by NMR and a paramagnetic surface probe, *Biochemistry*, 37 (1998) 9650-9657.
- [144] H.J. Kim, S.C. Howell, W.D. Van Horn, Y.H. Jeon, C.R. Sanders, Recent advances in the application of solution NMR spectroscopy to multi-span integral membrane proteins, *Prog Nucl Mag Res Sp*, 55 (2009) 335-360.
- [145] M. Scarselli, A. Bernini, C. Segoni, H. Molinari, G. Esposito, A.M. Lesk, F. Laschi, P. Temussi, N. Niccolai, Tendamistat surface accessibility to the TEMPOL paramagnetic probe, *J Biomol NMR*, 15 (1999) 125-133.
- [146] A.M. Petros, L. Mueller, K.D. Kopple, NMR identification of protein surfaces using paramagnetic probes, *Biochemistry*, 29 (1990) 10041-10048.
- [147] J.J. Falke, L.A. Luck, J. Scherrer, ¹⁹F nuclear magnetic resonance studies of aqueous and transmembrane receptors. Examples from the *Escherichia coli* chemosensory pathway, *Biophys J*, 62 (1992) 82-86.

- [148] G. Pintacuda, G. Otting, Identification of protein surfaces by NMR measurements with a paramagnetic Gd(III) chelate, *J Am Chem Soc*, 124 (2002) 372-373.
- [149] M. Respondek, T. Madl, C. Gobl, R. Golser, K. Zangger, Mapping the orientation of helices in micelle-bound peptides by paramagnetic relaxation waves, *J Am Chem Soc*, 129 (2007) 5228-5234.
- [150] F.L. Garcia, T. Szyperski, J.H. Dyer, T. Choinowski, U. Seedorf, H. Hauser, K. Wuthrich, NMR structure of the sterol carrier protein-2: implications for the biological role, *J Mol Biol*, 295 (2000) 595-603.
- [151] C.H. Papavoine, R.N. Konings, C.W. Hilbers, F.J. van de Ven, Location of M13 coat protein in sodium dodecyl sulfate micelles as determined by NMR, *Biochemistry*, 33 (1994) 12990-12997.
- [152] D.E. Woessner, Nuclear spin relaxation in ellipsoids undergoing rotational Brownian motion, *Journal of Chemical Physics*, 37 (1962) 647-654.
- [153] I. Bertini, F. Capozzi, C. Luchinat, G. Nicastro, Z.C. Xia, Water Proton Relaxation for Some Lanthanide Aqua Ions in Solution, *Journal of Physical Chemistry*, 97 (1993) 6351-6354.
- [154] I.I. Bertini, O. Galas, C. Luchinat, G. Parigi, G. Spina, Nuclear and Electron Relaxation in Magnetic Exchange Coupled Dimers: Implications for NMR Spectroscopy, *J Magn Reson*, 130 (1998) 33-44.
- [155] H.M. Al-Hashimi, H. Valafar, M. Terrell, E.R. Zartler, M.K. Eidsness, J.H. Prestegard, Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings, *J Magn Reson*, 143 (2000) 402-406.
- [156] R.M. Golding, M.P. Halton, Theoretical Study of N-14 and O-17 Nmr Shifts in Lanthanide Complexes, *Australian Journal of Chemistry*, 25 (1972) 2577-2581.
- [157] C.D. Barry, A.C. North, J.A. Glasel, R.J. Williams, A.V. Xavier, Quantitative determination of mononucleotide conformations in solution using lanthanide ion shift and broadening NMR probes, *Nature*, 232 (1971) 236-245.
- [158] P.E. Johnson, E. Brun, L.F. MacKenzie, S.G. Withers, L.P. McIntosh, The cellulose-binding domains from *Cellulomonas fimi* beta-1, 4-glucanase CenC bind nitroxide spin-labeled celooligosaccharides in multiple orientations, *J Mol Biol*, 287 (1999) 609-625.

CHAPTER 2

Structural studies on KCNE3 using paramagnetic restraints obtained from lanthanide tagging experiments

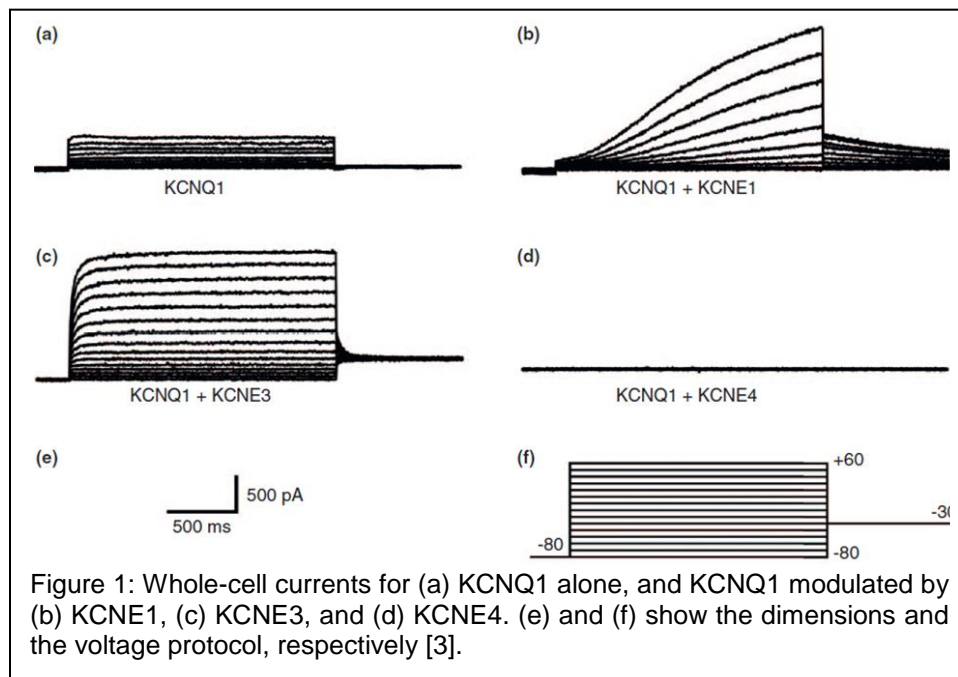
Introduction

KCNQ1 (also called Kv7.1 or KvLQT1) is a voltage-gated potassium channel expressed in the inner ear and the heart, where it is responsible for the delayed rectifier current I_{ks} necessary for the heartbeat [1-2]. KCNQ1 is a homo-tetrameric protein and each of the four α -subunits consists of six trans-membrane helices. The first four of the helices S1-S4 form the voltage sensor domain and S5-S6 form the pore domain. KCNQ1 is modulated by accessory β -subunits which are members of the KCNE family of single trans-membrane span proteins. This family encompasses five members, KCNE1 (minK) to KCNE5, where KCNE2 to KCNE5 are also referred to as MiRP1 (minK related protein 1) to MiRP4. Members of the KCNE family modulate a variety of different voltage-gated potassium channels and all five of them modulate KCNQ1 [4]. Mutations in either the channel or the accessory subunits can lead to cardiac arrhythmias such as familial atrial fibrillation or long QT syndrome [5-9]. This increases the risk of *torsade de pointes*, a form of irregular heartbeat that can lead to palpitations, fainting, and sudden death caused by ventricular fibrillation.

The stoichiometry of the KCNQ1/KCNE complex has been studied extensively without completely conclusive results [10-11]. Although there is evidence that four KCNQ1 subunits associate with two KCNE proteins [12-14], stoichiometries of 4:4 have also been suggested [15]. In a recent study Isacoff and co-workers suggest that the stoichiometry is flexible with up to four KCNE proteins associating with a KCNQ1 tetramer depending on the relative expression densities of the two proteins [15].

Interestingly, the authors also found that the kinetics of gating depends on the relative expression densities and propose that the rhythm of the heartbeat could be regulated by altering the expression densities of these two proteins in heart muscle tissue.

Each of the KCNE family members modulates KCNQ1 (among other potassium channels) in a different manner. Figure 1 shows the whole cell currents of KCNQ1 alone and in complex with KCNE1, KCNE3, and KCNE4 [3]. Association of KCNQ1 with KCNE1 leads to delayed channel activation with a slow opening of the channel enhancing the open state conductance and resulting in more positive potentials [4, 16-17]. KCNE3 opens the channel, enhances conductance and suppresses the voltage-dependence of the gating leading to a dramatic increase of the currents [18]. KCNE4, in contrast, abolishes function completely resulting in no current [19].



It has been shown that KCNE family members modulate channel activity at least partly by an interaction in the trans-membrane domains [20-22] and structural knowledge of the trans-membrane domains of KCNE proteins would allow for hypothesis of how

they specifically interact with the channel. A challenge is that except for KCNE1 there are no structures of KCNQ1 or the accessory subunits available so far, even though KCNQ1 has been modeled computationally [23]. The open state model was created using homology modeling based on Kv1.2 as a template structure. The closed state structure was created using the model of Kv1.2 as a template [24] that was in turn modeled using KcsA as a template. In both cases the modeling was accomplished by a mixture of homology modeling in MOE (Molecular Operating Environment), loop modeling and *de novo* modeling of missing residues in Rosetta, and refinement in Amber. The solution structure of KCNE1 has been determined using NMR spectroscopy

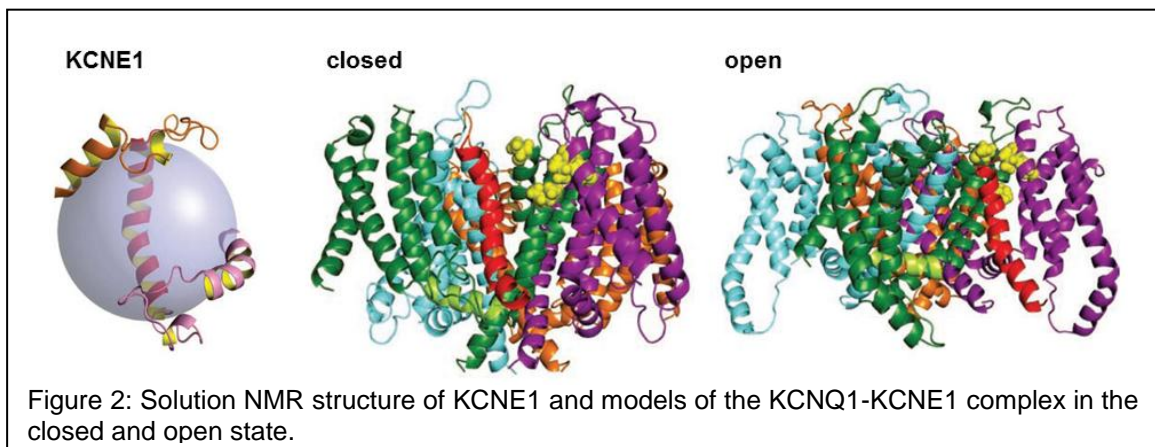
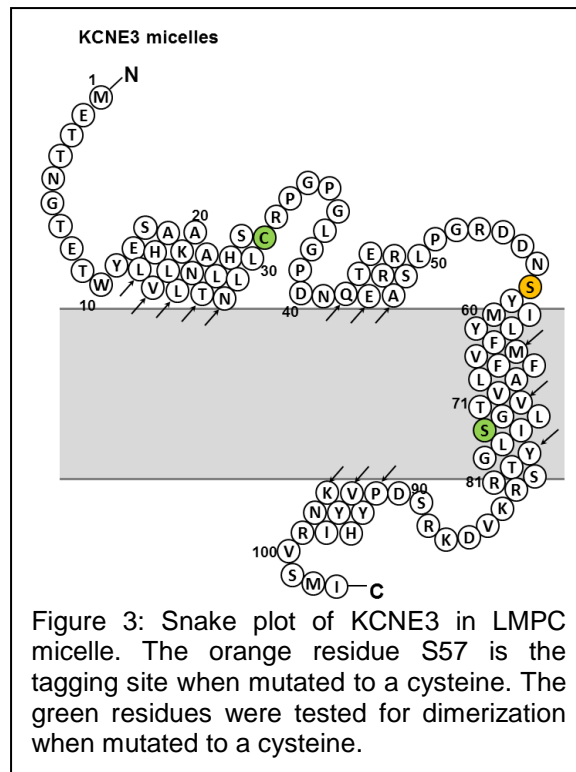


Figure 2: Solution NMR structure of KCNE1 and models of the KCNQ1-KCNE1 complex in the closed and open state. [25-26]. The protein has a flexible N- and C-terminus and a curved α -helical trans-membrane domain (Figure 2). KCNE1 was docked into the open and closed state models of KCNQ1 using Rosetta. These models led to the hypothesis that the bent trans-membrane helix of KCNE1 is 'sitting' on the S4-S5 linker helix such that it slows down motion of this helix during channel opening leading to decreased conductance at the beginning of channel opening. This hypothesis has to be tested by a number of electrophysiological measurements of mutant channels and accessory subunits. The challenge in experimental verification is the fact that the KCNQ1/KCNE1 complex can exist in a number of different equilibrium states where the existence of an open or closed

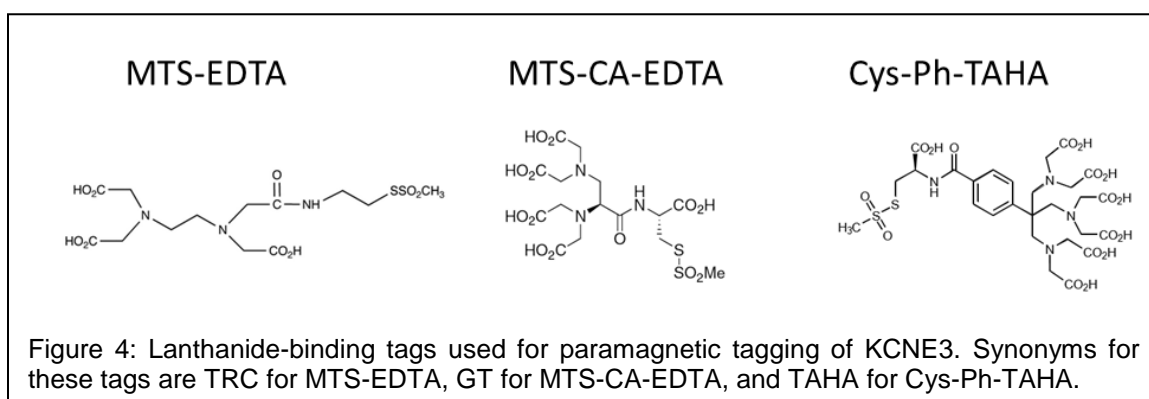
conformation with its associated KCNQ1/KCNE1 interactions depends both on the voltage applied as well as whether KCNE1 is associated with the channel.

KCNE3 and KCNE4 is currently being studied in the Sanders laboratory with structure determination efforts underway. The structure of KCNE3 in DHPC/DMPC bicelles will be determined shortly using conventional NMR restraints including PREs obtained using a MTSL spin label and RDC data acquired in polyacrylamide gels. A future aim is that this structure will be compared to the structure of KCNE3 in LMPC micelles that will be elucidated using paramagnetic restraints, such as PREs, RDCs, and PCSs. This comparison will yield structural differences that arise by incorporation of KCNE3 into micelles versus bicelles and would be the first comparison of a membrane protein structure in both micelles and bicelles using solution NMR. The determined structures will be docked into the KCNQ1 open- and closed state models to generate testable hypotheses for why KCNQ1 modulation by KCNE1 differs from its modulation by KCNE3. Furthermore, PRE and RDC data is available for KCNE3 in LMPC micelles



acquired using conventional techniques. The structure determined using these restraints could be compared to the structure determined solely from paramagnetic restraints obtained from lanthanide tagging to obtain error margins as well as advantages and disadvantages of both approaches.

KCNE3 consists of 103 residues (Figure 3) that are arranged in one trans-membrane helix, one N-terminal helix, and possibly one short C-terminal helix that is only observed by Chemical Shift Index (CSI) data in bicelles [27]. Over-expression of KCNE3 into LMPC micelles and DHPC/DMPC bicelles was achieved and assignments could be obtained for both media [27]. Whole cell current recordings of *Xenopus* oocytes injected with these micelles or bicelles showed that the bicelle injected cells contained KCNQ1 channel modulated by KCNE3 due to co-assembly of the two proteins, whereas in the micelle injected cells KCNQ1 was not functionally modulated by KCNE3 [27]. This points to a subtle structural difference between KCNE3 in these two environments that could result in a different trafficking behavior that would explain why KCNQ1 is modulated by KCNE3 in bicelles but not in micelles.



The paramagnetic NMR studies will be carried out using three different lanthanide-binding tags (Figure 4): a commercially available MTS-EDTA (methane-thio-sulfonyl-cysteaminy-ethylene-diamine-tetraacetic acid, MW = 429 Da) tag obtained from Toronto Research Chemicals, and manually synthesized MTS-CA-EDTA (methane-thio-sulfonyl-cysteaminy-carbonic-acid-ethylene-diamine-tetraacetic acid, MW = 517 Da) and

Cys-Ph-TAHA (cysteiny-phenyl-triaminohexaacetate, MW = 753 Da) tags. The use of tags similar to the commercially available MTS-EDTA tag has been described in the literature [28-30]. In our study we use MTS-EDTA in preliminary experiments to establish the tagging protocol. The MTS-CA-EDTA tag has the advantage of forming a single isomer when complexed with lanthanide ions which is of tremendous advantage in spectral analysis since a chiral complex can lead to the formation of stereoisomers in the sample that result in two sets of peaks in the spectra hampering the measurement of restraints. This tag also has the advantage of a shorter linker providing the basis for a stronger alignment and larger amplitudes of the restraints. It is long-term stable and binds metal ions with very high affinity in the picomolar range, and therefore can be used for metal-binding proteins. The Cys-Ph-TAHA tag has similar favorable characteristics with one more advantage: it has nine coordination sites for lanthanide ions which saturates the binding sites of the lanthanide. Since the lanthanide should be fully enclosed by the Cys-Ph-TAHA tag, direct interactions of the lanthanide ion with the protein should be minimized. This leads to a reduced perturbation of the electronic environment of the protein and hence smaller chemical shift disturbances in the NMR spectrum by the lanthanide ion resulting in restraints of higher quality.

Methods

Expression and purification of KCNE3

KCNE3 with a hexa-His tag encoded in a pET16b expression vector was transformed into *E.coli* BL21-DE3 Codon Plus RP cells. Cells were plated onto Luria Broth (LB) plates supplemented with ampicillin and chloramphenicol and incubated at 37°C for 24 hours. Single colonies were added to a small scale culture consisting of 5 ml LB supplemented with 100 µg/ml ampicillin, 68 µg/ml chloramphenicol, and shaken overnight at 250 rpm at 37°C. The small scale culture was transferred into 1L of

autoclaved minimal media composed of 1 μM CaCl_2 , 10 μM $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 10 ml/L 40% glucose, 10 ml/L MEM vitamin solution, 100 mg/L ampicillin, 34 mg/L chloramphenicol and grown at 37°C while shaking at 250 rpm until the OD600 reached 0.8. The culture was inoculated with 1mM IPTG and induction was continued 12-16 hours overnight. The cells were harvested by centrifugation at 4°C at 10,000 g for 15 mins. The pellet containing inclusion bodies was suspended in 50 ml Lysis buffer (70 mM Tris-HCl, 300 mM NaCl, pH 7.8). 2 mM β -mercaptoethanol (BME), 0.5 ml phenylmethanesulfonylfluoride (PMSF) and a mix of 0.2 mg/ml lysozyme, 0.02 mg/ml DNase, and 0.02 mg/ml RNase were added together with 5 mM magnesium acetate. The mixture was tumbled at 4°C for 30 mins and subsequently pulse sonicated for 10 mins (5 sec on, 5 sec off) while keeping on ice. The lysate was centrifuged at 20,000 g for 20 mins. The pellet was dissolved in 35 ml suspension buffer (8 M urea, 25 mM Tris-HCl, 150 mM NaCl, pH 8.0, 0.2% SDS, 4.9 μl BME) and tumbled at room temperature for 30 mins or until dissolved. The solution was centrifuged at 20,000 g for 20 mins at room temperature to remove debris. The supernatant was mixed with 5 ml Ni-NTA resin that was previously equilibrated with 5 column volumes suspension buffer and 2 mM BME. The supernatant-resin-mixture was rotated at room temperature for 3 hours and transferred to the column. The resin was equilibrated using 5 - 10 column volumes of suspension buffer with 2 mM BME and the protein was refolded by a subsequent wash step with 20 column volumes of wash buffer (25 mM Tris-HCl, 200 mM NaCl, pH 8.0, 0.2% SDS, 2mM BME) that did not contain any urea. The detergent was exchanged from SDS to LMPC by washing with 15 column volumes of rinse-exchange buffer (25 mM Tris-HCl, 200 mM NaCl, pH 8.0, 0.2% LMPC, 2mM BME) and the protein was eluted using 250 mM imidazole (ultra-grade), pH 6.5, 0.2% LMPC, 0.5 mM BME.

Reduction of intermolecular disulfide bonds

KCNE3 is prone to form intermolecular dimers which are very difficult to break. I examined which protocol is the most effective to reduce these dimers and under which conditions dimerization is more or less strong. Dimerization was studied under different conditions of pH, temperature, concentrations of reducing agent, incubation time, and location of the cysteine in the protein. Dithiothreitol (DTT) was used as a reducing agent. The tested conditions and their results are presented in the results section of this chapter.

Attachment of the lanthanide-binding tags

Since KCNE3 was prone to persistent dimer formation, remaining disulfide bonds in the protein elution fraction were reduced by adding 5 mM DTT and tumbling overnight at 37°C. The lanthanide-binding tag (MTS-EDTA, MTS-CA-EDTA, or Cys-Ph-TAHA) was dissolved in water to a stock solution of 50 mM. Lanthanide chloride was dissolved in water to a stock solution of 100 mM. The tag was pre-loaded with lanthanide ions at a ratio of 1:2 of tag to lanthanides. The solution was mixed by shaking for 2 hours at room temperature. In the meantime, the purified KCNE3 was desalted to remove DTT. This was accomplished using a size-exclusion column of 1 cm in diameter, filled to about 20 cm with Sephadex G25 resin. The resin was initially equilibrated with 50 ml PIPES buffer (10 mM PIPES, pH 6.5, 0.2% LMPC). The purified protein concentrated to 1 ml (MWCO 10 kDa) was loaded onto the column. The protein was eluted using 15 - 20 ml PIPES buffer while monitoring the absorbance at 280 nm. To prevent recurring dimer formation and for optimal tagging results, the eluted KCNE3 solution was quickly concentrated to ~0.5 ml followed by the determination of the protein concentration. The preloaded tag solution was then added at a ratio of 1:2:4 of protein:tag:lanthanide. The protein solution was tumbled overnight at room temperature. The unbound lanthanide and small-

molecule reaction products were removed by 5 times 7-fold buffer exchange using centrifugal ultrafiltration: repeated dilution to 3.5 ml using 100 mM imidazole, pH 6.5 followed by concentration to 0.5 ml (or to NMR sample volume in the final round). Samples were prepared in 3 mm NMR tubes with 10% D₂O and final protein concentrations of ~0.7 mM and ~4% LMPC.

Mass-spectrometry

Mass-spectrometry measurements using positive ion electrospray were carried out to verify that the tag is bound to the protein and that it is loaded with lanthanide ions. The measurements were carried out on a Waters Synapt hybrid quadrupole/orthogonal access time of flight mass spectrometer (Waters Corp., Milford, MA) coupled to a Waters Acquity UPLC system in the Vanderbilt mass-spectrometry core.

NMR experiments

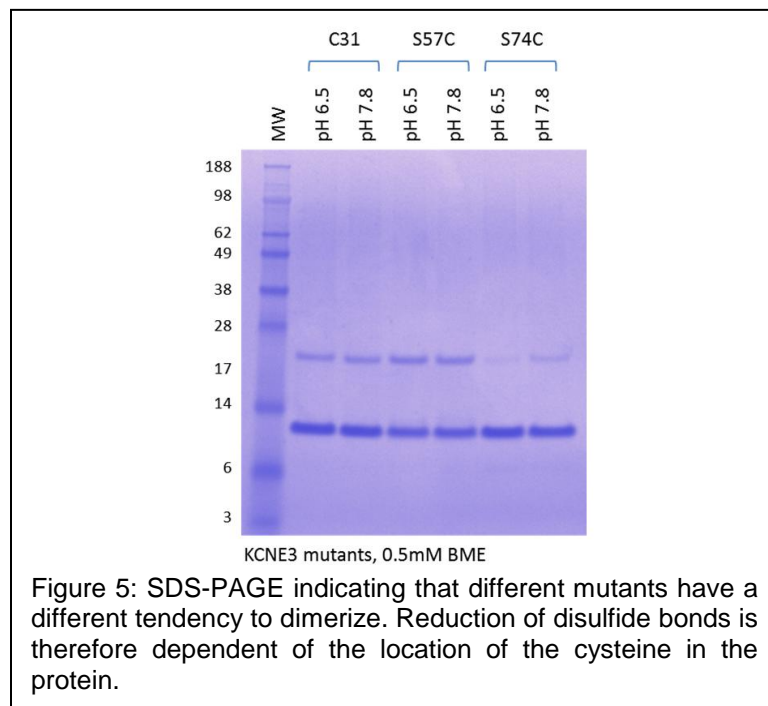
For each experiment, a set of two tagged samples was prepared: (1) KCNE3 tagged with a paramagnetic lanthanide and (2) a diamagnetic lanthanide as a reference. For each of these samples, which had similar protein concentrations around 0.7 mM, a TROSY and an HSQC experiment were carried out. NMR experiments were carried out at 40°C on a Bruker 800 MHz Avance spectrometer equipped with a triple resonance cryoprobe. The pulse programs used were a ¹H-¹⁵N-HSQC and a ¹H-¹⁵N-TROSY-HSQC from the Bruker standard pulse program library (hsqcetf3gpsi and trosytf3gpsi2). Data processing was accomplished using NMRPipe and NMRDraw software and analyzed using Sparky. For extraction of the paramagnetic restraints, especially for interpretation of the PREs, the peak intensities were normalized by the protein concentration. The PRE intensity ratios of the resonances were computed by dividing the Gaussian fitted peak heights of the paramagnetic resonances by the fitted peak heights of the

diamagnetic resonance. RDCs were extracted from the chemical shift difference of the TROSY to the HSQC component from the paramagnetic to the diamagnetic spectra ($=\frac{1}{2}(J + D)$). PCSs were obtained by measuring the chemical shift differences in both the ^1H and ^{15}N dimensions and computing an overall chemical shift difference by $\Delta\delta = \sqrt{(\Delta^{15}\text{N}/10)^2 + (\Delta^{1}\text{H})^2}$ between the paramagnetic and the diamagnetic resonances.

Results

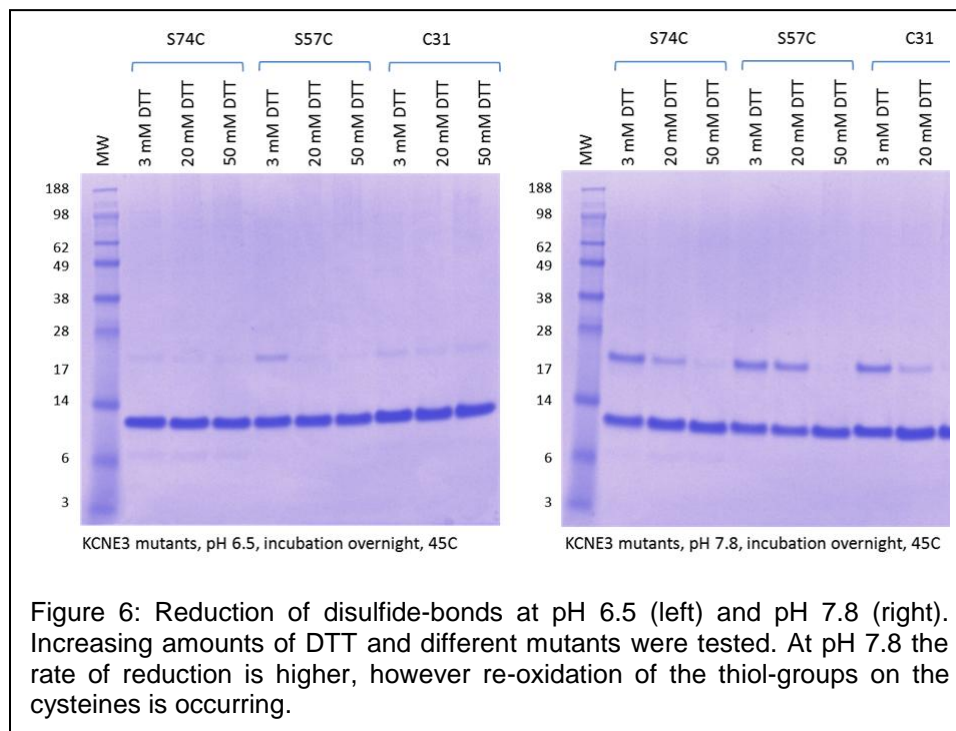
Reduction of intermolecular disulfide bonds is difficult

The strength of KCNE3 dimerization was studied under several conditions to develop a protocol for efficient reduction of the disulfide bonds. Conditions such as pH, temperature, amount of the reducing agent DTT, incubation time, and location of the cysteine in the protein were tested (Figure 5 and Figure 6). The following observations were made, as deduced from gel electrophoresis.



(a) Increasing amounts of DTT were tested: 3 mM, 20 mM, and 50 mM. The higher the concentration of DTT was in the sample, the higher was the reduction power. However, 3-5 mM of DTT have shown to be sufficient to reduce the disulfide bonds for KCNE3.

(b) Three different cysteine mutants were tested: C31, which is the native cysteine in the N-terminal amphiphatic helix, S57C being in the interface region at the



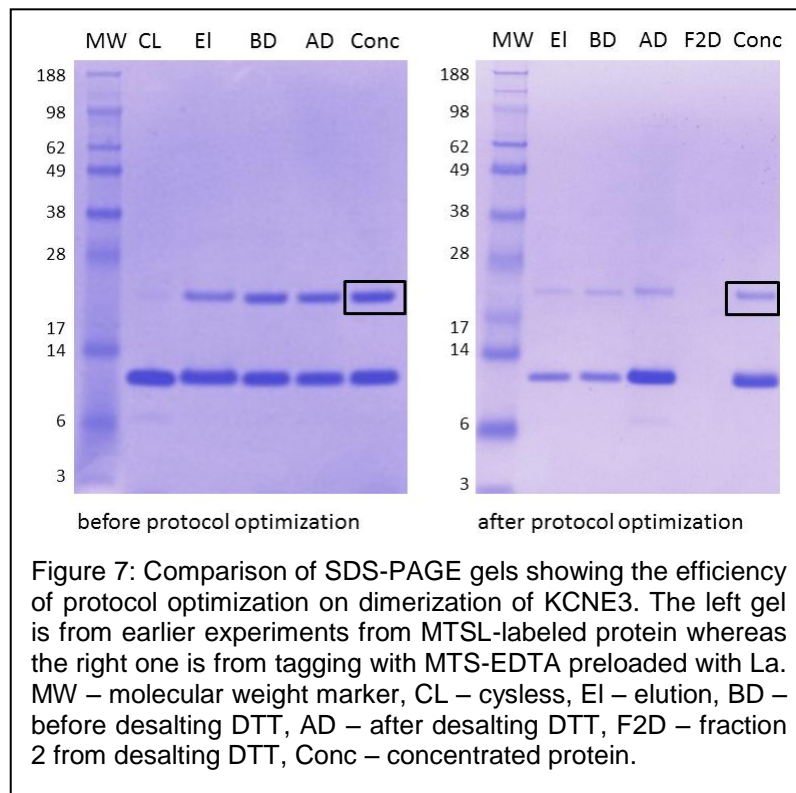
beginning of the trans-membrane span, and S74C which is buried in the membrane. It was found that the strength of dimerization depends on the cysteine mutant (Figure 5), i.e. the location of the cysteine in the protein. S74C in the transmembrane region exhibits the least dimerization and S57C, which is close to the membrane interface, exhibits most dimerization (see also left panel in Figure 6). This is plausible since the location of the cysteine in the protein correlates with its burial or exposure and therefore its availability for cysteines from other proteins to form disulfide bonds.

(c) Three different temperatures were tested: room temperature, 37°C, and 45°C. Higher temperature resulted in faster reduction.

(d) Four different incubation times were tested: no incubation, 1 h, 24 h, and 48 h. It was found that after 1 h most of the disulfide bonds were reduced for the tested mutants in KCNE3 where incubation at 37°C was more effective.

(e) Two different pH values were tested: pH 6.5 (acidic) and pH 7.8 (basic). It was found that reduction is possible at both pH, where the reaction occurs faster at pH 7.8 than at pH 6.5. However, when DTT is depleted at pH 7.8, re-oxidation of the cysteines occurs, which does not occur at pH 6.5. The time until re-oxidation at basic pH depends on the amount of DTT present, the temperature, and the location of the cysteine in the protein.

(f) The extent of dimerization was also tested as a function of the amount of resin used for purification. For protein from 2 g of wet cell mass, 2 ml and 6 ml of resin were



used. No influence of the amount of resin on the amount of dimerization could be observed (data not shown).

(g) It was found that DTT needs to be prepared fresh. Storage in aqueous buffer decreases its reduction power, due to oxidation. The half-life of DTT in aqueous buffers is about 40 h [31] at pH 6.5.

To summarize the findings about reducing persistent intermolecular disulfide bonds of KCNE3, we have settled on adding 5mM DTT to the purified protein at pH 6.5, followed by incubation overnight at 37°C. DTT is subsequently removed by a size-exclusion column before KCNE3 is tagged with the lanthanide-binding tag. Figure 7 shows a comparison of SDS-PAGE gels from before and after protocol optimization to increase the reduction of disulfide bonds.

Mass-spectrometry measurements

Mass-spectrometry was used to verify that the protein is fully tagged with the EDTA-based metal chelators and that the lanthanide ion is bound to the tag. It was found that KCNE3-S57C is fully tagged, indicated by a missing peak for untagged KCNE3 at 12,943 Da (Figure 8). Mass-spectrometry also indicates that the sample does not contain any dimer of KCNE3 (data not shown) and that the C-terminal isoleucine is typically cleaved off even though there are small populations where this isoleucine is present (at 5-20% maximal intensity using the MTS-EDTA and MTS-CA-EDTA tag). This population is not present in the samples with the Cys-Ph-TAHA tag. The reason for the isoleucine being cleaved is currently unknown, as none of the proteases that could potentially cleave after the preceding methionine (chymotrypsin, thermolysin, cyanogen bromide) are found in *E.coli*.

For the samples with the MTS-EDTA or MTS-CA-EDTA tags bound the largest populations exhibit a bound lanthanide ion (100% maximal intensity for the population

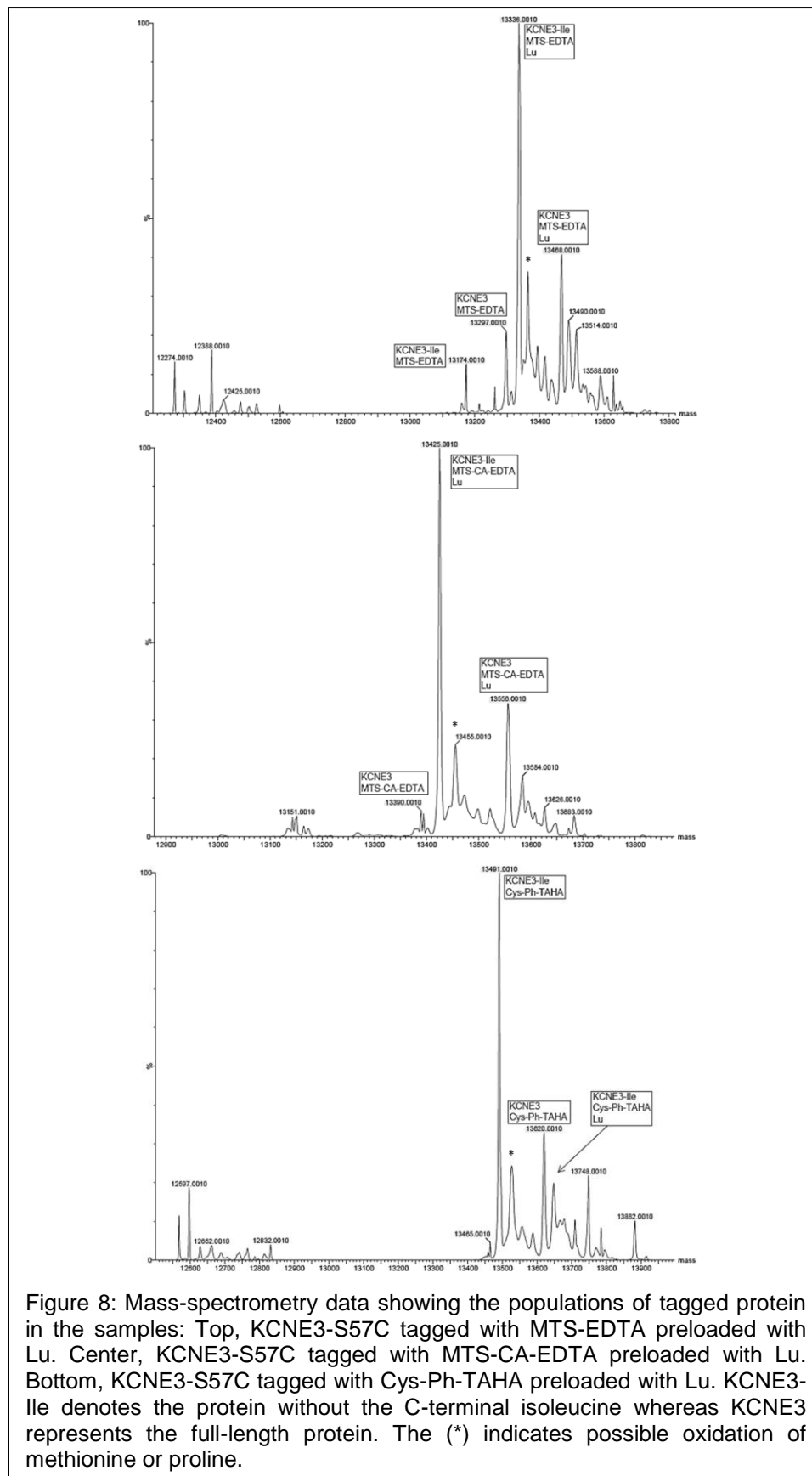


Figure 8: Mass-spectrometry data showing the populations of tagged protein in the samples: Top, KCNE3-S57C tagged with MTS-EDTA preloaded with Lu. Center, KCNE3-S57C tagged with MTS-CA-EDTA preloaded with Lu. Bottom, KCNE3-S57C tagged with Cys-Ph-TAHA preloaded with Lu. KCNE3-Ile denotes the protein without the C-terminal isoleucine whereas KCNE3 represents the full-length protein. The (*) indicates possible oxidation of methionine or proline.

without C-terminal Ile, and 35-50% maximal intensity for the population with C-terminal

lle). For the samples with Cys-Ph-TAHA bound, only a fraction of the tagged protein contains a bound lanthanide (15-20% maximal intensity, Figure 8). We argue that the lanthanide ion “flies off” during some of the electro-spray experiments since the percentage of tagged protein containing bound the lanthanide ion is not reproducible.

NMR spectroscopy

A spectral comparison is shown in Figure 9 where the spectra of WT KCNE3 and of KCNE3-S57C tagged with MTS-EDTA and loaded with Lu are overlaid. Overall, the spectra are similar, even though peak shifts are noticeable that are larger in the peripheral region of the spectra. These shifts can be attributed to the introduction of the tag that might affect the structure and/or dynamics of KCNE3 as it is sitting in the micelle. Similar results are found in Figure 10 that shows an overlay of untagged KCNE3-S57C in black versus protein tagged with MTS-EDTA preloaded with Lu in red. Notably, there are some changes in the spectrum that indicate a structural rearrangement of the protein due to the attachment of the tag. The shifting references are not only in the vicinity of the tagging site but distributed over different regions in the protein. It can be argued that the flexible cytosolic and periplasmic domains wrap around the micelle surface in a different manner, thereby affecting residues that are not in direct vicinity of the tagging site.

A spectral comparison of the protein with the different tags preloaded with Lu also points to some subtle changes in the spectra (Figure 11). These changes, however, are smaller than the ones originating from the attachment of the tag to the protein and mostly occur close to the C-terminus. The spectra of KCNE3 with the three different tags are overall very similar, indicating a similar structural rearrangement of residues around the tagging site of KCNE3.

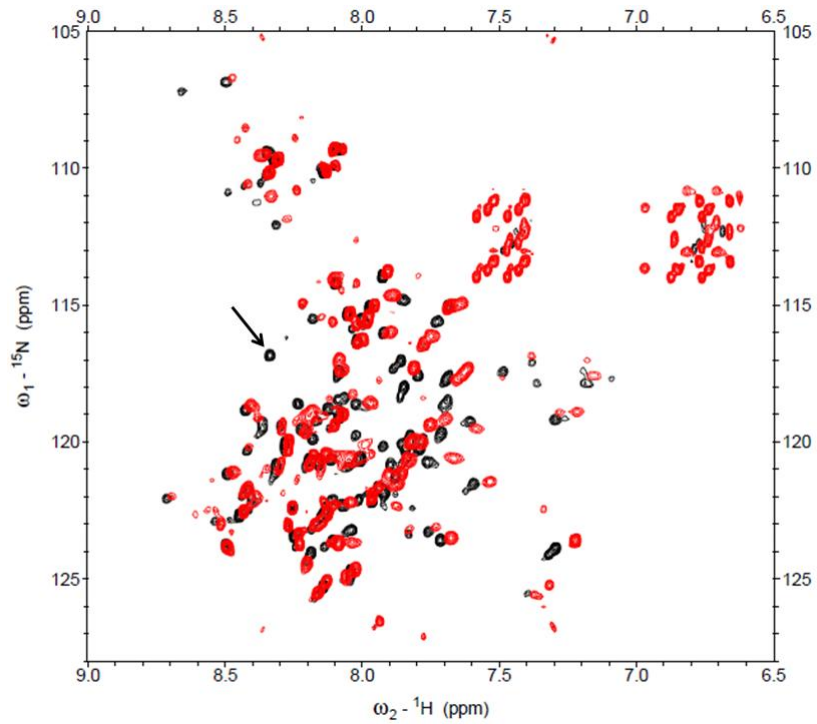


Figure 9: Spectral overlay of WT KCNE3 (black) with KCNE3-S57C tagged with MTS-EDTA loaded with Lu (red). The arrow indicates S57 which is replaced by cysteine for the attachment of the lanthanide binding tag.

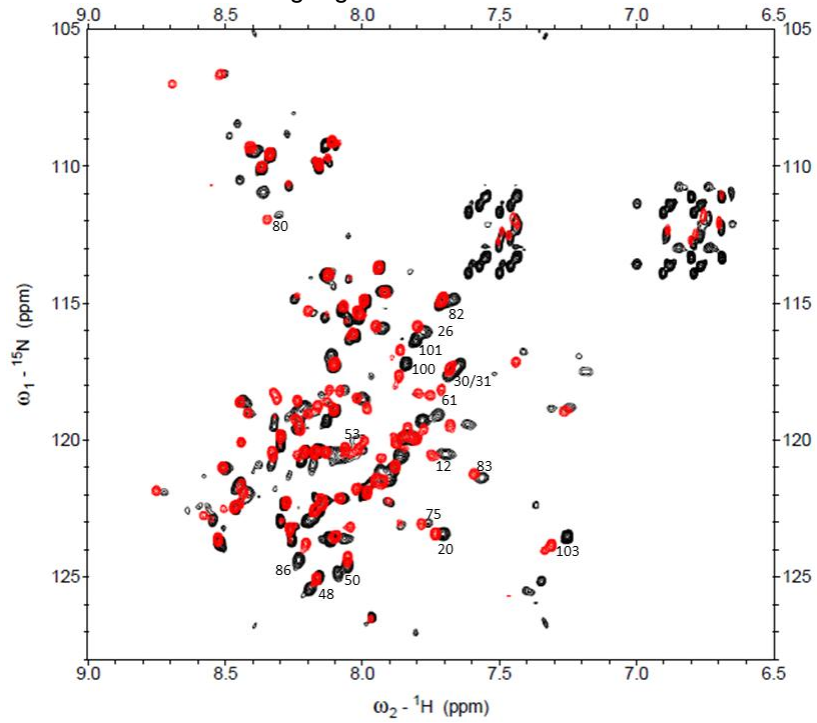
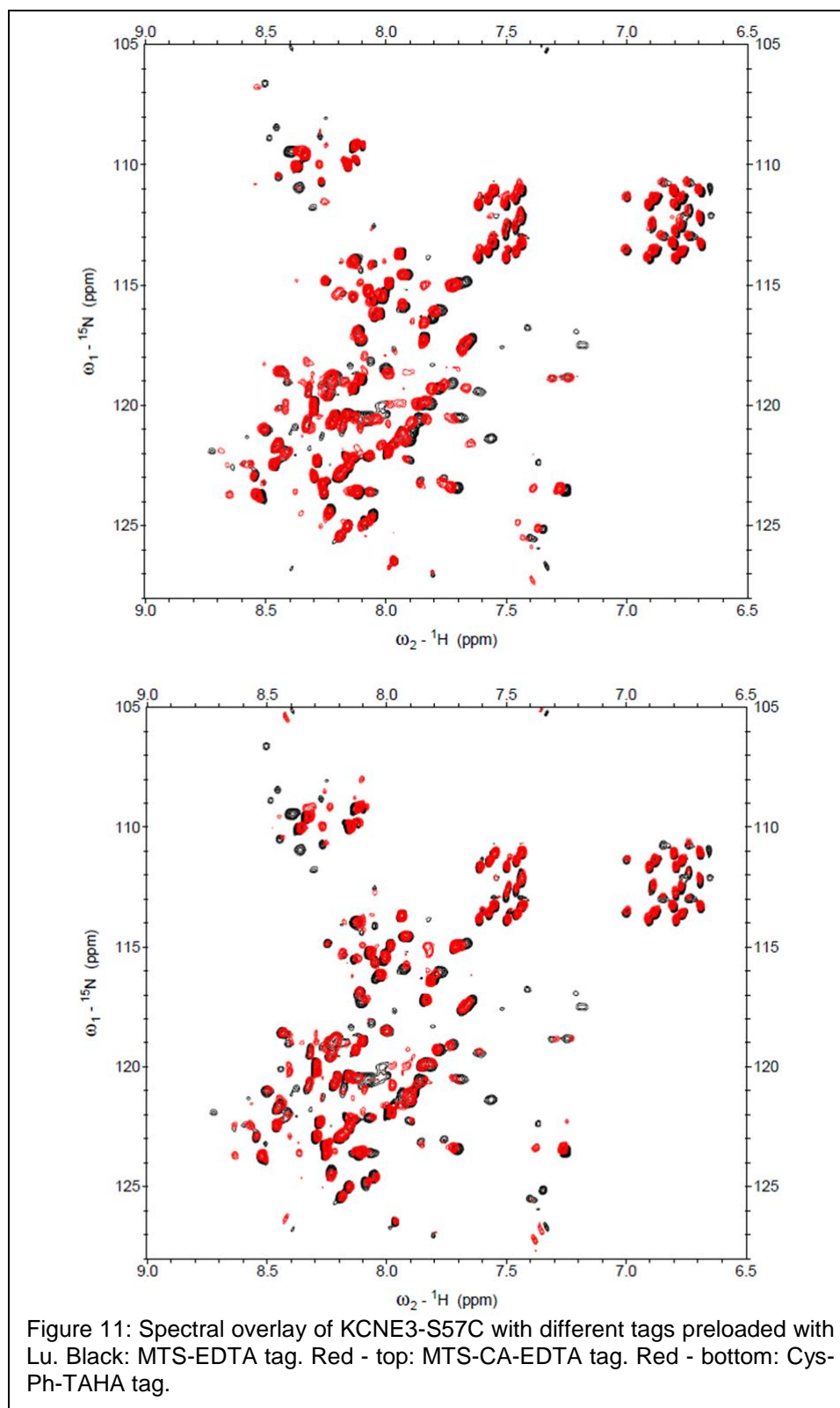


Figure 10: Spectral comparison of untagged KCNE3-S57C (red) with protein tagged with MTS-EDTA preloaded with Lu (black).

Overlays of paramagnetic (Yb) vs. diamagnetic (Lu) KCNE3-S57C tagged with the different tags are shown in Figure 12. Overall, the spectra are very similar, even though more peaks appear to be broadened in the Cys-Ph-TAHA tagged spectrum. This is likely due to a different contour level, which is difficult to reproduce due to the different protein concentrations in various samples.

The peak intensities, coupling differences as well as chemical shifts from the diamagnetic and paramagnetic spectra were extracted to compute PRE intensity ratios, RDCs, and PCSs. Figure 13 shows the PCSs extracted from the chemical shift differences between diamagnetic and paramagnetic references in the left panel and the ratios of the diamagnetic vs. paramagnetic peak intensities, which can be used to extract PREs, in the right panel. It can be seen that even though there are structural changes in the protein induced by the different structures of the tags, the overall pattern of the restraints vs. the residue number remains similar. The PCSs are missing around the tagging site at S57C due to PRE effects and are largest around residues 45, 70 and at the N-terminus. The PRE intensity ratio's are smallest or zero close to the tagging site due to extensive line-broadening. Small intensity ratio's can also be observed for the N-terminus around residue 10 whereas large amplitudes are seen around residue 20 and the C-terminus.

Figure 14 shows the histograms of the RDCs (left panel) and the PCSs (right panel). RDCs are consistently observed in the range from -20 Hz to 20 Hz and PCSs range from -0.06 ppm to 0.06 ppm. The shape of the RDC histograms represent the x, y, and z-components (the most populated, the smallest, and the largest components, respectively) of the susceptibility tensor of the whole molecule whereas the shape of the PCS histograms represent these components for the susceptibility tensor of the metal ion. Figure 15 shows the RDCs vs. the residue number for KCNE3-S57C tagged with the three different tags. This data was used to create the histograms in Figure 14.



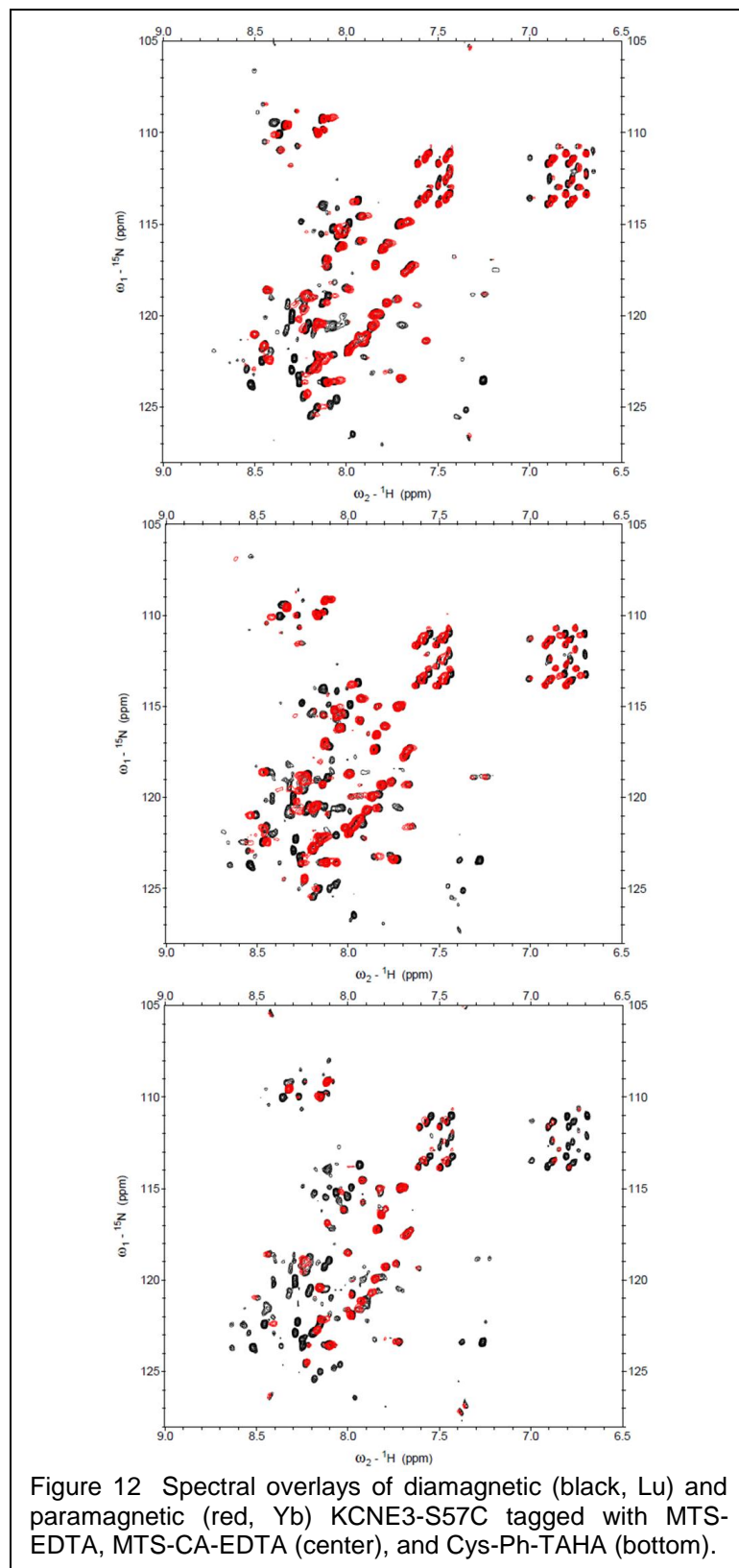


Figure 12 Spectral overlays of diamagnetic (black, Lu) and paramagnetic (red, Yb) KCNE3-S57C tagged with MTS-EDTA, MTS-CA-EDTA (center), and Cys-Ph-TAHA (bottom).

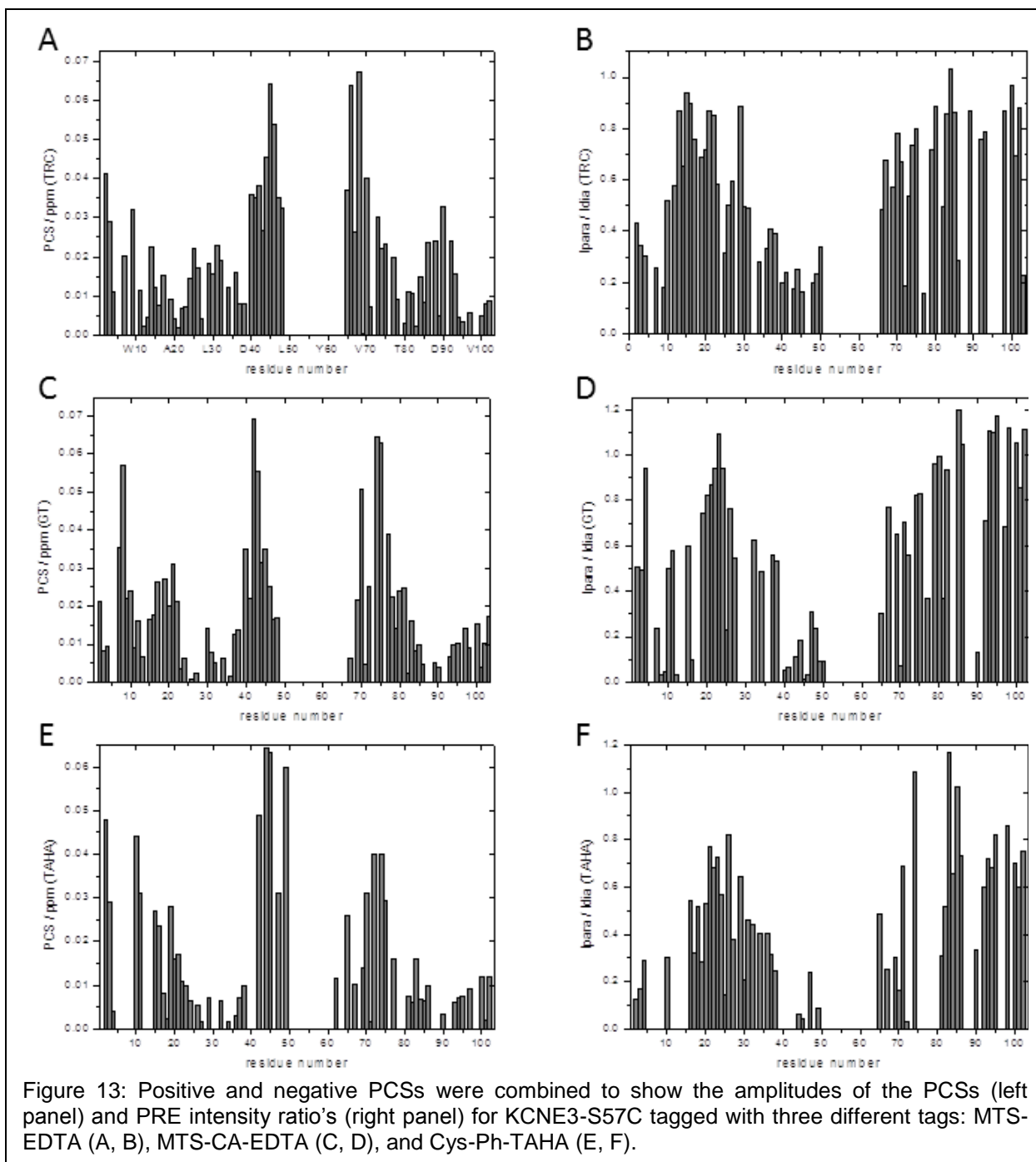


Figure 13: Positive and negative PCSs were combined to show the amplitudes of the PCSs (left panel) and PRE intensity ratio's (right panel) for KCNE3-S57C tagged with three different tags: MTS-EDTA (A, B), MTS-CA-EDTA (C, D), and Cys-Ph-TAHA (E, F).

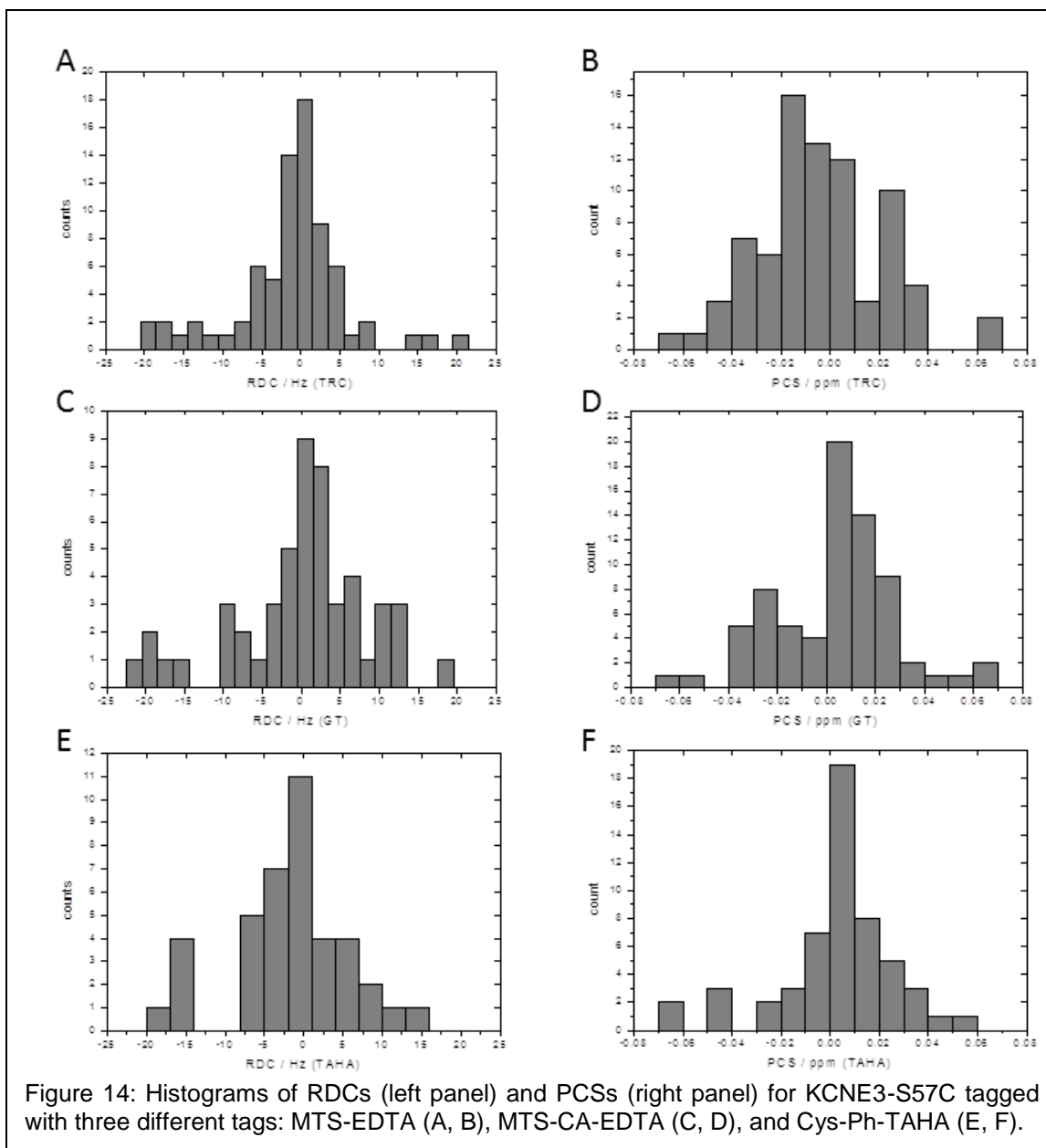
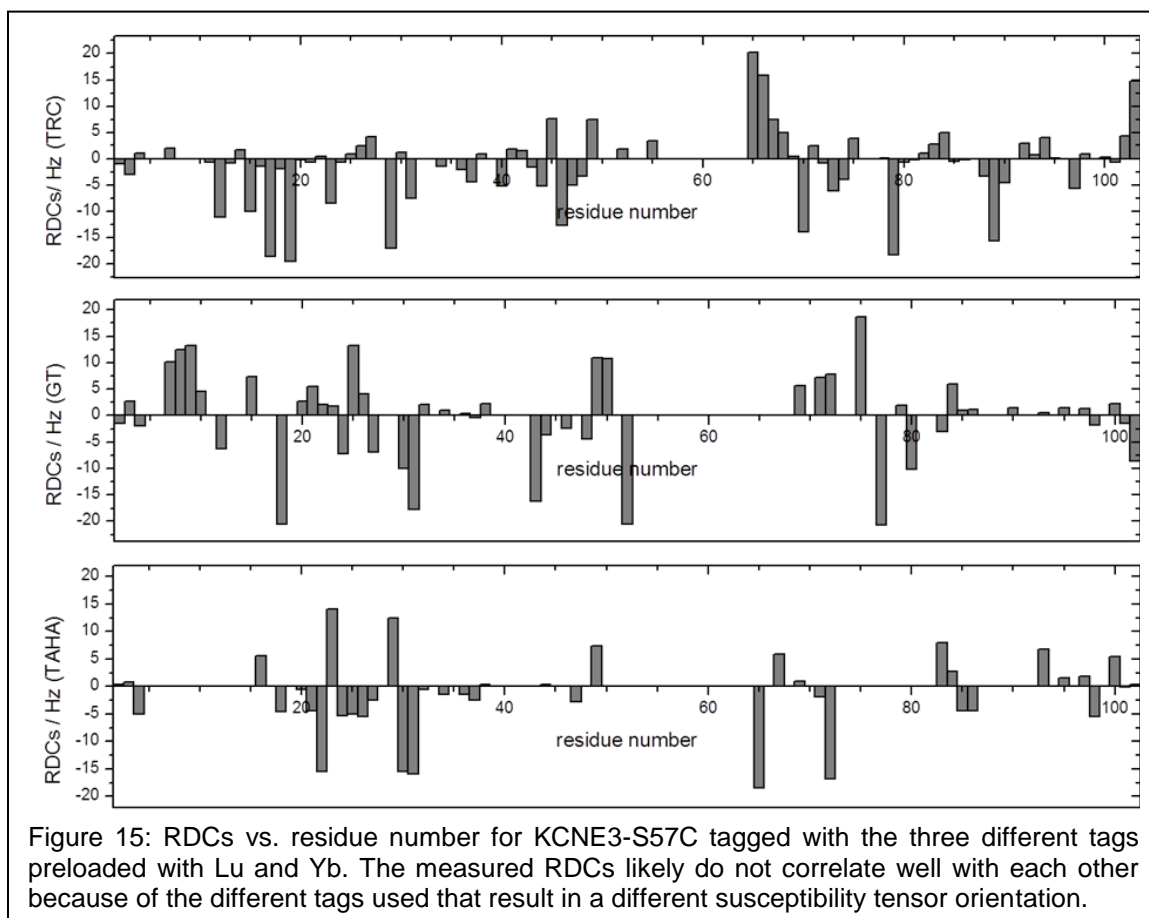


Figure 14: Histograms of RDCs (left panel) and PCSs (right panel) for KCNE3-S57C tagged with three different tags: MTS-EDTA (A, B), MTS-CA-EDTA (C, D), and Cys-Ph-TAHA (E, F).



Discussion

Reduction of intermolecular disulfide bonds

The results indicate that the reduction power of DTT to reduce disulfide bonds is dependent on pH, temperature, amount of DTT present, incubation time, and the location of the cysteine in the protein. Most of these variables are dependent on each other: at basic pH (pH 7.8 in our case) the reduction reaction occurs faster but the disulfide bonds can be re-oxidized when the reducing agent in the sample is depleted. At acidic pH (pH 6.5) the reduction reaction occurs slower but there is less danger of re-oxidation. Increasing temperature as well as increasing amounts of DTT accelerate the reduction, and in case of basic pH, large amounts of DTT decelerate re-oxidation because of the increased time span until DTT is depleted. The time interval in which

these reactions happen is also dependent on the location of the cysteine in the protein. For a less accessible cysteine (i.e. the deeper it is buried in the membrane or micelle) the dimerization occurs slower but it also takes longer for the disulfide bond to be reduced.

These findings in terms of pH and temperature dependence as well as the amount of DTT are not dependent on the protein. In contrast, the dependencies on incubation time and the location of the cysteine are protein dependent and need to be tested for each system to achieve optimal reduction.

It should be noted that the amount of dimer observed as judged by gel electrophoresis can be misleading. The denaturing agent lithium dodecyl sulfate (LDS) present in the SDS-Page loading buffer unfolds the protein and therefore exposes the cysteines making them available for disulfide bond formation. In such cases LC-MS data might be more accurate in detecting dimerization under real sample conditions. The mass spectrometry data on the NMR samples (for which spectra shown in Figures 9-15) confirmed that no dimerization was detected even though the SDS-Page showed a minimal amount of dimer.

Removing reducing agent before tagging

For MTSL-labeling, which is typically carried out in our lab and which was used as preliminary experiments to become familiar with tagging experiments, we use a standard PD10 size-exclusion column from GE Healthcare. However, monitoring the absorbance at 280 nm, it was found that the overlap between the protein peak and the MTSL peak is too large to safely assume that the small molecule (in this case MTSL) is completely removed from the sample (data not shown). This would be similar for removing DTT before tagging KCNE3 with the metal chelator. If trace amounts of DTT are left in the sample, the free thiol groups of DTT are also accessible for the reaction

with the tagging reagent and can, in case the metal chelator is not available in high concentrations, lead to insufficient labeling. This does not pose a problem for MTSL-labeling where complete labeling can be achieved by providing sufficient MTSL which is readily available through purchase. For MTS-CA-EDTA and Cys-Ph-TAHA that are manually synthesized, we tried to completely remove DTT in order to supplement the sample with as little tag as possible to achieve complete labeling. Therefore, a size-exclusion column filled 20 cm with Sephadex G25 has proven to be efficient in removing DTT before reacting KCNE3 with the metal chelators. We refrained from using Ni-NTA metal ion affinity resin for removing DTT from the samples because of the potential for the chelating tags to bind Ni(II).

Mass-spectrometry

Mass-spectrometry has shown that KCNE3 is fully tagged when working with all three tags: MTS-EDTA, MTS-CA-EDTA, and Cys-Ph-TAHA. The fact that the C-terminal Ile is cleaved in a large number of cases is not problematic, however the restraints extracted for this residue are probably incorrect since the percentage of cleavage is not exactly reproducible between samples. The mechanism of cleavage is currently unknown as none of the possible proteolytic cleavage agents (chymotrypsin, thermolysin, cyanogen bromide) occur in *E.coli*.

It was found that for KCNE3 tagged with Cys-Ph-TAHA most of the protein does not contain a lanthanide ion. Repeating these experiments with different tags showed that in several cases the lanthanide ion is not bound to the tag attached to the protein. However, the NMR data clearly indicates the effects of a paramagnetic metal ion present close to the protein. Mass-spectrometry data of the same samples showed that there are neither free lanthanide ions, nor lanthanide ions bound the free tag present in the sample (data not shown). One possibility is, that the lanthanide ion “flies off” during the

mass-spectrometry measurements, which is commonly seen for metal ions (Dr. Wade Calcutt, personal communication).

Mass-spectrometry data also indicates that there is no KCNE3 dimer population in these samples even though gel electrophoresis displays a faint dimer band. The cystless mutant does not exhibit any dimer band on SDS-PAGE indicating that the dimer has to be formed via a disulfide bond and not through hydrophobic interactions. One explanation is the different conditions under which the gel is run compared to the conditions under which NMR spectroscopy and mass-spectrometry are carried out. The latter definitely occur under more native-like conditions whereas for gel electrophoresis the protein is unfolded using a sample buffer containing denaturing reagent such as lithium dodecyl sulfate (LDS) at pH 8.4. Unfolding the protein makes the cysteine more accessible to bind to other KCNE3 molecules present in the sample and it would therefore not be surprising to have a small KCNE3 dimer population present that is observed as a faint dimer band on SDS-PAGE. Furthermore, our experiments indicate that at basic pH (the commercially available loading buffer has pH 8.4) free cysteines are more amenable to re-oxidation, i.e. formation of a dimer is more favorable than at acidic pH.

To summarize, mass-spectrometry is a suitable tool to verify (1) that the tag is bound to KCNE3; (2) the absence of dimer in the sample; and (3) the absence of free lanthanide ions and lanthanide ions bound to free tag in the protein sample indicating no impurities leading to perturbed NMR restraints due to free paramagnetic metal ions. In some cases the lanthanide tag attached to the protein is not loaded with lanthanide ion, however, it is expected that the lanthanide ion “flies off” during the measurements.

NMR spectroscopy

The changes in the protein structure upon attachment of the tag, as seen in Figure 10, are not only present in the vicinity of the tagging site. Chemical shift changes in the vicinity of the tag can be due to small back-bone and side-chain re-orientations and the presence of the tag, all of which have a small influence on the electronic environment of the amide bonds. Additionally, the small changes between KCNE3 tagged with different tags (Figure 11) originate in differences between the tags itself, their different sizes, and different coordination chemistries. All of these factors minutely affect the protein resonances in the spectra.

The intensity ratios available to extract PREs are smallest for residues close to the tagging site at S57C. Residues at the N-terminus close to W10 also exhibit small intensity ratio's indicating a proximity to the metal ion. Large intensity ratio's are observed around residues 25 and at the C-terminus. The profiles of KCNE3-S57C tagged with all three tags agree with each other. For the PRE intensity ratios the correlation coefficient of MTS-EDTA vs. MTS-CA-EDTA is 0.63 and for MTS-EDTA vs. Cys-Ph-TAHA is 0.66. Even though this is not a perfect correlation possibly due to the flexibility of the protein in the micelle combined with the flexibility of the tags, it is an encouraging result showing that the use of different tags yields similar restraints.

The PCSs in Figure 13 converge to a similar pattern for all three tags. The largest PCS amplitudes are observed close to the tagging site around residues 45, 70, and close to the N-terminus around residue 10. For residues close to the tagging site PCSs are largest because of the r^{-3} distance dependence of PCSs. Considering the angular dependence, PCSs are largest for an angle θ enclosing proton, metal-ion, and tensor z-axis, equal to zero. This means that if the amide proton lies on or close to the tensor z-axis, PCSs are largest. Observing that residues close to the N-terminus have large PCSs, it can be argued that these residues are winding back to S57C on the

micelle surface, which is also the entry site of the transmembrane helix into the micelle where the tag is attached.

For RDCs a quantitative interpretation of RDC amplitudes vs. residue number remains difficult. Quantitative statements are easiest to make and verify when a structural model is available. The fit of the RDCs to that model and calculation of a Q-value can evaluate the quality of the RDC data. The histogram of the RDCs can provide the principal components of the alignment tensor of the whole complex, which are approximately $D_{xx} \approx 1$ Hz, $D_{yy} \approx -20$ Hz, $D_{zz} \approx 18$ Hz for all three tags. From the histogram of the PCSs the principal components of the alignment tensor of the metal ion can be extracted, which are roughly $\delta_{xx} \approx 0.03$ ppm, $\delta_{yy} \approx -0.07$ ppm, $\delta_{zz} \approx 0.07$ ppm for all three tags. Since these two tensors differ by the diamagnetic susceptibility tensor, the shapes of these histograms do not need to be similar. Additionally, the tensors might be influenced by the coordination chemistry of the metal ion.

It should be noted that the extracted paramagnetic restraints on KCNE3 have to be confirmed by repeating the experiments which will also provide error bars to the measurements.

Conclusion and future directions

The studies described in this chapter are a good starting point for obtaining paramagnetic restraints on KCNE3. It has been established that the method as such is useful and efficient in obtaining the restraints. After the initial protocol has been established, the use of different metal ions will result in different restraints: paramagnetic metal ions with isotropic magnetic susceptibility will yield PREs only, whereas paramagnetic metal ions with anisotropic magnetic susceptibility will yield PREs, RDCs, and PCSs. Furthermore, paramagnetic metal ions have different magnetic susceptibility tensors and therefore yield restraints that are independent of each other.

Even though paramagnetic tagging is a very useful and efficient method for obtaining restraints, initial efforts of sample preparation and verification should not be underestimated. The challenge is to understand the interactions of the different components (protein, small molecule tag, metal ion) with each other and with the interacting media during sample preparation *on a molecular level*. For instance it remains unknown how loaded EDTA-tags would interact with Ni-NTA during purification. Knowing and understanding the sample preparation protocol in such detail is almost indispensable in preparing high-quality samples for NMR studies where the protein is monomeric, fully tagged, fully loaded with metal ions, and without impurities, such as additional tag or metal ions present. Studying these processes, however, requires much time and effort and the first step towards this understanding is described in this chapter.

In conclusion, a protocol has been developed that establishes the measurement of paramagnetic restraints on KCNE3 in LMPC micelles. The NMR spectra show that the cysteine mutation and the attachment of different tags result in small changes in the spectra which likely result in small structural rearrangements of the protein. Several factors interfere with efficient and reproducible measurement of paramagnetic restraints: (1) the flexibility of the single transmembrane span protein with long loop regions allows for structural rearrangements on the micelle surface; (2) this flexibility is maintained by the micelle environment which is less restrictive towards the protein fold than a bicelle system; (3) since KCNE3 contains only a single transmembrane spanning helix, its 3D fold is less restricted and therefore the electronic environment around the nuclei are more easily influenced by molecules not belonging to the protein; (4) four different components are present in the sample: flexible KCNE3, LMPC, lanthanide binding tag, and lanthanide, additionally to the sample buffer used. All components influence each other and are also influenced by the media used for sample preparation.

As future studies, the tagging will be repeated on KCNE3-S57C with the three tags to obtain a standard deviation of the measured restraints. Furthermore, tagging on the native C31 and S74C will be carried out to gather additional restraints for structure calculations. The paramagnetic restraints will be used and verified for structure calculations on KCNE3. Since the structure of KCNE3 will be determined in the Sanders lab using conventional methods, a comparison of these two structures – in bicelles and micelles – as well as a comparison of the usefulness of the restraints, will be carried out.

References

- [1] T. Jespersen, M. Grunnet, S.P. Olesen, The KCNQ1 potassium channel: from gene to physiological function, *Physiology (Bethesda)*, 20 (2005) 408-416.
- [2] D. Peroz, N. Rodriguez, F. Choveau, I. Baro, J. Merot, G. Lousouarn, Kv7.1 (KCNQ1) properties and channelopathies, *J Physiol*, 586 (2008) 1785-1789.
- [3] W.D. Van Horn, C.G. Vanoye, C.R. Sanders, Working model for the structural basis for KCNE1 modulation of the KCNQ1 potassium channel, *Curr Opin Struct Biol*, 21 (2011) 283-291.
- [4] A. Lundby, G.N. Tseng, N. Schmitt, Structural basis for K(V)7.1-KCNE(x) interactions in the I(Ks) channel complex, *Heart Rhythm*, 7 (2010) 708-713.
- [5] G. Seebohm, N. Strutz-Seebohm, O.N. Ureche, U. Henrion, R. Baltaev, A.F. Mack, G. Korniychuk, K. Steinke, D. Tapken, A. Pfeufer, S. Kaab, C. Bucci, B. Attali, J. Merot, J.M. Tavaré, U.C. Hoppe, M.C. Sanguinetti, F. Lang, Long QT syndrome-associated mutations in KCNQ1 and KCNE1 subunits disrupt normal endosomal recycling of IKs channels, *Circ Res*, 103 (2008) 1451-1457.
- [6] A. Krumerian, X. Gao, J.S. Bian, Y.F. Melman, A. Kagan, T.V. McDonald, An LQT mutant minK alters KvLQT1 trafficking, *Am J Physiol Cell Physiol*, 286 (2004) C1453-1463.
- [7] Y.H. Chen, S.J. Xu, S. Bendahhou, X.L. Wang, Y. Wang, W.Y. Xu, H.W. Jin, H. Sun, X.Y. Su, Q.N. Zhuang, Y.Q. Yang, Y.B. Li, Y. Liu, H.J. Xu, X.F. Li, N. Ma, C.P. Mou, Z. Chen, J. Barhanin, W. Huang, KCNQ1 gain-of-function mutation in familial atrial fibrillation, *Science*, 299 (2003) 251-254.
- [8] I. Splawski, J. Shen, K.W. Timothy, M.H. Lehmann, S. Priori, J.L. Robinson, A.J. Moss, P.J. Schwartz, J.A. Towbin, G.M. Vincent, M.T. Keating, Spectrum of

mutations in long-QT syndrome genes. KVLQT1, HERG, SCN5A, KCNE1, and KCNE2, *Circulation*, 102 (2000) 1178-1185.

- [9] L. Bianchi, Z. Shen, A.T. Dennis, S.G. Priori, C. Napolitano, E. Ronchetti, R. Bryskin, P.J. Schwartz, A.M. Brown, Cellular dysfunction of LQT5-minK mutants: abnormalities of IKs, IKr and trafficking in long QT syndrome, *Hum Mol Genet*, 8 (1999) 1499-1507.
- [10] W. Wang, J. Xia, R.S. Kass, MinK-KvLQT1 fusion proteins, evidence for multiple stoichiometries of the assembled IsK channel, *J Biol Chem*, 273 (1998) 34069-34074.
- [11] T. Tzounopoulos, H.R. Guy, S. Durell, J.P. Adelman, J. Maylie, min K channels form by assembly of at least 14 subunits, *Proc Natl Acad Sci U S A*, 92 (1995) 9593-9597.
- [12] T.J. Morin, W.R. Kobertz, Counting membrane-embedded KCNE beta-subunits in functioning K⁺ channel complexes, *Proc Natl Acad Sci U S A*, 105 (2008) 1478-1482.
- [13] H. Chen, L.A. Kim, S. Rajan, S. Xu, S.A. Goldstein, Charybdotoxin binding in the I(Ks) pore demonstrates two MinK subunits in each channel complex, *Neuron*, 40 (2003) 15-23.
- [14] K.W. Wang, S.A. Goldstein, Subunit composition of minK potassium channels, *Neuron*, 14 (1995) 1303-1309.
- [15] K. Nakajo, M.H. Ulbrich, Y. Kubo, E.Y. Isacoff, Stoichiometry of the KCNQ1 - KCNE1 ion channel complex, *Proc Natl Acad Sci U S A*, 107 (2010) 18862-18867.
- [16] J. Barhanin, F. Lesage, E. Guillemare, M. Fink, M. Lazdunski, G. Romey, K(V)LQT1 and IsK (minK) proteins associate to form the I(Ks) cardiac potassium current, *Nature*, 384 (1996) 78-80.
- [17] M.C. Sanguinetti, M.E. Curran, A. Zou, J. Shen, P.S. Spector, D.L. Atkinson, M.T. Keating, Coassembly of K(V)LQT1 and minK (IsK) proteins to form cardiac I(Ks) potassium channel, *Nature*, 384 (1996) 80-83.
- [18] B.C. Schroeder, S. Waldegger, S. Fehr, M. Bleich, R. Warth, R. Greger, T.J. Jentsch, A constitutively open potassium channel formed by KCNQ1 and KCNE3, *Nature*, 403 (2000) 196-199.
- [19] M. Grunnet, T. Jespersen, H.B. Rasmussen, T. Ljungstrom, N.K. Jorgensen, S.P. Olesen, D.A. Klaerke, KCNE4 is an inhibitory subunit to the KCNQ1 channel, *J Physiol*, 542 (2002) 119-130.

- [20] Y.F. Melman, S.Y. Um, A. Krumerman, A. Kagan, T.V. McDonald, KCNE1 binds to the KCNQ1 pore to regulate potassium channel activity, *Neuron*, 42 (2004) 927-937.
- [21] X. Xu, M. Jiang, K.L. Hsu, M. Zhang, G.N. Tseng, KCNQ1 and KCNE1 in the IKs channel complex make state-dependent contacts in their extracellular domains, *J Gen Physiol*, 131 (2008) 589-603.
- [22] D.Y. Chung, P.J. Chan, J.R. Bankston, L. Yang, G. Liu, S.O. Marx, A. Karlin, R.S. Kass, Location of KCNE1 relative to KCNQ1 in the I(KS) potassium channel by disulfide cross-linking of substituted cysteines, *Proc Natl Acad Sci U S A*, 106 (2009) 743-748.
- [23] J.A. Smith, C.G. Vanoye, A.L. George, Jr., J. Meiler, C.R. Sanders, Structural models for the KCNQ1 voltage-gated potassium channel, *Biochemistry*, 46 (2007) 14141-14152.
- [24] V. Yarov-Yarovoy, D. Baker, W.A. Catterall, Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels, *Proc Natl Acad Sci U S A*, 103 (2006) 7292-7297.
- [25] C. Tian, C.G. Vanoye, C. Kang, R.C. Welch, H.J. Kim, A.L. George, Jr., C.R. Sanders, Preparation, functional characterization, and NMR studies of human KCNE1, a voltage-gated potassium channel accessory subunit associated with deafness and long QT syndrome, *Biochemistry*, 46 (2007) 11459-11472.
- [26] C. Kang, C. Tian, F.D. Sonnichsen, J.A. Smith, J. Meiler, A.L. George, Jr., C.G. Vanoye, H.J. Kim, C.R. Sanders, Structure of KCNE1 and implications for how it modulates the KCNQ1 potassium channel, *Biochemistry*, 47 (2008) 7999-8006.
- [27] C. Kang, C.G. Vanoye, R.C. Welch, W.D. Van Horn, C.R. Sanders, Functional delivery of a membrane protein into oocyte membranes using bicelles, *Biochemistry*, 49 (2010) 653-655.
- [28] A. Dvoretzky, V. Gaponenko, P.R. Rosevear, Derivation of structural restraints using a thiol-reactive chelator, *FEBS Lett*, 528 (2002) 189-192.
- [29] G. Pintacuda, A. Moshref, A. Leonchiks, A. Sharipo, G. Otting, Site-specific labelling with a metal chelator for protein-structure refinement, *J Biomol NMR*, 29 (2004) 351-361.
- [30] D.E. Kamen, S.M. Cahill, M.E. Girvin, Multiple alignment of membrane proteins for measuring residual dipolar couplings using lanthanide ions bound to a small metal chelator, *J Am Chem Soc*, 129 (2007) 1846-1847.

- [31] R. Stevens, L. Stevens, N.C. Price, The Stabilities of Various Thiol Compounds Used in Protein Purifications, *Biochem Educ*, 11 (1983) 70-70.

CHAPTER 3

A Unified Hydrophobicity Scale for Multi-Span Membrane Proteins¹

Introduction

The hydrophobicity of an amino acid is related to its transfer free energy from a polar medium (such as water) to an apolar medium (like the membrane bilayer). While the transfer free energy depends on the chemical nature of the two solvents, it also depends on the structural context of the amino acid residue. An obvious influence is the degree of exposure to the solvent – which functional groups of an amino acid are exposed and hence available for interaction with the solvent. In turn the transfer free energy for a single amino acid will be much different from the transfer free energy in a model peptide, which again will differ from the transfer free energy of an amino acid in the structural context of a folded protein, given the various levels of exposure to the solvent.

This picture of direct interactions between amino acid and solvent is further complicated because the transfer from one medium into another may trigger structural changes in model peptides or proteins that will affect the free energy change of an amino acid.

Given the diverse biophysical properties of membranes and their hydration in various compartments of the cell, the challenge to model these complex systems accurately in experiments, and the different structural contexts in which amino acids are transferred from one medium into another make it impossible to design a single transfer

¹ This chapter has been published in: Koehler, J., et al., *A unified hydrophobicity scale for multispan membrane proteins*. *Proteins*, 2009. **76**(1): p. 13-29.

free energy scale that is optimal under all circumstances. The existence of many transfer free energy scales is a logical consequence.

An older experimental scale is that of Hopp & Woods (HW) [2], who described a hydrophilicity scale to predict antigenic sites on proteins. Goldman, Engelman and Steitz (GES) derived a hydrophobicity scale based on energetic considerations of residues in α -helices [3]. Wimley & White (WW) achieved a significant step forward [4-7] by introducing a three-state scale based on experimental hydrophobicities between water-interface and water-bilayer in model systems.

Although most hydrophobicity scales have been derived experimentally, there are also examples of knowledge-based approaches. A database of known protein structures is utilized to derive free energies from statistics using an inverse Boltzmann relation. Advantages of knowledge-based hydrophobicity scales include flexibility in the choice of the composition of the database (e.g. all folded, multi-span MPs). In turn, the reference point of the scale as well as the absolute size of the hydrophobicity values will match the chosen dataset and accurately describe the characteristics of amino acids in multi-span MPs. In contrast, an experimental scale (e.g. derived for α -helical peptides) will display a bias in absolute size of the hydrophobicity values as well as the reference point when used in the context of folded, multi-span MPs.

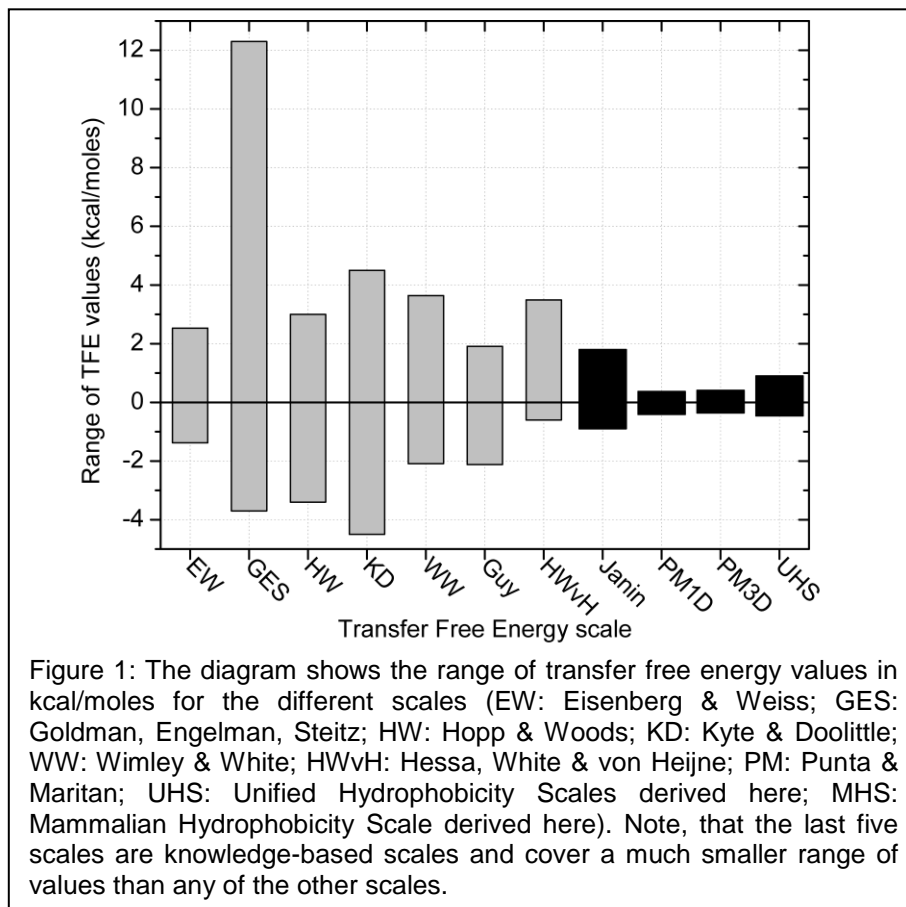
One of the oldest knowledge-based scales was published by Janin [8] who used the known X-ray structures of 22 soluble proteins and derived a scale based on burial versus solvent accessibility of residues. In 2003 Punta & Maritan (PM) [10] derived knowledge-based hydrophobicity scales from two databases containing 118 and 228 trans-membrane α -helices.

Very recently, Senes et. al from the DeGrado laboratory derived a knowledge-based membrane depth-dependent free energy potential from 24 α -helical MPs [12]. The

MPs were centered in the membrane by minimizing the projections of TM helices on the x-y-plane and by setting the center of mass from the non-backbone carbon atoms to zero. Using these structures, the amino acid propensities in 2Å bins were converted into free energies using an inverse Boltzmann relation. The energies were fit using sigmoidal or Gaussian functions (in case of Trp and Tyr) and from these fits the transfer free energies, midpoints, and steepness of the transitions were evaluated for all amino acids. In developing these potentials, the authors assumed that 24 helical MPs used for derivation are a good representation of the MP fold space. Furthermore, it is assumed that the amino acid frequencies in 2Å bins are sufficient to derive a reliable potential and that the fits are in good agreement with the data points. The authors found that Arg and Lys side chains can enter the membrane 4-5Å more deeply than the basic side chains. This is likely caused by the ability of Arg and Lys to snorkel back to the surface since these side chains are longer than the basic ones. Also, their interaction with the negatively charged phospholipid headgroups is more favorable than the repulsion of the latter with basic side chains. It was found that comparing the energy potential at the center of the membrane with the hydrophobicity scale of Eisenberg, with the Wimley & White octanol scale, and with the biological scale of Hessa & von Heijne resulted in a generally good agreement ($R(\text{Eisenberg}) = 0.94$, $R(\text{Wimley}) = 0.78$, $R(\text{Hessa}) = 0.88$). Hessa & von Heijne investigated the effect of symmetric mutations of a TM segment along the membrane bilayer (see below). Senes & DeGrado computationally repeated this experiment using their energy potential and found that the ranges of their energies are much smaller. This is expected since the potential was derived from fully folded structures while Hessa & von Heijne derived their hydrophobicity scale from a protein with at most three TM spans which therefore has a larger interaction surface with the lipids in the bilayer.

It has been common in the past to derive consensus hydrophobicity scales that seek to combine the advantages of several approaches. The scale by Kyte & Doolittle (KD) [9] is based on a variety of experimental observations from the literature [13-16] and uses the display method of Rose et al. [17-18] to detect trans-membrane spans along the protein sequence. Eisenberg et al. (EW) [1] published a consensus hydrophobicity scale derived from five different scales (Nozaki & Tanford, von Heijne & Blomberg, Janin, Chothia, Wolfenden). In 1985 Guy [11] developed a scale based on statistical and experimental results of several studies [8, 13-14, 19-22].

For most of the experimentally derived scales the range of hydrophobicity values is rather large in comparison to the knowledge-based ones (Figure 1). This is expected since experimentally derived scales use mostly model peptides that form α -helices where the residue in question is exposed and other structural context is removed. These



scales capture neatly the nature of the chemical interactions between apolar solvent and amino acid. In contrast, knowledge-based scales derive statistics from multi-span MPs to arrive at a hydrophobicity that might be biologically more relevant in the structural context of intact proteins. In multi-span MPs polar residues are somewhat more likely to occur in the membrane since these side chains can be buried from the interaction with the apolar membrane.

To this end, a remarkable series of experiments has been carried out by Hessa et. al in the von Heijne laboratory [23-24] leading to a 'biological' hydrophobicity scale for α -helical proteins. The authors inserted a so-called H-segments into the two TM span α -helical protein Leader peptidase. The H-segment was designed as a 19-residue stretch flanked by GGPG and a glycosylation site on either side. If the H-segment is inserted into the membrane, one of the two sites is glycosylated, if it is translocated, both sites are glycosylated. Quantifying the fractions of singly versus doubly glycosylated protein on a SDS-Page gel allowed the determination of the equilibrium constant that was converted into apparent transfer free energies. The residues in the H-segment were Ala or Leu with all of the 20 amino acids inserted at varying positions in the membrane. The authors also tested whether the length of the H-segment influences the apparent transfer free energy. For this, the number of Ala and Leu were chosen such that the apparent transfer free energy was kept approximately zero for the inserted segment. It was found that for each Leu removed from the segment, about three Ala have to be added to keep the apparent transfer free energy constant. Changing the flanking residues to charged residues lead to several observations: addition of Asp/Glu or Asn/Gln increased the apparent transfer free energy when inserted at the luminal, but not at the c-terminal end. Furthermore, inserting Arg or Lys decreased the apparent transfer free energy when inserted at the c-terminal but not at the luminal end. This is in agreement with the positive-inside rule which states that in MP structures positively charged residues are

more abundant in cytoplasmic regions than in the periplasm. It was also found that longer TM helices (up to 25 residues) flanked by charged residues are inserted in the same way as shorter helices indicating that the TM helix is treated as one segment during membrane insertion.

Furthermore, it was found that apparent free energy distributions are effective in distinguishing MPs from soluble or secreted proteins, even more so than other hydrophobicity scales. According to how the free energy scale is derived, lower free energies are predicted for single TM spanning proteins than for multi-span MPs. For multi-span MPs it was found that up to 1/4 of the TM helices have a predicted apparent transfer free energy larger than zero. The authors state that this indicates that proper positioning of these helices in the membrane depends on interactions with neighboring TM helices.

The first hydrophobicity scale derived from a completely folded protein structure was recently published by Moon & Fleming [25]. The authors derived a side chain hydrophobicity scale by guanidine HCl unfolding of the β -barrel MP OmpLA. They replaced an Ala residue at the center of the membrane with all of the 20 amino acids and measured guanidine HCl unfolding and refolding curves using Tryptophan fluorescence spectroscopy to extract side chain specific transfer free energies. From the hydrophobicity of the replaced residue and from the determined structure of OmpLA it is known that the side chains face the lipid and not the aqueous barrel interior. The authors also introduced Leu and Arg in a depth-dependent manner and fit a normal distribution to arrive at a energy potential for these two residues. Additionally, with the use of double mutants they demonstrated that the placement of two Arg residues in proximity of each other at the center of the bilayer leads to cooperative effects that decrease the energy penalty observed for placement of these residues in the membrane. The reason for this cooperativity is that the energy required for opening an aqueous bulge in the bilayer is

larger than extending it by the same amount of surface area. It was also shown that the removal of 1 Å² a hydrophobic residue leads to a stability enhancement of 23 cal/mol which is similar to the 24 cal/mol that Chothia found for soluble proteins [26].

Obviously, highly specialized hydrophobicity scales can be derived if assumptions regarding secondary structure (such as separation of α -helices from β -strands) or tertiary structure (such as level of exposure) are made. The ROSETTAMEMBRANE algorithm, for instance, features a knowledge-based potential for folding of α -helical MPs [27-29]. Beuming & Weinstein derived a knowledge-based prediction method to distinguish between the burial and exposure of certain amino acids [30-31]. They used a database of 28 α -helical MPs with a resolution $<4\text{\AA}$ to derive a surface propensity scale. The surface propensities for the 20 amino acids were calculated based on the solvent accessible surface area (SASA) of the residue side chains using a probe size of 1.4 Å for a water molecule and 2.0 Å for a CH₂ group. These SASA values were normalized by a reference value for the SASA of the side chain of a X residue in a GXG tri-peptide. The authors found that in this database the surface of MPs contains more hydrophobic residues (Ala, Ile, Leu, Val) than the interior, that aromatic residues occur more often on the surface, and that the interior is enriched with small residues to ensure proper packing. The number of charged residues in the interior is small but larger than for MP surfaces. The assumption for the derivation of this scale is that the database is sufficiently large to derive a reliable propensity. Moreover, it is assumed that 4Å resolution structures are sufficient in resolution to identify the orientation of the side chains reliably. The authors used this surface propensity scale together with a conservation index to develop the ProperTM algorithm that is able to derive residue properties from a multiple sequence alignment.

The objective of this work, however, is to derive a hydrophobicity scale for multi-span integral MPs with no *a priori* assumptions regarding secondary or tertiary structure

(structural context). This scale measures the likelihood of an amino acid to reside in membrane, transition, or soluble region within a folded protein. The scale can be used as absolute reference energy for folded multi-span MPs, applied in protein structure elucidation, for example as input for machine learning techniques for the prediction of secondary structure, trans-membrane spans, or other structural features, or as reference energy for MP folding simulations or design. The scale is optimized to describe the characteristics of both α -helical proteins and β -barrels equally well. One application of the scale is the prediction of trans-membrane spans from amino acid sequence only, a method that could be applied to detect integral MPs in ORFs of newly sequenced genomes where no structural information is available, or in the early stages of a MP structure determination project. In addition, the identification of a MP or membrane spanning regions within a sequence is of particular interest in the initial phase of *de novo* computational tertiary structure prediction of proteins[27]. Furthermore, we derived a specialized hydrophobicity scale from α -helical mammalian MPs only to be able to identify α -helical trans-membrane spans in the ORF of the human genome and the genome of other mammals.

To demonstrate the usefulness of these scales for such applications and to allow comparison with other hydrophobicity scales we implemented a simple version of such a prediction scheme for trans-membrane regions: The hydrophobicity values are averaged over a window of 15 amino acids. While we realize that this simple scheme is sub-optimal to achieve high-quality predictions in particular for β -barrel proteins, it proves efficient to benchmark these scales and compare it to other hydrophobicity scales.

Methods

Creation of the databases of non-redundant multi-span membrane proteins

Knowledge-based potentials are derived from a database of known properties and have shown to be especially suitable to describe features of proteins in structural biology (e.g., see [32] and [33]). For the derivation of such potentials, the ProteinDataBank (PDB) is an invaluable resource. It contains ~46,000 three-dimensional structures of soluble proteins and ~850 structures of MPs (as of 02/2008), about 70% of which are multi-span MPs. Tusnady et al. compiled the PDBTM [34], a sub-database of the PDB which contains all MPs and includes additional information such as the bilayer thickness for each protein determined by the TMDET algorithm [35-36]. In this database coordinates of symmetric domains were reconstructed from the crystallographic symmetry transformations (SYMTR) in the PDB entry and conversely coordinates of redundant atoms (from crystallization) are removed.

For the derivation of the UHS the complete list of multi-span MPs from the PDBTM was submitted to the PISCES server [37-38] to identify proteins with low sequence similarity. The input parameters used for culling are the following: sequence percentage identity $\leq 25\%$, resolution = 0.0Å - 3.0Å, R-factor = 0.3, sequence length 40 – 10,000 amino acids. The resulting database of unique structures contained 60 MPs. Before proceeding with the analysis, all non-standard amino acids were converted into the closest standard amino acid type. Further details about the composition of this database are given in the results section. For a complete list of all proteins see Supplementary Table (I).

For deriving the MHS all MPs in the PDBTM were classified according to their host organism. The list of mammalian proteins (156 PDB entries in total) was culled with the PISCES server using the following culling parameters: sequence identity $\leq 25\%$, resolution $\leq 0.0\text{Å} - 3.0\text{Å}$, R-factor 0.3, sequence length 40-10,000 amino acids. The

resulting database consisted of 16 α -helical proteins (from cattle, human, mouse, pig, rat, rabbit, and sheep) with 12,389 amino acids in total. The PDB codes of these proteins are: 1afo, 1okc, 1p49, 1ppj, 1u19, 1v54, 1vry, 1wpg, 1zll, 1zoy, 2b6o, 2hac, 2hfe, 2jwa, 2uui, 2z9a. Since these proteins have large extra-membrane domains only 2,563 amino acids were located in the membrane bilayer and the remaining 9,826 belonged to the soluble phase. For the three-state scenario, 2563 residues were located in the TM, 3122 in the TR, and 6704 in the SOL. These biases were corrected by appropriate normalization procedures (see below).

Definition of membrane, transition, and soluble regions

We distinguish between two different scenarios: (a) the two-state scenario, where only the trans-membrane (TM) and soluble region (SOL) is defined and no transition region exists and (b) the three-state scenario, where trans-membrane, transition (TR) and soluble region exist. A UHS was derived for both scenarios. While the two state scenario allows for comparison with most of the published hydrophobicity scales, the three state scenario gives a more comprehensive and detailed picture of free energies and can be compared to the Wimley & White hydrophobicity scale[6].

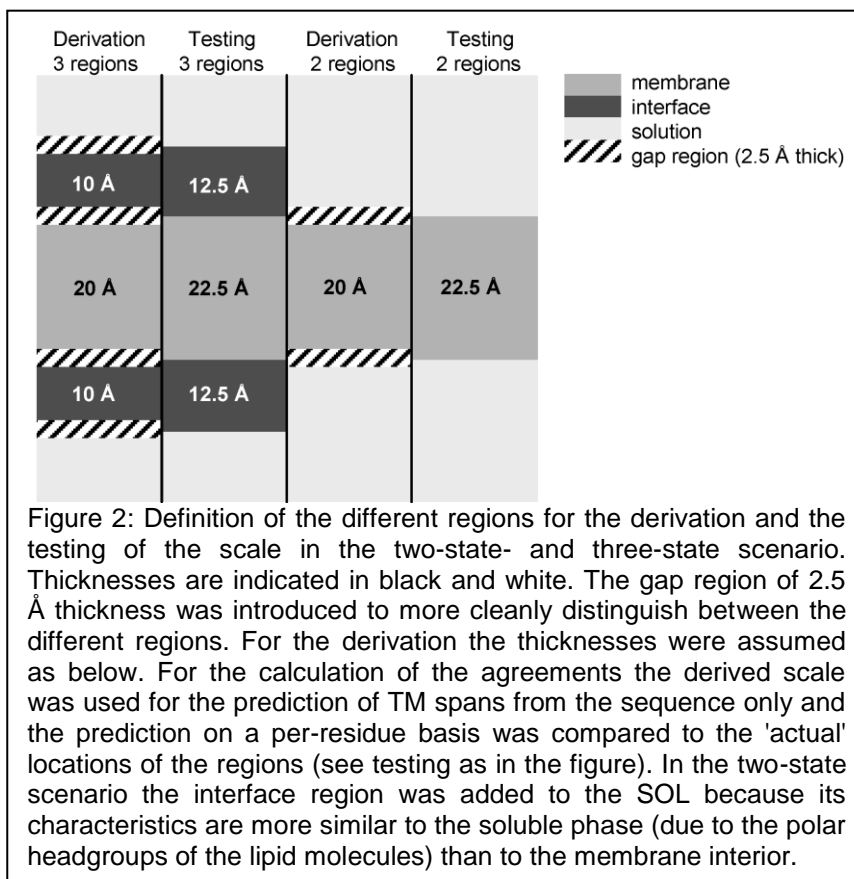
In the three-state scenario we assume a thickness of 20 Å for the TM core region [39]. On either side this region is flanked by a 2.5 Å buffer zone, before the TR regions begins. Its thickness is assumed with 10 Å on either side of the membrane and connects to another buffer zone of 2.5 Å. Adjacent to this second buffer zone the SOL regions starts (Figure 2).

In the two-state scenario the SOL and TR regions are combined and the buffer zone between them vanishes. This procedure was chosen since SOL and TR share a higher similarity when compared to their respective similarity to the TM region.

The buffer zones were added to distinguish more cleanly between the different regions and account for differences in the membrane thicknesses. We abstained from using the membrane layer thicknesses given in the PDBTM to avoid a somewhat recurrent influence of another prediction method on our results. We also found that usage of individual membrane thicknesses influenced the hydrophobicity values only marginally.

Derivation of amino acid propensities in the respective regions

To derive the free energies from the database the occurrence of each amino acid in each region was counted, which resulted in a total of 60 frequencies for the three-state scenario (20 amino acids x 3 regions) and 40 frequencies for the two-state scenario (20 amino acids x 2 regions). To eliminate a bias in the original data with respect to any region, the number of amino acids in each region was normalized to 20.



Afterwards the propensity as defined by Shortle[40] was computed:

$$P = \frac{\text{number}(\text{region}, AA) / \text{number}(\text{region})}{\text{number}(AA) / \text{number}(\text{total})}. \quad (1)$$

The expected propensity for a randomly selected cell in the resulting matrix is 1, which is important for the proper definition of the reference energy (see below).

Translation of propensities into free energies

The resulting propensities P were used to derive the free energies, ΔG , for each amino acid kcal/mol using the equation

$$\Delta G = -RT \ln P \quad (2)$$

with $R = k_B N_A$ (k_B being Boltzmann's constant and N_A being Avogadro's constant) at a temperature of $T = 293$ K. In the two-state scenario one can rewrite equation (2) to directly arrive at water to trans-membrane phase transfer free energies $\Delta\Delta G$ for each amino acid using the equation

$$\Delta\Delta G_{TM-SOL} = \Delta G_{TM} - \Delta G_{SOL} = -RT \ln\left(\frac{P_{TM}}{P_{SOL}}\right) \quad (3)$$

A corresponding equation applies for water to transition phase transfers.

Averaging of free energies over a sequence window of variable size for prediction

In order to obtain a prediction for a particular amino acid to be in one of the three regions (TM, TR, or SOL) the hydrophobicity values are averaged over a certain number of residues.

Two different approaches for averaging were tested: (a) all amino acids within the window have the same weight (rectangular weight function), and (b) the central residue has the highest weight with a linear decrease towards the edges of the window

where the weight is set to zero (triangular weight function). The resulting averaged free energy was utilized to predict the state of the central residue. Predictions over a complete sequence were achieved by sliding the window over the whole sequence (Supplementary Figure 2). Window sizes from 1 residue (no window) to 31 residues were tested. Only odd window sizes were considered to unambiguously assign a central residue.

Comparison of the hydrophobicity scales

In order to test the performance of the scale the average value of the free energies over a certain window size was calculated for SOL, TR, and TM free energies in the three-state scenario and TM, and SOL region in the two-state scenario. The amino acid in the center of the window was assigned the state that corresponds to the lowest of the average energies. Agreement for a specific region was computed as percentage of correctly predicted amino acids. The overall agreement was computed by averaging the agreements in all regions. For the assignment of the correct state the 2.5 Å buffer zones were split in half, i.e. the membrane was 22.5 Å and the TR was 12.5 Å thick with no buffer zones in between (Figure 2).

Construction of datasets for cross-validation

To perform cross-validation and to obtain standard deviations for the free energies and transfer free energies the database was divided into subsets. For the UHS the dataset was divided into five subsets, where four sets were taken for the derivation and the performance was tested on the fifth independent set. All experiments were repeated five times with the independent test-set permuting through the five datasets. The subsets were chosen to contain approximately the same number of α -helix, β -strand, and coil residues (Supplementary Table (I)). Since the proteins vary considerably

in size the numbers of proteins within the subsets fluctuate. A two-fold cross-validation was set up for the MHS as the dataset was significantly smaller with only 16 proteins.

Testing of the scale on four proteins

To test the algorithm four different example proteins from the PDBTM which were not present in the MP-database of 60 proteins, were investigated. The examples comprise the voltage-gated potassium channel KcsA (PDB code 1K4C), the chloride channel CIC (PDB code 1KPK), the Glycerol facilitator protein GlfP (PDB code 1LDI), and the outer membrane protein W OmpW (PDB code 2FIT). The examples were chosen so as to test both α -helical and β -barrel proteins. Furthermore, the α -helical proteins present difficult examples because of short or broken α -helices as in the selectivity filter of KcsA and in GlfP, and the unusually large tilt angles of the α -helices in CIC.

Results and Discussion:

Composition of the database of 60 non-redundant multi-span membrane proteins

The database of 60 non-redundant multi-span MPs encompasses a total of 43,523 amino acids. 31.4% of which reside in the TM region, 33.6% reside in the TR region, and 35.0% reside in the SOL region. Including the extra-membrane domains 21 proteins were purely α -helical, 5 were purely β -strand, and 34 were mixed α -helical/ β -strand proteins. Around 50% of all secondary structure elements reside in extra-membrane domains. In total 977 α -helices (605 of which were TM) and 1056 β -strands (405 of which were TM) were present in the database. For a summary of these data see Supplementary Table (I). When deriving the free energy scales, amino acid counts were normalized by region to avoid a bias in the hydrophobicity values that resulted from an imbalanced database.

Table (I): Values of the water-membrane transfer free energies in kcal/mol.

| | 2-state scales | | | | | | | | | | 3-state scales | | | | | | | |
|----------------|----------------|-------|-------|-----------|-------|-------|-----------------|-------|------------------|------------------|-------------------|-------|------|--------------------|-------------------|--------------------|-------------------|------|
| | experimental | | | consensus | | | knowledge-based | | | | WW _{int} | | | UHS _{int} | | | | |
| | HW* | GES* | WW | HWvH | EW* | KD* | Guy | Janin | PM _{1D} | PM _{3D} | UHS | SD | MHS | STD | WW _{int} | UHS _{int} | SD _{int} | |
| <i>polar</i> | C | -1.00 | -2.00 | -0.02 | -0.13 | -0.29 | -2.50 | -1.42 | -0.90 | -0.06 | -0.15 | 0.01 | 0.15 | 0.01 | 0.02 | -0.24 | 0.78 | 0.07 |
| | N | 0.20 | 4.80 | 0.85 | 2.05 | 0.78 | 3.50 | 0.48 | 0.50 | 0.18 | 0.22 | 0.50 | 0.03 | 0.03 | 0.42 | -0.04 | 0.01 | |
| | Q | 0.20 | 4.10 | 0.77 | 2.36 | 0.85 | 3.50 | 0.95 | 0.70 | 0.26 | 0.03 | 0.46 | 0.07 | 0.07 | 0.58 | 0.18 | 0.04 | |
| | S | 0.30 | -0.60 | 0.46 | 0.84 | 0.18 | 0.80 | 0.52 | 0.10 | 0.05 | 0.16 | 0.06 | 0.04 | 0.39 | 0.13 | 0.06 | 0.04 | |
| | T | -0.40 | -1.20 | 0.25 | 0.52 | 0.05 | 0.70 | 0.07 | 0.20 | 0.02 | -0.08 | -0.01 | 0.01 | -0.05 | 0.01 | 0.14 | 0.02 | 0.01 |
| <i>charged</i> | D | 3.00 | 9.20 | 3.64 | 3.49 | 0.90 | 3.50 | 0.78 | 0.60 | 0.37 | 0.41 | 0.73 | 0.05 | 1.10 | 1.23 | 0.13 | 0.04 | |
| | E | 3.00 | 8.20 | 3.63 | 2.68 | 0.74 | 3.50 | 0.83 | 0.70 | 0.15 | 0.30 | 0.70 | 0.03 | 1.09 | 2.02 | 0.41 | 0.02 | |
| | K | 3.00 | 8.80 | 2.80 | 2.71 | 1.50 | 3.90 | 1.40 | 1.80 | 0.32 | 0.24 | 0.90 | 0.04 | 1.24 | 0.99 | 0.09 | 0.04 | |
| | R | 3.00 | 12.30 | 1.81 | 2.58 | 2.53 | 4.50 | 1.91 | 1.40 | 0.37 | 0.32 | 0.55 | 0.05 | 1.24 | 0.81 | 0.04 | 0.02 | |
| <i>apolar</i> | A | -0.50 | -1.60 | 0.50 | 0.11 | -0.62 | -1.80 | 0.10 | -0.30 | -0.17 | -0.15 | -0.16 | 0.03 | -0.24 | 0.17 | 0.02 | 0.02 | |
| | G | 0.00 | -1.00 | 1.15 | 0.74 | -0.48 | 0.40 | 0.33 | -0.30 | 0.01 | 0.08 | -0.20 | 0.03 | -0.10 | 0.06 | 0.01 | -0.25 | 0.02 |
| | I | -1.80 | -3.10 | -1.12 | -0.60 | -1.38 | -4.50 | -1.13 | -0.70 | -0.28 | -0.29 | -0.39 | 0.03 | -0.43 | 0.01 | -0.31 | -0.06 | 0.03 |
| | L | -1.80 | -2.80 | -1.25 | -0.55 | -1.06 | -3.80 | -1.18 | -0.50 | -0.28 | -0.36 | -0.30 | 0.03 | -0.48 | 0.07 | -0.56 | -0.08 | 0.03 |
| | M | -1.30 | -3.40 | -0.67 | -0.10 | -0.64 | -1.90 | -1.59 | -0.40 | -0.26 | -0.19 | -0.20 | 0.02 | -0.17 | 0.09 | -0.23 | -0.12 | 0.06 |
| | P | 0.00 | 0.20 | 0.14 | 2.23 | -0.12 | 1.60 | 0.73 | 0.30 | 0.13 | 0.15 | 0.50 | 0.04 | 0.50 | 0.11 | 0.45 | 0.07 | 0.03 |
| | V | -1.50 | -2.60 | -0.46 | -0.31 | -1.08 | -4.20 | -1.27 | -0.60 | -0.17 | -0.24 | -0.25 | 0.02 | -0.35 | 0.04 | 0.07 | 0.09 | 0.02 |
| | F | -2.50 | -3.70 | -1.71 | -0.32 | -1.19 | -2.80 | -2.12 | -0.50 | -0.41 | -0.22 | -0.46 | 0.04 | -0.43 | 0.10 | -1.13 | -0.36 | 0.02 |
| | H | -0.50 | 3.00 | 2.33 | 2.06 | 0.40 | 3.20 | -0.50 | 0.10 | -0.02 | 0.06 | 0.38 | 0.05 | 0.64 | 0.47 | 0.96 | -0.07 | 0.05 |
| | W | -3.40 | -1.90 | -2.09 | 0.30 | -0.81 | 0.90 | -0.51 | -0.30 | -0.15 | -0.28 | -0.03 | 0.03 | -0.06 | 0.01 | -1.85 | -0.38 | 0.04 |
| Y | -2.30 | 0.70 | -0.71 | 0.68 | -0.26 | 1.30 | -0.21 | 0.40 | -0.09 | -0.03 | -0.12 | 0.04 | 0.23 | 0.17 | -0.94 | 0.01 | 0.02 | |

HW: Hopp & Woods, GES: Goldman, Engelman, Steitz, WW: Wimley & White, HWvH: Hessa, White & von Heijne, EW: Eisenberg & Weiss, KD: Kyte & Doolittle, Guy: Guy, Janin: Janin, PM1D and PM3D: Punta & Maritan, UHS: the knowledge-based scale derived in this paper with its standard deviation (SD), MHS: the mammalian scale derived here with its standard deviation. The last three columns show the values for the transition between water-interface from the Wimley & White scale and the values from the UHS with its standard deviations (SD). Shaded regions are negative.

* The values from the literature have been inverted to match the direction of transfer from water to bilayer.

The two-state scale allows direct comparison with other hydrophobicity scales

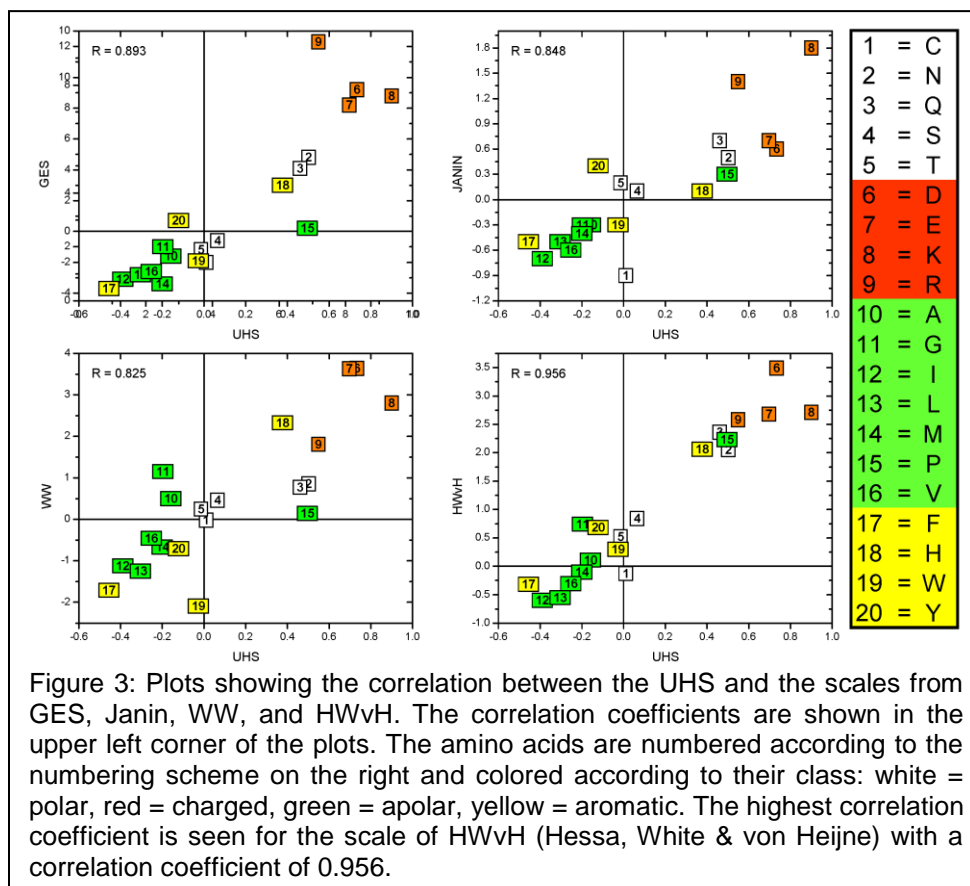
Most of the hydrophobicity scales in the literature have been derived for two regions, i.e. no TR is defined (see equation (2)). Although we strongly encourage ultimate usage of a three-state scale, a two-state UHS was derived in order to facilitate comparison with other methods. All hydrophobicity values are summarized in Table (I) and the characteristics of the different scales are given in Table (II). Correlation with other hydrophobicity scales is plotted in Figure 3 and Supplementary Figure (1).

Table (II): Chart summarizing the different hydrophobicity scales and their applicability.

| scale | ref | year | derivation* | α/β | 2-state/ 3-state | characteristics/applicability |
|---------------------------|---------|---------------|-------------|----------------|---------------------|---|
| Hopp & Woods | [2] | 1981 | exp | n/a | 2 | <ul style="list-style-type: none"> • hydrophilicity scale for antigenic sites on the protein surface; • derived from the values of Levitt[41]; • some values were adjusted to fit immunochemical data of 12 proteins; • for the proteins only the primary sequence was available!; • window used is 6 residues \approx length of antigenic determinant |
| Goldman, Engelman, Steitz | [3] | 1986 | exp | α | 2 | <ul style="list-style-type: none"> • hydrophobicity scale for single trans-membrane helices; • semi-theoretical approach based on energetic considerations of residues undergoing hydrogen bonds in helices derived from experimental data in the literature; • hydrophobicity scale as a sum of hydrophilic and hydrophobic components |
| Wimley & White | [4-5] | 1996 | exp | α | 2 + 3 | <ul style="list-style-type: none"> • derived by measuring the partitioning energies of host-guest penta-peptides; • whole residue scale that considers the polar peptide bond; • interface: POPC vesicle interface; bilayer: n-octanol; • for unfolded peptides in all 3 phases (solution, interface, bilayer) |
| Hessa et al. | [23-24] | 2005/ 2007 | exp | α | 2 / pot | <ul style="list-style-type: none"> • designed TM helix within the Lep protein that is inserted via the Sec61 translocon; • TM helix is 19-residue helix with amino acid in question incorporated in the center; • measured fraction of singly vs. doubly glycosylated Lep molecules to derive the scale; • therefore applicable to folded MPs; • scale has been extended |

| | | | | | | |
|---------------------|------|------|------|----------------|---------|---|
| | | | | | | to position-dependent free energy scale (2007) |
| Eisenberg & Weiss | [42] | 1982 | cons | n/a | 2 | • normalized consensus scale of five different scales |
| Kyte & Doolittle | [9] | 1982 | cons | n/a | 2 | • normalized consensus scale based on experimental observations of different scales; • refinement by studying hydrophathy plots of proteins of known X-ray structure; |
| Guy | [11] | 1985 | cons | n/a | 2 | • based on experimental and statistical results from several studies; • considers solvent accessibility according to accessible layers of amino acids in globular proteins |
| Janin | [8] | 1979 | KB | n/a | 2 | • derived from X-ray structures of 22 soluble proteins; • looked at molar fraction of buried and accessible residues |
| Punta & Maritan | [10] | 2003 | KB | α | 2 | • derived two membrane propensity scales from two TM helix databases using a simple perceptron algorithm; • databases contained 118/228 TM helices; • sequence identity of the proteins was 30% |
| Beuming & Weinstein | [30] | 2004 | KB | α | n/a | • calculated surface propensities of amino acids (probability of finding a residue on the surface of a TM protein); • based on surface fractions of residues; • considered 28 α -helical MPs |
| Senes et al. | [12] | 2007 | KB | α | 2 / pot | • calculated membrane depth-dependent potential for amino acid side chains; • considered 24 α -helical MPs |
| UHS | | 2008 | KB | α/β | 2 + 3 | • derived from 60 known structures of folded MPs; • considers folded structures both in solution and membrane bilayer; • both α , β , and α/β structures were taken into account with approximately equal distribution of helices and strands; • considers only depth in membrane bilayer and no accessibility or secondary structure |
| MHS | | 2008 | KB | α | 2 + 3 | • derived from 16 known structures of folded MPs from mammalian organisms; • only α -helical structures could be taken into account; • considers folded structures both in solution and membrane bilayer; • considers only depth in membrane bilayer and no accessibility or secondary structure |

*exp: experimental; cons: consensus; KB: knowledge-based; pot: potential



Three-state scale demonstrates the preference of Trp for interface region

Table (III) shows the free energy values in kcal/mol for all 20 amino acids and for all three regions (TM, TR, SOL). As for the two-state scenario, Cys has a large standard deviation for TM (0.09) and SOL (0.06) and its large value within TR indicates that it does not prefer to be in the TR. Ser and Thr have almost no preference for any of the three regions (Ser: TM = 0.02, TR = 0.02, SOL = -0.04; Thr: TM = -0.01, TR = 0.02, SOL = 0.00), which agrees with the findings of Senes [12] and Hessa [23]. The fact that Trp is often found in the TR [12, 43-45] is confirmed by our results. It has been previously noted that Tyr also has a preference for the interface between TM and TR region. However, this preference of Tyr for the interface region is less distinct when compared to Trp [12, 23]. In the UHS Tyr shows a slight preference for residing within the TM region

which could be a result of a slightly larger membrane thickness in our definition when compared to other scales [23]. Further, we find strong preferences for Ile, Phe, Leu, Val, and Met to be in the TM region and for Glu, Lys, Cys, Asp, and Gln to be in the SOL region.

Table (III): Free energy values of the UHS and MHS in kcal/mol

The table shows the knowledge-based values for the free energies in kcal/mol and their corresponding standard deviations (SD) for the 20 amino acids in the three regions of the membrane bilayer (TM), the transition region (TR) and the soluble region (SOL) for both scales, the UHS and the MHS. The shaded cells indicate the preference of the amino acid for that region. Note that Serine and Threonine in the UHS show almost no preference for any of the three regions.

| | | UHS | | | | | | MHS | | | | | |
|----------|---|-------------|------|-------------|------|--------------|------|-------------|------|-------------|------|--------------|------|
| | | TM \pm SD | | TR \pm SD | | SOL \pm SD | | TM \pm SD | | TR \pm SD | | SOL \pm SD | |
| polar | C | -0.07 | 0.09 | 0.56 | 0.03 | -0.22 | 0.06 | -0.01 | 0.01 | 0.14 | 0.14 | -0.09 | 0.10 |
| | N | 0.37 | 0.03 | -0.14 | 0.01 | -0.10 | 0.01 | 0.41 | 0.04 | -0.15 | 0.03 | -0.12 | 0.04 |
| | Q | 0.31 | 0.05 | -0.01 | 0.03 | -0.19 | 0.03 | 0.60 | 0.09 | -0.13 | 0.07 | -0.19 | 0.04 |
| | S | 0.02 | 0.03 | 0.02 | 0.03 | -0.04 | 0.02 | 0.13 | 0.28 | -0.01 | 0.07 | -0.06 | 0.14 |
| | T | -0.01 | 0.01 | 0.02 | 0.01 | 0.00 | 0.01 | -0.02 | 0.01 | -0.04 | 0.04 | 0.06 | 0.04 |
| charged | D | 0.52 | 0.04 | -0.08 | 0.03 | -0.21 | 0.02 | 0.82 | 0.28 | 0.05 | 0.18 | -0.34 | 0.05 |
| | E | 0.47 | 0.02 | 0.10 | 0.02 | -0.31 | 0.01 | 0.84 | 0.05 | 0.06 | 0.10 | -0.36 | 0.04 |
| | K | 0.69 | 0.03 | -0.13 | 0.03 | -0.22 | 0.03 | 0.99 | 0.15 | -0.08 | 0.03 | -0.30 | 0.04 |
| | R | 0.40 | 0.04 | -0.11 | 0.02 | -0.15 | 0.01 | 1.06 | 0.13 | -0.22 | 0.08 | -0.18 | 0.10 |
| apolar | A | -0.10 | 0.02 | 0.07 | 0.02 | 0.05 | 0.01 | -0.15 | 0.01 | 0.12 | 0.02 | 0.07 | 0.01 |
| | G | -0.09 | 0.02 | -0.06 | 0.01 | 0.19 | 0.02 | -0.07 | 0.03 | 0.05 | 0.03 | 0.03 | 0.06 |
| | I | -0.22 | 0.02 | 0.12 | 0.02 | 0.18 | 0.02 | -0.24 | 0.01 | 0.13 | 0.08 | 0.23 | 0.07 |
| | L | -0.16 | 0.01 | 0.06 | 0.02 | 0.14 | 0.02 | -0.26 | 0.03 | 0.15 | 0.07 | 0.24 | 0.00 |
| | M | -0.12 | 0.03 | 0.01 | 0.05 | 0.13 | 0.04 | -0.07 | 0.04 | -0.06 | 0.02 | 0.16 | 0.02 |
| | P | 0.34 | 0.04 | -0.08 | 0.02 | -0.15 | 0.02 | 0.36 | 0.07 | -0.09 | 0.07 | -0.14 | 0.09 |
| | V | -0.16 | 0.01 | 0.15 | 0.02 | 0.06 | 0.01 | -0.22 | 0.03 | 0.22 | 0.05 | 0.09 | 0.01 |
| aromatic | F | -0.19 | 0.02 | -0.02 | 0.02 | 0.34 | 0.01 | -0.22 | 0.05 | 0.00 | 0.01 | 0.35 | 0.11 |
| | H | 0.29 | 0.03 | -0.14 | 0.03 | -0.07 | 0.04 | 0.55 | 0.41 | -0.28 | 0.03 | 0.03 | 0.12 |
| | W | 0.08 | 0.02 | -0.19 | 0.02 | 0.19 | 0.03 | 0.07 | 0.11 | -0.24 | 0.15 | 0.36 | 0.24 |
| | Y | -0.07 | 0.02 | 0.04 | 0.01 | 0.03 | 0.02 | 0.20 | 0.09 | -0.13 | 0.04 | -0.02 | 0.12 |

The absence of structural context leads to less distinct free energy values

relevant for multi-span membrane proteins

It can be seen from Table (I) and Figure 1 that the values of knowledge-based hydrophobicity scales are in general not as pronounced as in scales that were derived experimentally. This observation holds for the newly derived hydrophobicities: for

example, while the GES scale ranges from -3.70 (Phe) to 12.30 (Arg), the Wimley & White scale from -2.09 (Trp) to 3.64 (Asp), the Hessa, White & von Heijne ranges from -0.60 (Ile) to 3.49 (Asp), the values in the UHS derived here range only from -0.46 (Phe) to 0.90 (Lys). However, the correlation diagrams indicate that despite the deviation in absolute values the scales agree very well in general trends with correlation coefficients between $R=0.804$ and $R = 0.956$. The UHS has higher correlation coefficients to knowledge-based scales such as PM1D and PM3D, and surprisingly, the highest correlation coefficient ($R = 0.956$) is found for the Hessa, White & von Heijne scale, the most recent experimental scale considered (see below).

By disregarding structural context such as the level of exposure when deriving the scale the absolute size of the free energies derived is reduced. This originates in the MP database used for derivation containing only multi-span MPs. These proteins have both hydrophobic cores and active polar sites within the TM shielded from direct contact with the membrane lipids, as f.ex. in ion channel proteins. Similarly, the extra-membrane domains of these proteins also have both hydrophobic cores and polar active sites shielded from direct interaction with the solvent. This reduces the absolute size of the free energies obtained. This has to be compared to e.g. an experimental scale that was observed for model peptides forming single α -helices within the membrane exposing their amino acid side chains almost completely to the lipid and having no extra-membrane domains with hydrophobic interior.

When compared to all the other tested experimental scales, the 'biological' transfer free energies from Hessa et al.[24] match the knowledge-based ones very closely in size and distribution. The scale yields the highest correlation coefficient of $R = 0.956$ to the UHS (compare Figure 3 and Supplementary Figure (1)). The reason for the smaller range is the measurements on an intact protein (*E. coli* leader peptidase)

consisting of three TM segments where the structural context for the residue in question is maintained.

Triangular window function of 15 residues used for
predicting trans-membrane spans

In order to test the usefulness of the derived UHS it was applied towards predicting the state of an amino acid (TM, TR, or SOL) from primary sequence only. To achieve increased prediction accuracies the hydrophobicities were averaged over a sequence window. This procedure allows for identification of spans of similar dielectric environments along the protein sequence.

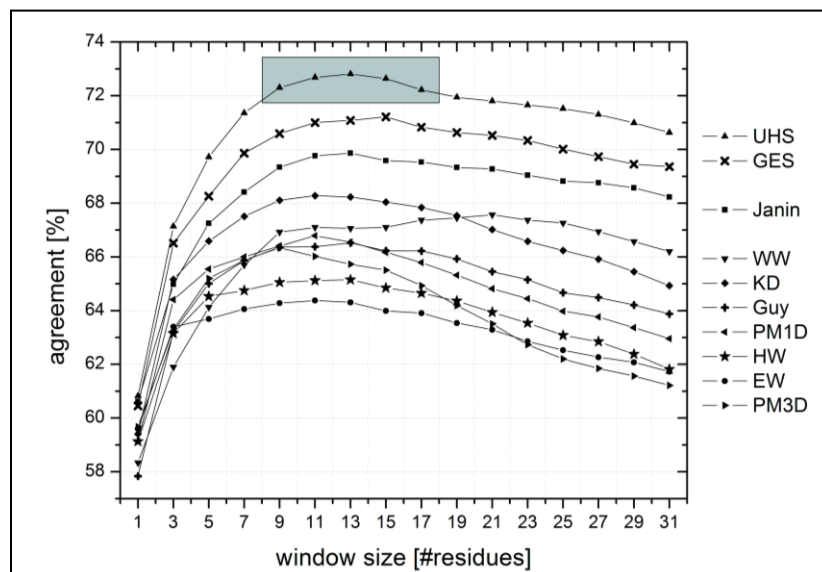


Figure 4: For predicting trans-membrane spans from the sequence, the hydrophobicity values have to be averaged over a certain number of residues ("window"). The percent per-amino acid agreements between prediction and known location of the residues were computed as a function of window size for the scales from the literature (EW: Eisenberg & Weiss [1], GES: Goldman, Engelman, Steitz [3], HW: Hopp & Woods [2], KD: Kyte & Doolittle [9], WW: Wimley & White [6], Guy: Guy [11], Janin: Janin [8], PM1D and PM3D: Punta & Maritan [10]) and for the Unified Hydrophobicity Scale (UHS). The shaded region indicates a range of window lengths for the UHS, which all yield similarly good performance. The best performance is seen for the UHS scale and the scale from GES.

Preliminary prediction trials have shown that the triangular window function performed better than the rectangular window. Since hydrophobicity is a local measure and therefore depends more on neighboring residues than on residues further away, this result is expected. In addition, these preliminary trials showed that the prediction accuracy in the two-state scenario (distinguishing TM from SOL region) is dependent on the window size as can be seen in Figure 4. Note, that there is a plateau range between 9 and 17 residues where all scales gave consistently good results. Therefore, for all further experiments we chose a window size of 15 residues. This number agrees with the average length of an α -helix spanning the core region of the membrane (15 residues \times 1.5Å rise = 22.5Å membrane thickness).

Two-state scenario: UHS achieves 72.6% correct classifications

Table (IV) displays the percentages of agreement for the TM and SOL with their average value. The scales of EW, HW, PM3D, PM1D, and Guy display a bias towards predicting an amino acid within the TM (>80% agreement) but poorly agree in the SOL (<50% agreement). Conversely, the scales of WW and HWvH bias towards the SOL (87% and 99%) with a lower performance in the TM (48% and 11%). These biases are indicative of offsets in the absolute TFE values when applied to intact multi-span MPs and may not exist in other applications. This is not unexpected given that the reference point for every experimental scale is imposed by the experimental setup. For example, the bias in the WW scale originates from the fact that the scale was derived for unfolded peptides in both solution and membrane bilayer.

Table (IV): Per amino acid agreements for the two-state scenario

The table shows the percentage per amino acid agreements for the two-state scenario between the prediction and the PDB for the hydrophobicity scales from the literature and the UHS. (TM) membrane bilayer; (SOL) soluble phase (SOL); (avg) average value of agreement between TM and SOL. The values are computed for a window size of 15 residues for averaging. The first four scales on the left show similar performances for the TM and the SOL, whereas the other scales exhibit an uneven distribution.

| | | PDB | | | PDB | | |
|------|-----|---------|--------|--------|------|-----|-----|
| | | TM | SOL | avg | TM | SOL | avg |
| pred | TM | UHS | | | Guy | | |
| | SOL | 70 ± 10 | 25 ± 7 | 73 ± 2 | 81 | 49 | 66 |
| | | 30 ± 10 | 75 ± 7 | | 19 | 51 | |
| pred | TM | GES | | | PM1D | | |
| | SOL | 66 | 23 | 71 | 86 | 53 | 66 |
| | | 34 | 77 | | 14 | 47 | |
| pred | TM | Janin | | | PM3D | | |
| | SOL | 72 | 32 | 70 | 83 | 52 | 66 |
| | | 28 | 67 | | 17 | 48 | |
| pred | TM | KD | | | HW | | |
| | SOL | 76 | 39 | 68 | 89 | 59 | 65 |
| | | 24 | 61 | | 11 | 41 | |
| pred | TM | WW | | | EW | | |
| | SOL | 48 | 13 | 67 | 88 | 60 | 64 |
| | | 52 | 87 | | 12 | 40 | |
| pred | TM | HWvH | | | | | |
| | SOL | 11 | 1 | 55 | | | |
| | | 89 | 99 | | | | |

The other scales predict amino acids in an approximately balanced distribution ($KD_{TM} = 76\%$, $KD_{SOL} = 61\%$; $Janin_{TM} = 72\%$, $Janin_{SOL} = 67\%$; $GES_{TM} = 66\%$, $GES_{SOL} = 77\%$; $KB_{TM} = 70\%$, $KB_{SOL} = 75\%$). While the good performance of our UHS scale is remarkable considering the simple approach it was derived with, it should be acknowledged that particularly good performance is expected in this experiment since the scale was derived with particular focus on such applications.

Even though the improvement of UHS above the GES scale is small in the two-state scenario, this translates into a significant improvement when the accuracy of detecting full-length TM spans from the sequence is analyzed. Here the UHS identifies 81.1% of the TM spans, the GES scale identifies 76.6%, and the WW scale identifies 59.9%.

False positive rate on soluble proteins is comparable to GES scale

To assess the over-prediction of regions in soluble proteins as being in the TM region the scale was tested on a non-redundant set of soluble proteins (<25% sequence identity). This set was created by culling the PDB with the PISCES server with the same culling parameters as for the MHS and UHS (see Methods section). The database comprised 2,569 proteins with 3,538 chains and 526,422 amino acids.

Detailed results can be found in Supplementary Table (II). The scales of Hessa et al. and of Wimley & White predict amino acids as being in the SOL more than 95% of the time and hence have a corresponding false positive rate for prediction TM spans of smaller than 5%. This originates in the tendency of these scales to over-predict amino acids as being in the SOL. In result both scales have a significantly reduced accuracy in the TM (compare Table (IV)). The scales of GES, Janin, KD, and the UHS have (according to Table (IV)) an approximately balanced distribution between SOL and TM and have a high agreement in SOL with a small number of false positives. Among these four scales, the UHS performs comparably well to the GES with an accuracy of ~86% in solution and ~14% over-prediction. Both scales are significantly better than the scales of Janin or KD in this experiment. The remaining scales (Punta & Maritan, Guy, Hopp & Woods, Eisenberg & Weiss) have a lower agreement in the SOL coupled with an increased rate of false positives caused by the tendency of these scales towards over-predicting amino acids as being in the TM.

The over-prediction of amino acids in soluble proteins as being in the TM region is reduced by about 10% when compared to the SOL of MPs (Supplementary Table (II, IV)). In MPs many residues close to the membrane surface are counted as soluble in the two-state scenario. These residues are difficult to be accurately predicted as they often interact with the membrane surface and not only with the solvent. Further, the window for averaging will include some membrane amino acids for these residues. The absence of such difficult residues improves the prediction accuracy when looking at soluble proteins.

Comparison of UHS and GES for individual amino acids

Supplementary Figures (3) and (4) summarize the results for individual amino acids for the two-state scenario in comparison to the Goldman, Engelman, Steitz (GES) scale, which gave, according to Figure 4, for this experiment the best results besides the UHS. Both scales over-predict the polar amino acids Arg, Asn, Asp, Glu, Gln, and Lys in the SOL region. For the UHS the average agreements are higher for the polar residues Arg, Asp, Glu, and Lys.

Comparing the GES scale with the UHS, the average agreements have increased most for Arg (51% to 58%), Cys (72% to 78%), and Glu (58% to 62%). Note that the average agreement in the UHS is lower than in the GES scale only for His (72% to 69%). This indicates a slightly better representation of polar residues in the present UHS.

Three-state scenario: UHS displays agreement of 57.1%

As discussed earlier, one strength of the UHS scale is that in contrast to many existing methods it distinguishes three regions. Only one of the nine scales used for comparison was derived with a TR region. Hence, comparison for the three-state

scenario is limited to the Wimley & White (WW) scale. The data are summarized in Table (V). For classifying an amino acid correctly in one of the three regions TM, TR, SOL the UHS scale achieves 57.1% as compared to 49.8% obtained for the WW scale. For the UHS the agreements for the different regions are relatively balanced (TM = 63.2%, TR = 43.8%, SOL = 64.4%). As already observed for the two-state scenario, the WW scale is biased in its prediction towards the SOL with an agreement of 89.1%. However, the agreement drops to 24.4% for the TR and 35.9% in the TM. Again, we wish to emphasize that these biases result from a different experimental setup and occur when applied to intact multi-span MPs, and may not exist in other applications.

Supplementary Figures (3) and (4) illustrate the individual amino acid agreements for the three-state scenario in comparison to the Wimley & White scale. As in the two-state scenario, the polar residues Arg, Asn, Asp, Glu, Gln, His, Lys, and Ser are predicted in a more balanced manner in the UHS than in the WW scale. When comparing the overall prediction accuracies, all amino acids either display an improvement or at least a similar accuracy for the UHS. Highest changes are observed for Asp and Glu (from 36% to 47%), Asn (from 41% to 50%), and His (from 44% to 53%).

Table (V): Per amino acid agreements for the three-state scenario

The table shows the percentage per amino acid agreements between the prediction and the PDB for the different regions for the UHS (with its standard deviation) in comparison to the Wimley & White scale. The performance of the MHS is also shown. The window size for averaging is 15 residues and (TM) represents the trans-membrane, (TR) the transition and (SOL) the soluble region. The percentages were calculated by dividing the correctly predicted number of amino acids by the total number of amino acids in that region. An average agreement (avg) was calculated by averaging the percentages of agreement for the diagonal elements of the matrix. While the average prediction agreement seems to be relatively low, note that there are three regions defined, so that the threshold between a good and a bad percentage of agreement would be 33% and not 50% as in the two-state system. For the Wimley & White scale both the octanol and the interface scale were used to establish a scale for three regions. The standard deviations for the UHS and MHS arise from cross-validation, whereas the scale of WW was tested on the whole dataset without cross-validation.

| | | TM | TR | SOL | avg |
|------|-----|---------|---------|--------|--------|
| pred | | UHS | | | |
| | TM | 63 ± 11 | 29 ± 10 | 9 ± 6 | |
| | TR | 23 ± 7 | 44 ± 3 | 26 ± 4 | |
| | SOL | 13 ± 6 | 27 ± 9 | 64 ± 8 | 57 ± 3 |
| pred | | WW | | | |
| | TM | 36 | 14 | 2 | |
| | TR | 29 | 24 | 9 | |
| | SOL | 35 | 62 | 89 | 50 |
| pred | | MHS | | | |
| | TM | 71 ± 1 | 17 ± 4 | 5 ± 2 | |
| | TR | 19 ± 3 | 48 ± 1 | 30 ± 4 | |
| | SOL | 10 ± 2 | 35 ± 2 | 65 ± 2 | 61 ± 0 |

It should be noted that the Wimley & White scale was derived for unfolded peptides in all three phases (solution, interface, and membrane bilayer). In contrast to folded secondary structure elements or domains where most backbone amide and carbonyl groups are undergoing hydrogen bonds, unfolded peptides can only engage in hydrogen bonds with polar solvents such as water, not with hydrophobic solvents or the membrane core. This fact offsets the WW scale towards a preference of the SOL region which explains the over-prediction for that region. Obviously, the Wimley & White scale

was not derived for the current application of predicting TM spans from the sequence only, [4-6, 46] and is an exceptional scale in its own right. We focus on its performance since it is the only available scale for three-state scenario we can use for comparison. The lack of suitable scales for the present application presents another justification for the development of the UHS.

The UHS enables prediction of TM spans from sequence only

We realize that different and more specialized hydrophobicity scales can be derived if assumptions on secondary structure (like separation of α -helices from β -strands) or tertiary structure (like level of exposure) were made. On purpose, such assumptions were forgone to make the hydrophobicity scale applicable in the absence of any structural information about the sequence of interest.

We also abstained from use of secondary structure prediction techniques since their accuracy is limited and most of these tools are highly specialized. However, we appreciate that the incorporation of secondary structure (e.g. the separate prediction for α -helices and β -strands) and/or the exposure of an amino acid is likely to be superior to the presented scale for certain applications.

The UHS is largely independent of the protein fold

To date, only a small fraction of the proteins stored in the PDB are MPs and only about 60 MP folds are known. When deriving a knowledge-based scale from such a limited database, the question arises whether this scale is applicable to the MP universe whose folds have not been elucidated yet. The scale could for example have a compositional bias of certain amino acid types due to the under-representation of distinct folds in the database.

While we believe that such a bias is unavoidable given the very limited number of MP structures known, we argue that it is small as the hydrophobicity scale is governed by more general rules of MP fold formations such as α -helix/ α -helix packing or β -barrel formation. We tested this hypothesis by excluding folds one by one when deriving the UHS scale and analyzing the effects on the hydrophobicity values. Note that for some MP folds multiple representatives are found in the database of 60 proteins, as it was culled purely by sequence and not by fold identity. Further we tested the prediction accuracies of these "leave-one-fold-out" UHS scales on the excluded folds. The details of this experiment are included in the supplementary data section. Briefly, we find the hydrophobicity values robust with respect to exclusion of a single fold (changes in hydrophobicity values are on average well below one standard deviation) and the prediction accuracy for TM and SOL regions is within 2.4% to the one observed with the UHS scale.

Limitations of the UHS compared to other hydrophobicity scales

The UHS described in this chapter is derived only in a membrane depth-dependent manner and does not take into account the SASA of the residues in the membrane. Also, the scale is derived from both α -helical MPs as well as β -barrels which both can have aqueous interior. This scale therefore under-estimates the penalty for charged residues in the membrane bilayer if they are lipid-exposed. Since both single and multispan α -helical MPs and β -barrels were used for derivation, our hydrophobicity scale combines characteristics of lipid-exposed surfaces, water-exposed surfaces, and protein interior. Even though the UHS might not accurately describe any of these very specific scenarios, it captures the characteristics of MP hydrophobicity in a statistical manner and is useful for *de novo* protein structure prediction where no structural information is known.

The energy potential derived by Senes & DeGrado [12] was created from 24 α -helical MPs. Creating the potential the amino acid propensities in 2 Å bins along the membrane bilayer were converted into free energies and fit using a sigmoidal or Gaussian distribution. The database only contained multi-span MPs excluding single-span MPs. The derived potential combines the characteristics of side chains facing the lipid bilayer and MP interior, even though the possibility is excluded that the MP interior is an aqueous pore. Furthermore, it is unclear of how the database was derived as neither a sequence-similarity or structure-similarity criteria were used to identify suitable representatives. This leads repeatedly to the question whether the used MP database is an accurate representation of the fold space. Moreover, deriving a potential from frequencies binned at 2Å using such a small database results in a noisy potential.

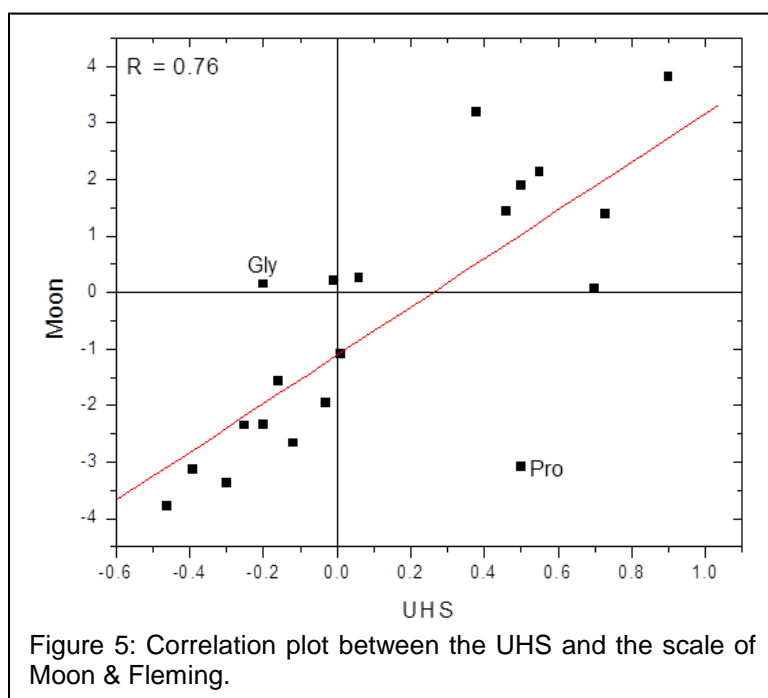
The surface propensity scale from Beuming & Weinstein [30] is derived from a database of 28 α -helical MPs. Compared to all other scales described here, this scale does not measure the transfer from water to bilayer but investigates the differences when the side chains face the lipid bilayer versus the interior of the protein. These differences are much more subtle than described by the other hydrophobicity scales that describe water/bilayer transitions. It would be interesting to know how well this database represents the MP fold space and whether the distribution of residues towards the lipid bilayer is different in β -barrel MPs. Even the question whether there is a difference in amino acid distribution between aqueous pore-facing residues and surface residues of soluble proteins would be interesting to investigate.

Hessa et. al derived an apparent free energy potential derived from singly versus doubly glycosylated Lep proteins as described above [23]. The authors state that about 1/4 of the TM helices in multi-span MPs have an apparent free energy value larger than zero indicating that it is less favorable for these helices to exist in the membrane. This can be attributed to how these free energies are derived, namely using a three TM span

helical protein, if the H-segment inserts into the membrane. In this case, 2/3 of the side chains of this helix are facing lipids while 1/3 are facing the side chains of the other two helices, which are very likely hydrophobic. The scale does not take into account proteins that have a large aqueous interior and therefore should favor the bilayer more than our UHS scale. Our scale and Hessa's scale show an overall high correlation with $R = 0.93$. Hessa's scale describes the transition between translocon and bilayer and not water/bilayer and the free energies might be affected by the interactions with the other TM helices in the protein. Therefore this scale does not separate side chain/lipid interactions and side chain/side chain interactions. It is also uncertain how accurate the fraction of singly vs. doubly glycosylated protein can be quantified from a SDS-Page electrophoresis.

The side chain hydrophobicity scale derived by Moon & Fleming [25] is the first scale that describes the free energy transfer of residue side chains from water into the membrane bilayer if the residue is embedded in a fully folded and "fully functional" β -barrel MP. The authors claim that OmpLA is fully functional even though the unfolding curves were recorded at pH 3.8 whereas the activity measurements were carried out at pH 8.0. They reason that the folded state of OmpLA is identical at both pH'es as judged by Tryptophan fluorescence and SDS-page electrophoresis measurements, however, high-resolution information is lacking to support this hypothesis. Another limitation, even though less likely, is the possibility that the strand containing the mutation flips orientations in the membrane such that the side chain faces the aqueous phase.

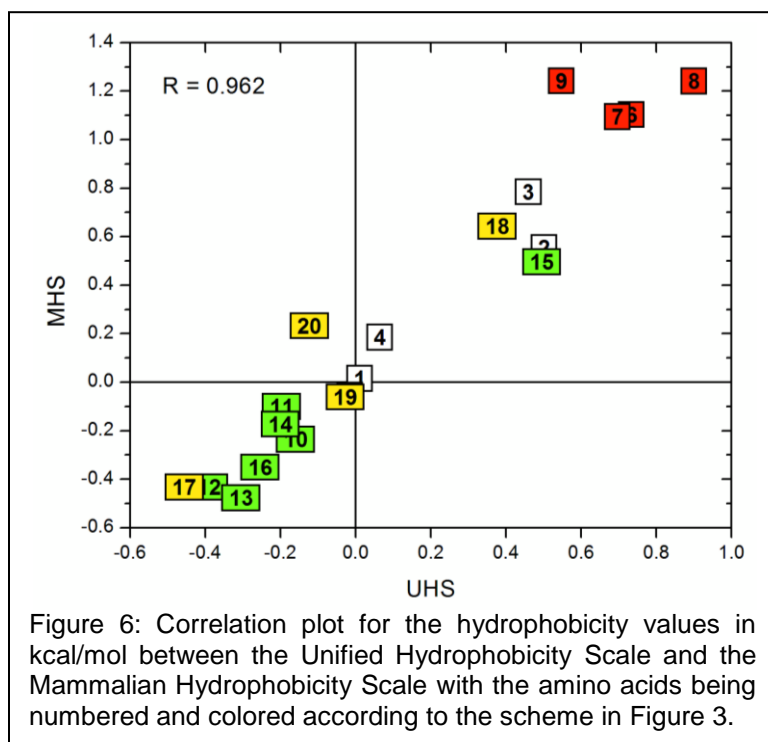
Comparing the UHS with the scale of Moon & Fleming shows an overall good correlation between these scales ($R = 0.76$). The scale from Moon is shifted 1.1 kcal towards negative energies indicating a shift towards the membrane bilayer. This can be explained by the derivation of the scales: Moon & Fleming investigate the transfer from water to bilayer where the side chains face the lipid phase. Our hydrophobicity scale does not distinguish between side chains facing lipids, protein interior, or even aqueous pores. The occurrence of charged residues in the protein interior and aqueous pores accounts for a shift of our scale towards water. Furthermore, this accounts for the range of our scale being smaller than the range of Moon's scale. Another major difference is that the UHS does not distinguish α -helical MPs and β -barrels whereas Moon's scale is derived from side chains facing the lipid in a β -barrel. Proline tends towards water in our scale whereas it is favoring the lipid phase in Moon's scale (see Figure 5). This could be attributed to the large number of α -helical MPs in our database that barely tolerate the characteristics of a helix-breaking residue in the membrane bilayer. Conversely, Gly favors the lipid bilayer in our case whereas it favors water in Moon's scale. This is likely



due to the fact that Glycines are often found in the interior of MPs due to their small size. At the same time, they avoid lipid contacts due to the lack of hydrogen bonding ability as would be possible if Glycine were found in water.

Applications of the scale

The UHS is optimized for usage as reference hydrophobicity values in computational protein structure prediction of α -helical and β -strand multi-span MPs. The scale fills a gap since most existing scales were optimized for usage with α -helical MPs only and distinguish only two states (TM and SOL). Furthermore, it can be used for the prediction of trans-membrane spans from genomic data (see below), in the early stages of a MP structure determination project when no structural information is available, or to assess the overall and local stability of folded multi-span MPs. To exemplify the latter, the UHS values for Trp, Tyr, and Phe were compared to Lukas Tamm's thermodynamic free energy changes (see Table 1 in [47]) by measuring the unfolding of wt OmpA and



OmpA mutants as described in [47]. The correlation coefficients are 0.715 for the single mutants, and 0.759 for the double mutants, excluding one outlier Y168A. No extensive conclusions can be drawn from the moderate agreement with these 11 data points for three amino acids, however, we believe that this possible application of the UHS warrants further investigation.

A hydrophobicity scale for mammalian α -helical membrane proteins

Since the amino acid occurrences are variable among organisms and there is interest in applying hydrophobicity scales to ORFs of mammalian genomes (specifically the human genome), a hydrophobicity scale was derived only from mammalian proteins. A database of 16 mammalian MPs was created as described in the Methods section and a mammalian scale (Mammalian Hydrophobicity Scale - MHS) was derived using two-fold cross-validation. The MHS was established for the two- and three-state scenario. The scale will be most applicable to α -helical proteins since the database used for derivation contained exclusively multi-span α -helical MPs. The hydrophobicity values and their standard deviations are given in Tables (I) and (III). The standard deviations are somewhat larger for the MHS when compared to the UHS because of the smaller dataset and the only two-fold cross-validation.

Overall amino acid abundance is quite similar between bacterial and mammalian MPs (data not shown) with an average difference of 0.18 when the amino acid abundances for all amino acids are normalized to 20. Ala, Asn, and Gly are somewhat more abundant in the bacterial dataset with differences of 0.32, 0.25, and 0.60 respectively, whereas Leu and Pro are more abundant in the mammalian dataset (-0.32 and -0.26). Furthermore, comparing the distribution between TM and SOL, it was found that Arg, Gly, Phe, and Tyr tend to be more abundant in the TM in the bacterial dataset

than in the mammalian dataset. The differences for these amino acids are 0.29, 0.29, 0.26, and 0.50 when the occurrence for each amino acid is normalized to 2.

Overall UHS and MHS are similar with deviations of 1.3 standard deviations on average and 3.8 standard deviations at maximum for Glu. Even though these seem to be relatively large changes, the change in actual numbers remains small, because of the small standard deviations for the UHS. A correlation plot of the two scales is shown in Figure 6 with a correlation coefficient of 0.962.

Comparison of the hydrophobicity values from the UHS with the MHS reveals that the largest deviations occur for Arg (UHS: 0.55 / MHS: 1.24), Asp (0.73 / 1.10), Glu (0.70 / 1.10), Leu (-0.30 / -0.48), Lys (0.90 / 1.24), and Tyr (-0.12 / 0.23). A test of the prediction accuracy of the MHS is available in Table (V) and in the supplement (Supplementary Table (V)). Briefly, the scale achieves an average prediction accuracy for SOL and TM of 83.1% in the two-state scenario (see Supplementary Table (V)) and 61% in the three-state scenario (see Table (V)).

It is important to note, that although this is the first mammalian hydrophobicity scale ever derived, care has to be taken in its application. The dataset used for its derivation is with only 16 MPs very small and does not guarantee very accurate hydrophobicity values. A refinement of the scale is to be expected when more MPs structures are elucidated. Nevertheless, we hope the MHS will find widespread application in the scientific community.

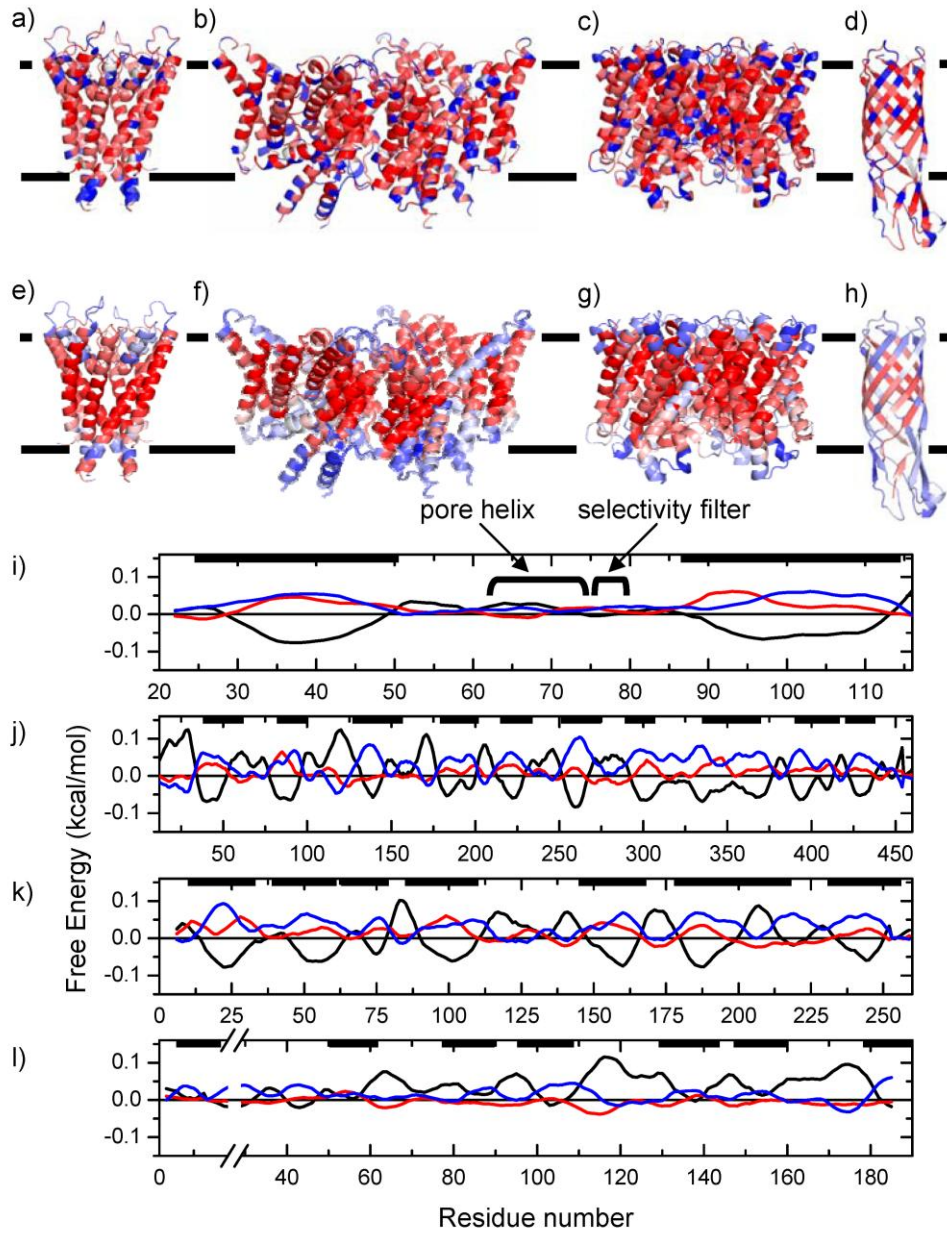


Figure 7: The derived UHS has been used to calculate the free energies (with a window length of 15 residues) for four examples: a), e) and i) KcsA – potassium channel (PDB code 1K4C), b), f) and j) CIC – chloride channel (PDB code 1KPK), c), g) and k) GlfP – Glycerol facilitator protein (PDB code 1LDI), and d), h) and l) OmpW – outer membrane protein W (PDB code 2F1T). The upper panels a) to d) show the three-state predictions from the sequence without any averaging procedure mapped onto the known crystal structure. The central panels e) to h) display the predictions for a window length of 15 residues. Dark blue indicates a prediction for the aqueous phase, white indicates interface, and dark red indicates a prediction for the TM. Lighter colors refer to a lower confidence in the prediction (as seen by smaller differences between the lowest and second lowest free energy in the bottom panels of the figure). The location of the membrane is displayed by the black lines. The lower panels i) to l) show the predictions of the free energies vs. the residue number as in panels e) to h) (black is TM, red is TR, and blue is SOL). Membrane locations are indicated by the black bars at the top. Panel i) shows one of four identical chains, j) shows one of two chains (chain A), k) shows one of four identical chains, and l) shows the whole protein sequence.

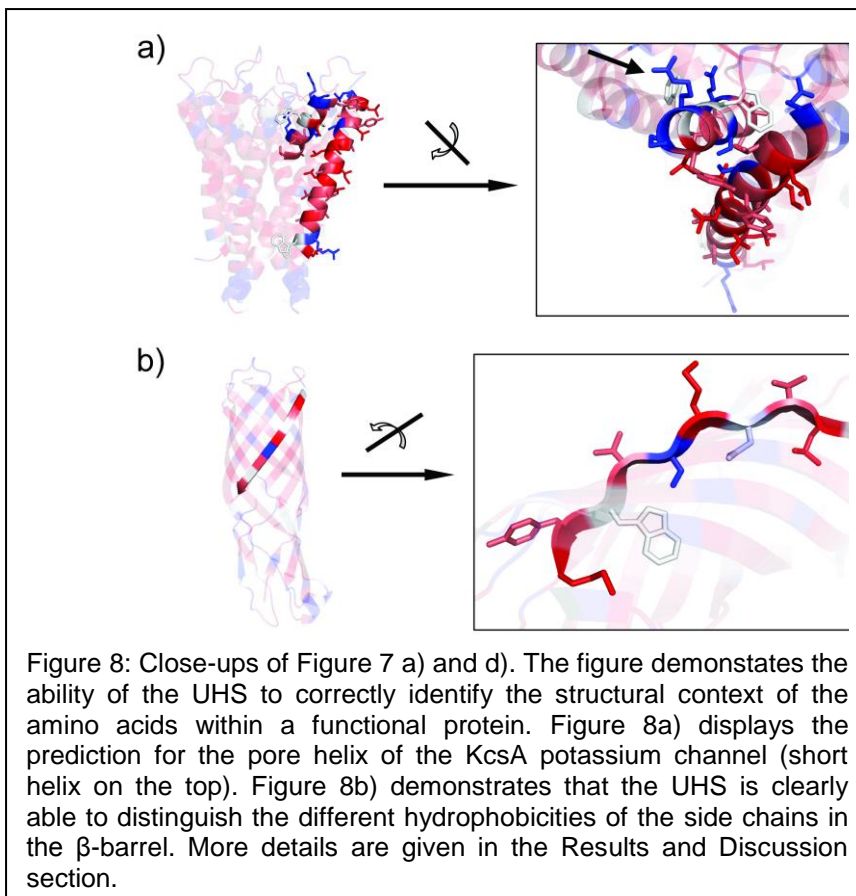
Four examples show that the UHS accurately reflects
the character of α -helical and β -barrel MPs

Four proteins not present in the MP database (used for derivation of the UHS) were used as examples to demonstrate the usefulness of the UHS and the prediction algorithm. The hydrophobicity values of the UHS for all residues in the sequence were mapped onto the known crystal structures in Figure 7a) to d). In Figure 7e) to h) the average free energy values were mapped onto the same crystal structures to illustrate the prediction of TM, TR, and SOL regions. Panels i) to l) in Figure 7 show the predicted free energies averaged over a window length of 15 residues (compare to panels e) to h)) vs. the residue number.

UHS distinguishes core TM α -helices from functional sites
in potassium channel KscA

The first example (Figure 7a), e) and i)) is the crystal structure of the potassium channel KscA which was determined by Roderick McKinnon et al. at a resolution of 2.0 Å (PDB code 1K4C). This example demonstrates the ability of the UHS to distinguish a typical hydrophobic, membrane-spanning α -helix from a functional site such as the pore α -helix and the selectivity filter. The pore α -helix is too short to even reach the center of the bilayer and the attached loop region returns to the extra-cellular side. This region is rich in polar amino acids as it is exposed to the SOL and has no direct contact to the membrane. The UHS clearly identifies the pore α -helix as an amphiphilic helix (short helix on the top of Figure 8a)) where the polar side chains point to the aqueous cavity (arrow) and the apolar side chains are in contact with other hydrophobic α -helices. This compares to a fully hydrophobic α -helix (long helix at the bottom) where all non-polar side chains interact with the hydrophobic environment. It illustrates that the UHS is well

able to identify the structural context of the individual residues even though no structural information is used in its derivation.



The prediction algorithm is clearly able to distinguish the membrane region from the sequence only. Figure 7i) demonstrates that the TMs are perfectly identified with a very high confidence and with their approximate lengths. However, the pore helix and selectivity filter of the protein have a small preference for the SOL region, which is indicated by the light blue α -helices at the top of the molecule (6e). This is not surprising because – as detailed above – both structural features are not in contact with the membrane at all but form a polar pore filled with water and ions (see Figure 8a)).

Chloride channel CIC

The second example (Figure 7b), f) and j)) is the crystal structure of the chloride channel CIC determined by Roderick McKinnon and co-workers at a resolution of 3.5 Å (PDB code 1KPK). In this case all TM α -helices are reliably identified and the predicted membrane locations agree well with the actual ones. However, the lengths deviate from the predicted spans slightly more than in the first example.

Glycerol facilitator protein GlfP

The third example (Figure 7c), g) and k)) is the crystal structure of the glycerol facilitator protein GlfP determined by Robert Stroud et al. at a resolution of 2.7 Å (PDB code 1LDI).

Generally, the UHS is able to identify polar residues within the trans-membrane domains of the protein which mostly face the interior of the protein and are therefore protected from the hydrophobic environment of the membrane bilayer. The α -helix at residues 204-217 is a short helix dipping into the membrane and the attached loop residues return to the same side of the membrane. The UHS clearly identifies this short α -helix as an amphiphilic helix where the polar side of this short helix faces inwards into one of the four channels of the homo-tetramer. Again, this shows the capability of the UHS to distinguish between regular, fully hydrophobic trans-membrane α -helices and functional sites in the protein.

Figure 7k) shows that the TM α -helices are correctly identified with a high reliability. The lengths of the α -helices agree well with the actual lengths except for the one α -helix at residue numbers 175-215 which is predicted to be too short. As discussed, under-prediction is unsurprising due to the amphiphilicity of this short α -helix

that faces one of the pores of the channel. In Figure 7g) this α -helix is the light blue helix on the lower left side of the protein.

The UHS identifies alternative hydrophobicity pattern in the β -barrel of the outer membrane protein W

The fourth example (Figure 7d), h) and l)) is a β -barrel protein which is the crystal structure of the outer membrane protein W (OmpW) determined by van den Berg and Tamm et al. at a resolution of 3.0 Å (PDB code 2F1T). Figure 8b) shows that the UHS correctly identifies the polarity of the side chains pointing to the aqueous interior of the β -barrel whereas apolar side chains face the hydrophobic milieu of the membrane bilayer. It can be seen, that consecutive side chains along the β -strand alternately face the polar interior and apolar membrane environment. These patterns are nicely detected by the UHS (Figure 7d)). This demonstrates the efficiency of the UHS to depict structural features of the amino acids although no structural information is required for the application of the UHS.

Figure 7 h) and l) show, that for β -barrels the prediction has lower confidence and only some of the TM spans are identified. However, this behavior is expected because this simple window function is insufficient to reliably identify trans-membrane spans if an alternating pattern of hydrophobicity values complicates the prediction, as is the case for β -barrel proteins. To optimize the prediction accuracies for β -barrels, we plan to utilize the UHS as an input for an artificial neural network or a hidden Markov model in the future.

In summary these four examples illustrate the ability of the UHS scale to accurately reflect the hydrophobicity of a certain residue within a folded protein. In particular the scale distinguishes nicely between the core of the protein and functional

sites and highlights the alternating hydrophobicity pattern seen in β -barrel proteins. The scale is therefore suitable as input for MP secondary and tertiary structure prediction tools. This was demonstrated by usage of the UHS for prediction of TM spans from sequence only. Although the prediction accuracies are somewhat lower for β -barrel proteins when using such a simple averaging scheme, they are better than random (60% average prediction accuracy in the two-state scenario and 45% in the three-state scenario). For α -helical bundles the prediction accuracy increases up to 77% in the two-state scenario and even 66% in the three-state scenario (compared to 33% for a random prediction).

Conclusions

In this chapter we derive a three-state Unified Hydrophobicity Scale (UHS) exclusively from multi-span membrane proteins of known structure. The database of membrane proteins contained both α -helical and β -barrel proteins. The absolute hydrophobicity values in the UHS range between -0.46 for Phe and 0.90 for Lys. This reduced amplitude when compared to most experimentally derived scales was previously observed for other knowledge-based scales and results from averaging over a wide variety of structural contexts, in particular different degrees of burial in the protein core or different types of secondary structure. This makes the UHS applicable for the prediction of trans-membrane spans from the proteins primary sequence only.

The UHS is derived only in a membrane depth-dependent manner and does not take into account the solvent accessible surface area in the membrane. Since pore-forming proteins were included in the database for derivation, the penalty for transferring polar or charged residues into the bilayer are likely under-estimated, if these residues are exposed to the lipid. However, the UHS is applicable as an unbiased average hydrophobicity value for an amino acid that is equally valid for both α -helical and β -

strand multi-span membrane proteins, which is of high importance for computational protein structure prediction. Furthermore, it can be used in the early stages of a membrane protein structure determination project when no structural information is available. The overall and local stability of folded multi-span membrane proteins can be assessed as demonstrated for OmpA. It can also be used for the prediction of trans-membrane spans from genomic data. For this application we specifically derived a hydrophobicity scale only from mammalian proteins (Mammalian Hydrophobicity Scale - MHS) to be applicable to mammalian genomes or the human genome in particular. This scale is optimized for α -helical multi-span membrane proteins and reaches average accuracies of up to 83%.

In general, we observe a bias in many existing hydrophobicity scales when applied to folded, multi-span membrane proteins. This offset applies to both the reference point of the scale (which we chose to be multi-span membrane proteins) as well as the absolute size of the free energy values. These biases are imposed by the respective experimental setup and may not exist in other applications. It emphasizes the importance to carefully choose the hydrophobicity scale based on the given task.

The UHS scale was tested for predicting trans-membrane spans from primary sequence only. It was found that prediction improves when free energies are averaged over a window of 9-17 amino acids with a triangular weight giving the central amino acid the highest influence. For a two-state prediction scenario (classifying an amino acid as being either in the TM or SOL) it was found that in comparison to other hydrophobicity scales the UHS yields an average prediction accuracy of 73%. The scales of GES (71%) and Janin (70%) perform almost as well. For a three state scenario that includes a TR region the UHS performs at an accuracy of 57%. This is significantly better than the WW scale (50% correct classifications).

Application of the UHS scale to four proteins illustrates its ability to very accurately map the hydrophobicity of a certain residue within a folded protein. In particular, the scale distinguishes nicely between the core of the protein and functional sites and highlights the alternating hydrophobicity pattern seen in β -barrel proteins. The scale is therefore suitable as input for membrane protein secondary and tertiary structure prediction tools. This was demonstrated by the usage of the UHS for prediction of TM spans from the sequence only.

When predicting TM spans in these four proteins, the lengths and positions of the predicted α -helices agree well with the actual lengths and locations. For β -barrel proteins the prediction tool is less reliable because the alternating hydrophobicity pattern thwarts the effectiveness of the simple averaging procedure. This is a general observation for β -barrel proteins across the scales and does not imply that the UHS poorly describes the characteristics of β -barrel proteins. It rather emphasizes the fact that the type of window function is not optimal for the prediction of β -barrels.

References

- [1] D. Eisenberg, R.M. Weiss, T.C. Terwilliger, The hydrophobic moment detects periodicity in protein hydrophobicity, *Proc Natl Acad Sci U S A*, 81 (1984) 140-144.
- [2] T.P. Hopp, K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, *Proc Natl Acad Sci U S A*, 78 (1981) 3824-3828.
- [3] D.M. Engelman, T.A. Steitz, A. Goldman, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu Rev Biophys Chem*, 15 (1986) 321-353.
- [4] W.C. Wimley, S.H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, *Nat Struct Biol*, 3 (1996) 842-848.
- [5] W.C. Wimley, T.P. Creamer, S.H. White, Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides, *Biochemistry*, 35 (1996) 5109-5124.

- [6] S.H. White, W.C. Wimley, Membrane protein folding and stability: physical principles, *Annu Rev Biophys Biomol Struct*, 28 (1999) 319-365.
- [7] S. Jayasinghe, K. Hristova, S.H. White, Energetics, stability, and prediction of transmembrane helices, *J Mol Biol*, 312 (2001) 927-934.
- [8] J. Janin, Surface and inside volumes in globular proteins, *Nature*, 277 (1979) 491-492.
- [9] J. Kyte, R.F. Doolittle, A simple method for displaying the hydrophobic character of a protein, *J Mol Biol*, 157 (1982) 105-132.
- [10] M. Punta, A. Maritan, A knowledge-based scale for amino acid membrane propensity, *Proteins*, 50 (2003) 114-121.
- [11] H.R. Guy, Amino acid side-chain partition energies and distribution of residues in soluble proteins, *Biophys J*, 47 (1985) 61-70.
- [12] A. Senes, D.C. Chadi, P.B. Law, R.F. Walters, V. Nanda, W.F. Degrad, E(z), a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices, *J Mol Biol*, 366 (2007) 436-448.
- [13] C. Chothia, The nature of the accessible and buried surfaces in proteins, *J Mol Biol*, 105 (1976) 1-12.
- [14] Y. Nozaki, C. Tanford, The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale, *J Biol Chem*, 246 (1971) 2211-2217.
- [15] J. Hine, P.K. Mookerjee, Intrinsic Hydrophilic Character of Organic Compounds - Correlations in Terms of Structural Contributions, *Journal of Organic Chemistry*, 40 (1975) 292-298.
- [16] R.V. Wolfenden, P.M. Cullis, C.C.F. Southgate, Water, Protein Folding, and the Genetic-Code, *Science*, 206 (1979) 575-577.
- [17] G.D. Rose, Prediction of chain turns in globular proteins on a hydrophobic basis, *Nature*, 272 (1978) 586-590.
- [18] G.D. Rose, S. Roy, Hydrophobic basis of packing in globular proteins, *Proc Natl Acad Sci U S A*, 77 (1980) 4643-4647.

- [19] J.L. Fauchere, K.Q. Do, P.Y. Jow, C. Hansch, Unusually strong lipophilicity of 'fat' or 'super' amino-acids, including a new reference value for glycine, *Experientia*, 36 (1980) 1203-1204.
- [20] R. Wolfenden, L. Andersson, P.M. Cullis, C.C. Southgate, Affinities of amino acid side chains for solvent water, *Biochemistry*, 20 (1981) 849-855.
- [21] D.H. Wertz, H.A. Scheraga, Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule, *Macromolecules*, 11 (1978) 9-15.
- [22] B. Robson, D.J. Osguthorpe, Refined models for computer simulation of protein folding. Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor, *J Mol Biol*, 132 (1979) 19-51.
- [23] T. Hessa, N.M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, G. von Heijne, Molecular code for transmembrane-helix recognition by the Sec61 translocon, *Nature*, 450 (2007) 1026-U1022.
- [24] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S.H. White, G. von Heijne, Recognition of transmembrane helices by the endoplasmic reticulum translocon, *Nature*, 433 (2005) 377-381.
- [25] C.P. Moon, K.G. Fleming, Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers, *Proc Natl Acad Sci U S A*, 108 (2011) 10174-10177.
- [26] C. Chothia, Hydrophobic bonding and accessible surface area in proteins, *Nature*, 248 (1974) 338-339.
- [27] V. Yarov-Yarovoy, J. Schonbrun, D. Baker, Multipass membrane protein structure prediction using Rosetta, *Proteins-Structure Function and Bioinformatics*, 62 (2006) 1010-1025.
- [28] V. Yarov-Yarovoy, D. Baker, W.A. Catterall, Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels, *Proc Natl Acad Sci U S A*, 103 (2006) 7292-7297.
- [29] P. Barth, J. Schonbrun, D. Baker, Toward high-resolution prediction and design of transmembrane helical protein structures, *Proc Natl Acad Sci U S A*, 104 (2007) 15682-15687.
- [30] T. Beuming, H. Weinstein, A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins, *Bioinformatics*, 20 (2004) 1822-1835.

- [31] <http://icb.med.cornell.edu/crt/ProperTM/>.
- [32] M.J. Sippl, Knowledge-based potentials for proteins, *Curr Opin Struct Biol*, 5 (1995) 229-235.
- [33] A.M. Poole, R. Ranganathan, Knowledge-based potentials in protein design, *Curr Opin Struct Biol*, 16 (2006) 508-513.
- [34] G.E. Tusnady, Z. Dosztanyi, I. Simon, PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank, *Nucleic Acids Res*, 33 (2005) D275-278.
- [35] G.E. Tusnady, Z. Dosztanyi, I. Simon, TMDet: web server for detecting transmembrane regions of proteins by using their 3D coordinates, *Bioinformatics*, 21 (2005) 1276-1277.
- [36] G.E. Tusnady, Z. Dosztanyi, I. Simon, Transmembrane proteins in the Protein Data Bank: identification and classification, *Bioinformatics*, 20 (2004) 2964-2972.
- [37] G.L. Wang, R.L. Dunbrack, PISCES: recent improvements to a PDB sequence culling server, *Nucleic Acids Research*, 33 (2005) W94-W98.
- [38] G.L. Wang, R.L. Dunbrack, PISCES: a protein sequence culling server, *Bioinformatics*, 19 (2003) 1589-1591.
- [39] M.C. Wiener, S.H. White, Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of x-ray and neutron diffraction data. III. Complete structure, *Biophys J*, 61 (1992) 434-447.
- [40] D. Shortle, Composites of local structure propensities: evidence for local encoding of long-range structure, *Protein Sci*, 11 (2002) 18-26.
- [41] M. Levitt, A simplified representation of protein conformations for rapid simulation of protein folding, *J Mol Biol*, 104 (1976) 59-107.
- [42] D. Eisenberg, R.M. Weiss, T.C. Terwilliger, W. Wilcox, Hydrophobic Moments and Protein-Structure, *Faraday Symposia of the Chemical Society*, (1982) 109-120.
- [43] W.M. Yau, W.C. Wimley, K. Gawrisch, S.H. White, The preference of tryptophan for membrane interfaces, *Biochemistry*, 37 (1998) 14713-14718.
- [44] M.B. Ulmschneider, M.S. Sansom, A. Di Nola, Properties of integral membrane protein structures: derivation of an implicit membrane potential, *Proteins*, 59 (2005) 252-265.

- [45] S.H. White, G. von Heijne, Transmembrane helices before, during, and after insertion, *Curr Opin Struct Biol*, 15 (2005) 378-386.
- [46] S.H. White, W.C. Wimley, Hydrophobic interactions of peptides with membrane interfaces, *Biochim Biophys Acta*, 1376 (1998) 339-352.
- [47] H. Hong, S. Park, R.H. Jimenez, D. Rinehart, L.K. Tamm, Role of aromatic side chains in the folding and thermodynamic stability of integral membrane proteins, *J Am Chem Soc*, 129 (2007) 8320-8327.

CHAPTER 4

Improved prediction of trans-membrane spans in proteins using an Artificial Neural Network¹

Introduction

Membrane proteins (MPs) account for about 30% of the proteins in the human genome and are involved in many essential functions in the cell. For instance, they act as transporters, participate in signaling pathways and function as ion-channels. Even though almost 50,000 protein structures are deposited in the ProteinDataBank (PDB), only about 900 belong to the class of MPs. This discrepancy reflects the difficulty of crystallizing MPs and they often exceed the size limitation for NMR spectroscopy. In contrast, the structures of MPs are arguably easier to predict computationally because of the constraints the membrane imposes on their fold [1].

First attempts to identify membrane spanning regions along the sequence utilize hydrophobicity scales. A free energy value of transfer from a polar medium (the cytosol) to an apolar medium (the membrane) is assigned to each of the 20 amino acids. Depending on the preference of an amino acid for a specific environment the sign of this transfer free energy value changes. For the prediction of trans-membrane (TM) spans the transfer free energy values are added over a sequence window (Figure 1).

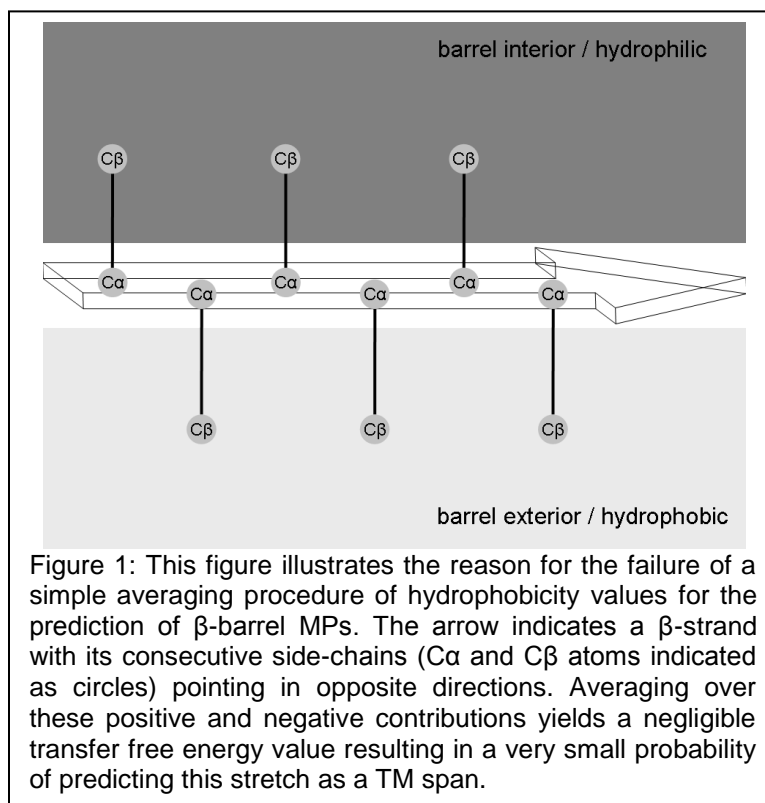
There is a wealth of hydrophobicity scales available that were derived using experimental (for example Wimley & White [2-3] or GES [4]), knowledge-based (UHS [5]), and consensus approaches (Kyte & Doolittle [6]). The scales are mostly derived

¹ This chapter has been published in: Koehler, J., R. Mueller, and J. Meiler, *Improved prediction of trans-membrane spans in proteins using an Artificial Neural Network*. IEEE Comp. Intel. Bioinf. Comp. Biol., 2009: p. 68-74.

considering two phases: solution (SOL) and membrane (TM). Only two scales [5, 7] include a third interface or transition region (TR). Potentials considering the depth of the residue in the membrane bilayer have also been reported for α -helical MPs [8-9].

The differences between various hydrophobicity scales can be explained by the different experimental setups used during their derivation. Wimley & White for instance examine unfolded peptides in solution and membrane bilayer [2, 7] whereas Hessa et al. consider folded proteins [8, 10].

Hydrophobicity scales that include an interface region between solution and membrane are rare even though three-state scales have a higher information content than two-state scales. In addition, three-state scales are able to provide information about the location of the polar headgroups of the membrane lipids, which are distinctly



different than the soluble phase. Further, three-state scales can identify amphipathic helices located in the interface region. Wimley & White experimentally derived a

hydrophobicity scale using penta-peptides that were unfolded in all three phases [2, 7]. For this reason the unsaturated hydrogen-bonds in the membrane bilayer lead to a bias of this scale towards solution. On average ~50% of the residues are correctly predicted in this three-state scenario. We derived a knowledge-based hydrophobicity scale from a database of known MP structures containing both α -helical and β -barrel proteins [5]. This Unified Hydrophobicity Scale (UHS) yields accuracies of ~57% in the three-state prediction scenario. Both scales were tested on a database containing both α -helical bundles and β -barrel proteins. A Mammalian Hydrophobicity Scale (MHS) was derived from 16 α -helical bundles and yields accuracies of ~61% tested on an only α -helical database [5].

Subsequent specialized prediction tools for TM spans use machine learning techniques such as Hidden Markov Models (HMMs), Artificial Neural Networks (ANNs), or Support Vector Machines (SVMs). According to Cuthbertson et al. [11] Split4 [12], TMHMM2 [13], and HMMTOP2 [14-15] are the most successful TM α -helix prediction tools available. Split4 [12] uses basic charge clusters and amino acid attributes to define the correct topology of the helices. TMHMM2 [13] is an HMM trained on a dataset of 160 both single- and multi-spanning proteins and has according to their developers 97% accuracy. HMMTOP [14] utilizes the evolutionary information of multiple-sequence alignments and is based on the notion that topology is governed by the difference of the amino acid distributions in different parts of the protein rather than the amino acid composition itself. The successor HMMTOP2 [15] incorporates experimental information into the topology prediction. Other methods include PhDhtm [16] (which uses two consecutive ANNs and multiple-sequence alignments), TMMOD [17] (which is based on TMHMM, but differs in training procedure and loop models), and TopCons [18] (a consensus prediction server combining five different predictors).

The most successful methods for β -barrel proteins are according to Bagos et al. [19] HMM-B2TMR [20] and PROFtmb [21-22], both HMM-based methods. HMM-B2TMR is sequence-profile based and therefore uses multiple-sequence alignments. A dynamic programming algorithm is employed for optimization of the location of TM segments. PROFtmb is also profile-based and is trained on eight non-redundant β -barrels. Their developers state a four-state accuracy of 86%. Bagos and co-workers tested the performance of various combinations of β -barrel predictors and implemented the best-performing consensus predictor as ConBBPRED [19].

The objective of this work is to establish the first integrated tool that identifies both α -helical and β -strand TM spans in a single three-state prediction for the residue being either in TM, TR, or SOL region. Advantage of this method is that sequences can be screened for TM spans with a single tool. Furthermore, synergistic effects during the ANN training lead to an increased prediction accuracy.

Methods

Creation of the databases of non-redundant protein structures

For the MP database all TM chains from the PDBTM [23] were culled using the PISCES server [24-25] with the following parameters: sequence identity $\leq 25\%$, resolution 3\AA , R-factor 0.3, sequence length 40-10,000 residues, non-X-ray entries as well as C α -only entries were included, and the PDB was culled by chains. Thereafter, structures derived from electron-microscopy data were excluded due to low resolution resulting in a database of 102 proteins with 136 polypeptide chains. The PDB files were downloaded from the PDBTM.

For the definition of the TM, TR, and SOL regions a fixed membrane thickness of 20\AA (TM region) followed by a 10\AA TR region was used. Furthermore, a 2.5\AA gap region

between the TM/TR regions as well as TR/SOL regions was introduced to more cleanly distinguish between the different environments (see ref. [5]). This procedure was implemented rather than using the membrane thickness given by the PDBTM (determined by the TMDET algorithm [26]) in order to avoid a recurrent influence of this predicted membrane thickness onto our method. The resulting database contained 28,379 residues in total, 9,510 residues being in the TM region, 9,079 classified as TR, and 9,790 classified as SOL. A total of 3,882 residues residing in the gap region were excluded from the training process to minimize noise due to incorrect assignment to regions.

Even though the MP database contained a large fraction of soluble residues a soluble protein database was established to account for different properties of soluble proteins that are not equally represented by the soluble parts of the MPs (like solvent-accessible surface area, compactness, and length of secondary structure elements).

For the soluble protein database the entire PDB was culled with the PISCES server [24] using the same parameters as above with two exceptions. Due to the much larger size of the database a resolution limit of 2Å was used. Moreover, we excluded non-X-ray and C α -only entries. The resulting database contained 3,499 proteins with a total of 3,623 polypeptide chains and 820,485 residues.

Both the MP as well as the soluble protein database were used as a basis for the input to the ANN.

Knowledge-based free energies for secondary structure type and membrane location were used as input

The MP database served as a basis for the derivation of knowledge-based free energies. The procedure is the same as described in [5] but updated databases allowed

for more data to be included. Briefly, three-state free energies for the regions TM, TR, and SOL were derived by normalizing the amino acid frequencies in each region to 20. The propensities P [27] were then calculated by

$$P = \frac{\text{number}(\text{region}, AA) / \text{number}(\text{region})}{\text{number}(AA) / \text{number}(\text{total})} \quad (1)$$

and the free energies ΔG were computed using

$$\Delta G = -RT \ln P \quad (2)$$

with R being the gas constant, and $T=293\text{K}$.

The same procedure was applied to obtain the three-state free energies for the secondary structures helix, strand, and coil. The nine-state free energies for each combination of region and secondary structure type were calculated as in the three-state scenario but normalizing the amino acid occurrences to nine instead of three.

We chose to include the free energies for the secondary structure types for the prediction of the TM region since the two phenomena are interrelated: when a nascent polypeptide chain in solution reaches the membrane interface the influence of the altered dielectric environment (as described by the free energies) leads to an increased formation of backbone hydrogen bonds and therefore to the formation of secondary structure.

The obtained free energies for these different scenarios were taken as input parameters for the ANN. Furthermore, several amino acid properties such as the steric parameter, polarizability, volume, iso-electric point, the solvent-accessible surface area

[28], and the position-specific scoring matrices obtained from PSI-BLAST [29] were used as input parameters as they increased prediction accuracy in previous experiments [28]. PSI-BLAST was run with three iterations and an E-value cutoff of 0.001.

Training procedure

For each dataset (i.e. for each residue) the above mentioned input parameters were employed over a sequence window of 31 residues. Therefore (20 property descriptors + 20 numbers in the PSI-BLAST profile) x 31 residues = 1240 inputs were used for each dataset. The MP database (28,379 residues) served as a basis for the TM and TR region datasets, whereas the soluble protein database (820,485 residues) together with the MP database were used for the SOL region datasets. To construct the input files the residues were randomly chosen from the databases. In addition, the residues were chosen as to equally represent TM, TR, and SOL residues using an over-sampling procedure. Three dataset sizes of 9,000, 90,000, and 450,000 datasets (i.e. residues) were used for training where the training was started on the smallest dataset and consecutively increased to larger dataset sizes.

This balancing procedure was chosen to avoid an intrinsic bias of the method to predict one region over the other. It also maximizes the entropy in the training data and therefore the information content added by the ANN prediction.

For the training procedure the datasets were shuffled and then split into three subsets: 80% were used for training, 10% for monitoring the training progress, and 10% as an independent test set. Two ANNs were trained with 32 and 64 nodes in the hidden layer, respectively. The ANN with 64 nodes performed best in this case and the results are shown for this network.

The ANN is a feed-forward network with bias neurons trained with back-

propagation of errors. Other network architectures have not been tested. In initial training phases the resilient propagation algorithm [30] displayed accelerated training behavior, faster convergence and higher robustness with respect to the initial training parameters than simple propagation. Therefore, the ANN was trained using the resilient propagation algorithm whereas simple propagation was used for final optimization of the weights.

Four examples illustrate the performance of the prediction tool

The ANN prediction was applied to four MPs not included in the training phase: two α -helical bundles and two β -barrel proteins. The crystal-structures of the potassium channel KcsA (PDB ID 1k4c) elucidated by Rod McKinnon at a resolution of 2Å was chosen as first helical example protein. Furthermore, we chose lens aquaporin-0 (PDB ID 2b6p) in the open state that was determined by Walz and co-workers at 2.4Å. Unusual structural features in both proteins are half-helices with their adjacent loops returning to the extra-membrane region. As β -barrel proteins the Outer Membrane Protein W (OmpW – PDB ID 2f1t) crystallized by Tamm and van den Berg at 3Å and the NMR structure of OmpA (PDB ID 2ge4) determined by Tamm and Bushweller were selected.

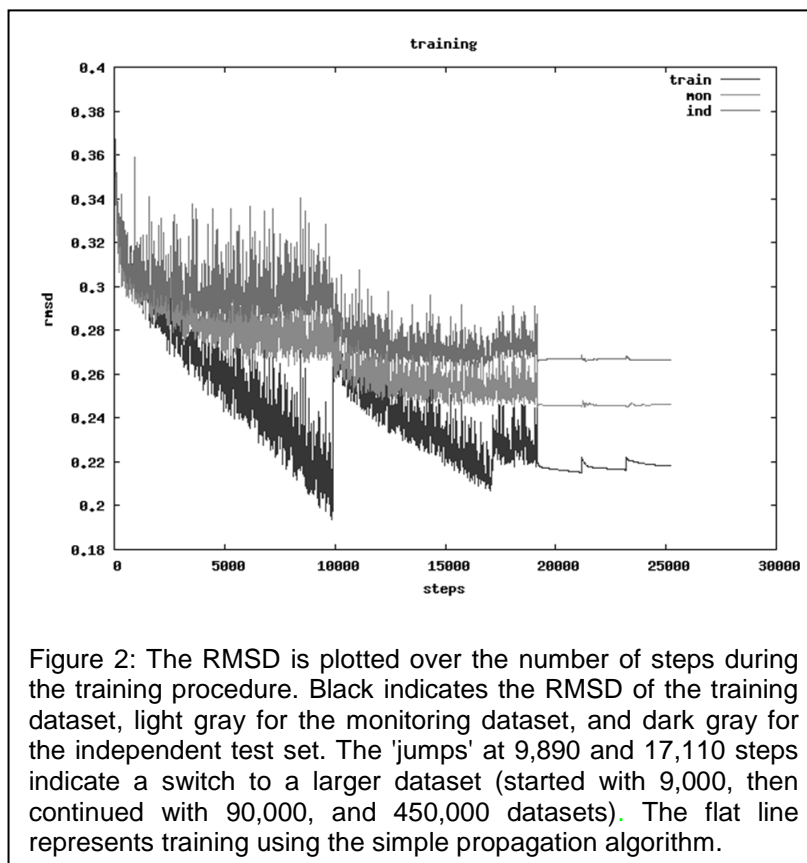
Results and Discussion

Most of the TM prediction methods are specialized methods for α -helical proteins. β -strand TM spans, on the other hand, are much more difficult to predict because a simple averaging procedure is less effective when consecutive side-chains alternate in facing the polar interior and the apolar exterior of the barrel (see Figure 1). This obstacle can be overcome using machine learning techniques such as ANNs, HMMs, or SVMs that are capable of recognizing such alternating patterns while

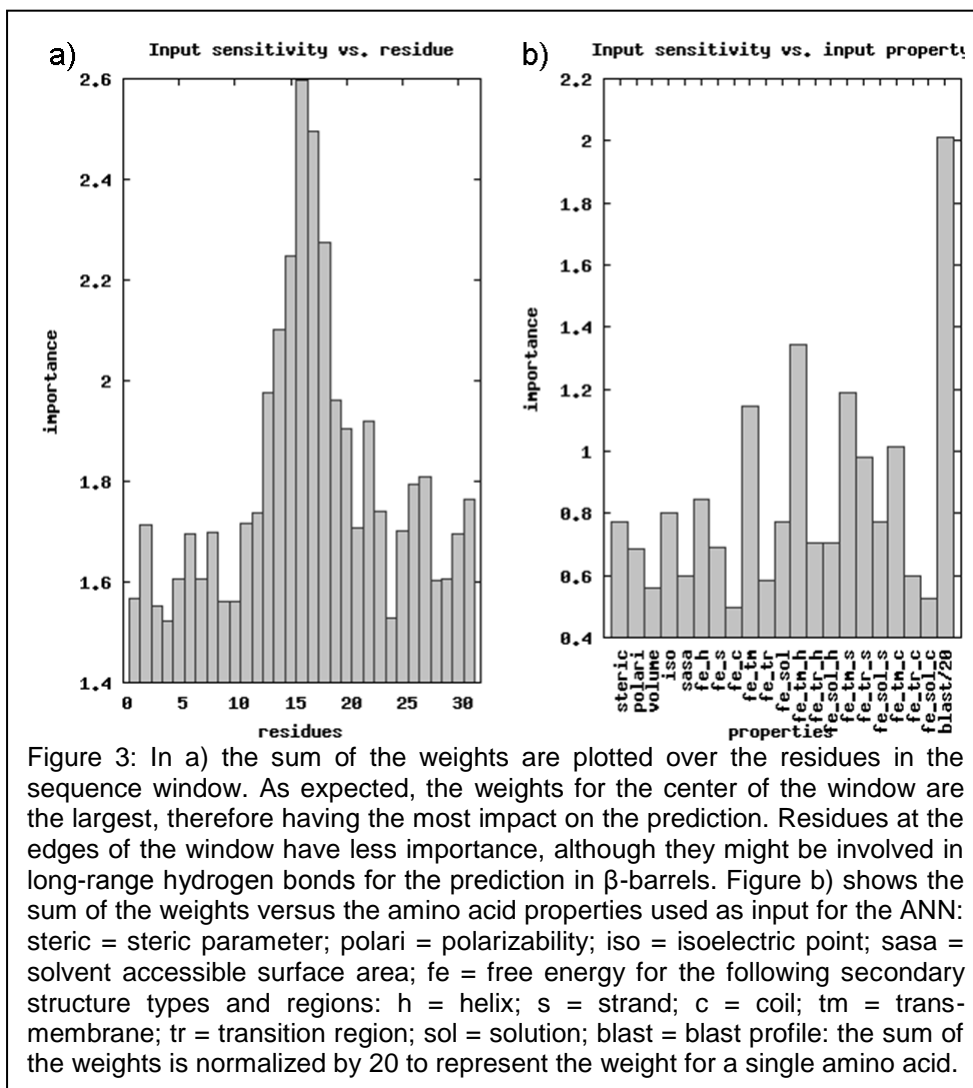
distinguishing between α -helices and β -strands at the same time. In addition, α -helices require ~ 19 residues to cross the lipid bilayer while β -strands require only ~ 9 residues. This difference results in a different optimal sequence window size for simple linear averaging strategies. However, non-linear functions like ANNs can be optimized on a single larger window (here 31 residues) to work equally well for both scenarios.

Resilient propagation accelerates training

The ANN is implemented within the Bio-Chemical-Library developed in the Meiler laboratory (www.meilerlab.org) and written in the C++ programming language. It serves as a framework for a wide variety of biomedical applications, such as *de novo* protein tertiary structure prediction [31-32] and virtual high-throughput screening. The training was started with a small dataset (9,000 datasets). Subsequently the number of datasets



was increased to 90,000 and 450,000 datasets. The ANN was trained on each dataset using the resilient propagation algorithm until the error of the monitoring dataset was minimized (see Figure 2). Afterwards the ANN was trained in simple propagation mode for several 100 iterations to reach the RMSD minimum. This procedure became necessary as resilient propagation is known to display unstable minimization behavior close to minima in the target function [30].



Trans-membrane free energies are important for training

Figure 3a) shows the sum of the input sensitivities plotted over the 31 residues in the sequence window used for input. The input sensitivity is defined as a partial derivative of an output value with respect to an input variable. The values are determined numerically after ANN training is completed. As expected, the center of the sequence window has the highest impact as reflected in the increased input sensitivities. This represents the importance of the pattern immediately adjacent to the residue of interest within an α -helix or β -strand. The sensitivities converge to a smaller constant value towards the edges of the window which reflects the significance of long-range interactions within the protein. Such interactions are attributed to backbone hydrogen-bonds that stabilize β -barrel proteins as well as helix-helix contacts in α -helical bundles. The large window size facilitates capturing part of this effect. The optimal window size was determined by testing window sizes of 15, 23, 31, 39, and 47 residues with 31 residues performing best.

Figure 3b) shows the sum of the input sensitivities for the individual input properties. The highest sensitivity is observed for the PSI-BLAST position-specific scoring matrices with a sensitivity of 2.0. The profile reflects evolutionary information of the protein sequence which is important for the distinction between α -helical bundles and β -barrel proteins. Furthermore, it is essential for the identification of TM spans because the likelihood for mutations contained in this profile provides information about the exposure to the polar solvent, membrane bilayer, or protein core.

Considerable influence have the free energies for the TM region, both in the three-state scenario (sensitivities TM = 1.2, TR = 0.6, SOL = 0.8) and in the nine-state scenario in conjunction with secondary structure types (see below). When considering secondary structure types the free energy for helices (sensitivity = 0.8) contains more

information than for strands (0.7). Both have a higher weight than the free energy for coil residues (0.5). Similarly, if the free energies for the secondary structure types are summed over TM, TR, and SOL regions, strands contain with 3.0 more information than helices with 2.8.

The sensitivities for the free energies of the TM region in the 9-state scenario sum up to 3.6, whereas for the TR and SOL these sums are smaller (2.3 and 2.0, respectively). The sum of the six amino acid properties (excluding the PSI-BLAST matrices) is 3.4 reflecting a smaller per property influence when compared to the free energy values. It is known, that the environment of residues plays a critical role in the formation of secondary structure. We therefore speculate that the ANN uses the free energy patterns efficiently for the identification of TM spans.

Per-residue accuracy is highest for soluble region

We have shown previously [5] that the per-residue accuracy of the Wimley-White hydrophobicity scale is ~50% for the three-state prediction scenario using a simple averaging strategy. The UHS correctly classifies up to 57% of the residues. However, it was also shown, that this averaging procedure is much less effective when identifying TM β -strands in β -barrel proteins due to the alternating hydrophobicities of consecutive amino acids (Figure 1). Furthermore, such a simple scheme is not able to incorporate different window lengths for helices and strands, as discussed above.

Table I shows the percentage of per-residue predictions for the three regions TM, TR, and SOL using the ANN method. The data is shown for both the independent and the training dataset. The diagonal matrix elements indicate correct predictions whereas off-diagonal elements represent false classifications. The agreement for the SOL is broken down into the accuracy for soluble proteins and MPs. It can be seen that the

highest agreement is achieved in SOL for soluble proteins where 92% of the residues in the independent dataset and 91% of the residues in the training dataset are correctly predicted. For MPs the percentage agreement is lower with 75% for the independent and 81% for the training dataset. The interface region has an agreement of 75% and 77% correct predictions, respectively. This is expected since the interface region has two adjacent regions that detract correct predictions. In addition, the usage of a fixed membrane thickness will reduce prediction accuracy in this region [1]. The TM region has an agreement of 73%. Therefore, the prediction accuracies for MPs are similar for all of the three regions. The smaller agreement in the SOL for soluble parts of MPs than for soluble proteins has been observed earlier [5] and can be attributed to the difficulty of accurately pinpointing the exact beginnings and ends of the TM spans. In other words, the residues on the membrane surface are more often predicted as TM although they belong to the SOL region. Such residues are absent in soluble proteins resulting in a better performance.

We chose a fixed membrane thickness for training of our method because we wanted to avoid a circular influence of other algorithms (that predict the membrane thickness from protein structures) onto our prediction tool. If, however, the used membrane thickness is too short to cover all hydrophobic protein surface, incorrect predictions for some of the residues in the membrane may result because these residues are then likely predicted to reside in the TR. Furthermore, our algorithm disregards assumptions about solvent accessible surface area and only considers depth in the membrane, even though some of the membrane proteins have large pores with aqueous interior. This might lead to incorrect predictions since polar or charged side chains are energetically favorable to face the aqueous pore but would be unfavorable if only membrane depth is considered.

Table I: Prediction Accuracy

Accuracies of the prediction method on the independent and training datasets with the percentage of predicted residues in these regions. The percentage of correctly predicted residues is 79.6% for the independent and 81.3% for the training dataset. sol = solution, tr = transition region, tm = trans-membrane.

| | | prediction | | |
|-------------------------|--------------------|------------|------|------|
| | | sol | tr | tm |
| observed independent | sol (SOL proteins) | 92.2 | 5.5 | 2.3 |
| | sol (MPs) | 74.9 | 17.7 | 7.4 |
| | tr | 10.4 | 74.7 | 14.9 |
| | tm | 5.4 | 22.1 | 72.6 |
| observed training | sol (SOL proteins) | 91.4 | 6.2 | 2.4 |
| | sol (MPs) | 80.5 | 14.0 | 5.5 |
| | tr | 15.5 | 76.8 | 7.8 |
| | tm | 4.2 | 19.4 | 76.5 |

Four examples illustrate a successful prediction

The algorithm was tested on four examples: two α -helical proteins and two β -barrel proteins. Only the sequence of the proteins was used as input and the prediction was mapped onto the known protein structures as shown in Figure 4.

Panel a) shows the crystal structure of the potassium channel KcsA (PDB ID 1k4c). The figure shows the correct prediction of the membrane location. The structure contains a half-helix (selectivity filter) with the adjacent loop returning to the extra-cellular side of the channel (see close-up). Since the correct prediction of such half-helices represents a particular challenge to the algorithm this indicates the ANNs ability to identify the correct location of these pore helices in the membrane and interface region. For this example the ANN predicts 83% of the residues correctly. 95% of the TR residues and 90% of the TM residues are correctly identified. The unified hydrophobicity

scale in conjunction with the simple window function implemented earlier [5] identifies 68% of the residues correctly with an accuracy of 21% for SOL, 55% for TR, and 90% for TM.

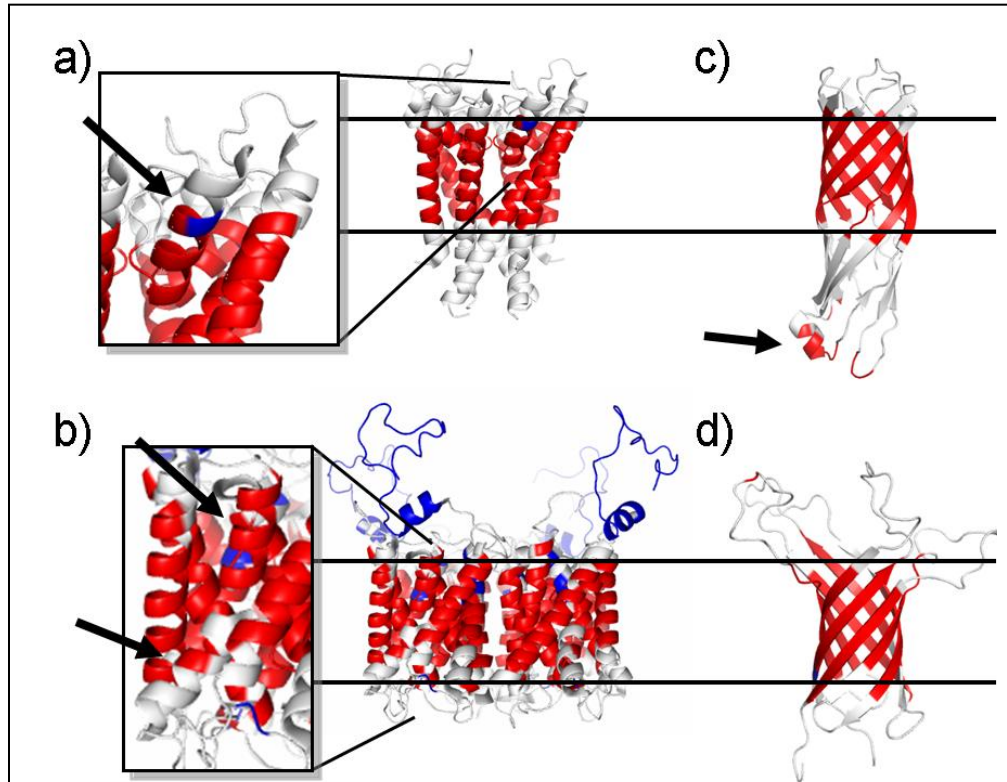


Figure 4: The algorithm was applied to the sequence of four proteins and mapped onto the known protein structures. a) KcsA potassium channel (PDBID 1k4c) – 83% of the residues correctly predicted; b) lens aquaporin-0 (PDBID 2b6p) – 75% correctly predicted residues; c) Outer membrane protein W (PDBID 2f1t) – 73% accuracy; d) Outer membrane protein A (PDBID 2ge4) – 81% accuracy. Red indicates a prediction for being in TM, white represents a prediction for TR, and blue indicates a prediction for SOL. The membrane location is indicated by the black lines. The arrow in the close-up of panel a) points to the pore helix of the tetrameric channel which is a half-helix with the adjacent loop (representing the selectivity filter) returning to the extra-cellular side.

The prediction for the crystal structure of lens aquaporin-0 in the open state (PDB ID 2b6p) is shown in panel b). Again, all of the three regions are correctly identified. Overall, 75% of the residues are correctly classified. The accuracy is 93% for SOL, 81% for TR, and 68% for TM. The lower agreement in TM is due to the fact that there are isolated residues in the membrane that are predicted to be in SOL. One of the two half-helices is correctly predicted to be in the membrane (as seen by the upper arrow in the

inset). The half-helices dip into the membrane and the adjacent loops return to the extra-membrane region. This represents a particular challenge for prediction algorithms since TM helices are usually much longer (~19 residues) and can be confused with hydrophobic regions in soluble proteins. This difficulty might be addressed by feeding the output of this prediction algorithm into a second ANN to obtain the final output. Such a procedure was applied in PSIPRED, one of the best secondary structure prediction algorithms to date [33].

Panel c) shows the structure of the Outer Membrane Protein W (OmpW – PDB ID 2f1t). The algorithm is able to correctly identify the location of TM strands. Overall, 73% of the residues are correctly identified with an accuracy of 100% for the TR, and 86% for the TM. The soluble region is not predicted as such since 71% of these residues are predicted to be in TR and 29% in the TM. This is indicated by the small helix at the bottom (see arrow) which is predicted to be in TM although it resides in SOL. For comparison, the unified hydrophobicity scale in conjunction with the simple window function implemented earlier [5] identifies 43% of the residues correctly with an accuracy of 29% for SOL, 75% for TR, and 27% for TM.

Panel d) shows the Outer Membrane Protein A (OmpA – PDB ID 2ge4). Also this example suggests that the algorithm is able to distinguish the different regions for β -barrel proteins. In this protein the overall prediction accuracy averages to 81%. 97% of the TR residues are correctly identified and 77% of the TM residues are correctly predicted. The algorithm identifies all of the 12 soluble residues as being in TR. However, they constitute only ~7% of the total residues in this small β -barrel.

Conclusion

An artificial neural network was trained to predict the location of trans-membrane

spans from the protein sequence. In contrast to earlier prediction tools which are specialized for either α -helical or β -barrel proteins, the method represents the first tool that predicts trans-membrane spans for both classes of proteins.

The artificial neural network was trained on a membrane protein and soluble protein database. As input served several amino acid properties and the position-specific scoring matrices from PSIBLAST. Furthermore, we used the free energies for (1) the three-state scenario of the residue being in helix, strand, and coil, (2) the three-state scenario of the residue being in trans-membrane, transition, and soluble region, and (3) the nine-state scenario with pair-wise combinations of the former. We found that the position-specific scoring matrices and the free energies for the trans-membrane region (both for individual secondary structure types as well as combined) had the highest impact on the prediction. In contrast, other amino acid properties were less important for the prediction.

Soluble residues were correctly predicted in 92% of the cases, for interface residues the accuracy was 75%, and for trans-membrane residues 73%. Therefore, in the three-state scenario, on average 79% of the residues are correctly predicted, which is a remarkable improvement compared to the prediction using simple hydrophobicity scales.

The algorithm was applied to four membrane proteins, two of α -helical nature and two β -barrel proteins. In these examples the prediction tool is able to classify 78% of the residues correctly. Even though half-helices are intrinsically difficult to predict, the predictor correctly identified two of three half-helices as trans-membrane spans. Since the tested proteins lack large soluble domains, the network has difficulties to identify short soluble loops and correctly classifies them only for one of the four examples.

References

- [1] J.U. Bowie, Solving the membrane protein folding problem, *Nature*, 438 (2005) 581-589.
- [2] W.C. Wimley, T.P. Creamer, S.H. White, Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides, *Biochemistry*, 35 (1996) 5109-5124.
- [3] S.H. White, W.C. Wimley, Membrane protein folding and stability: physical principles, *Annu Rev Biophys Biomol Struct*, 28 (1999) 319-365.
- [4] D.M. Engelman, T.A. Steitz, A. Goldman, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu Rev Biophys Chem*, 15 (1986) 321-353.
- [5] J. Koehler, N. Woetzel, R. Staritzbichler, C.R. Sanders, J. Meiler, A unified hydrophobicity scale for multispan membrane proteins, *Proteins*, 76 (2009) 13-29.
- [6] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J Mol Biol*, 157 (1982) 105-132.
- [7] W.C. Wimley, S.H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, *Nat Struct Biol*, 3 (1996) 842-848.
- [8] T. Hessa, N.M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, G. von Heijne, Molecular code for transmembrane-helix recognition by the Sec61 translocon, *Nature*, 450 (2007) 1026-U1022.
- [9] A. Senes, D.C. Chadi, P.B. Law, R.F. Walters, V. Nanda, W.F. Degrado, E(z), a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices, *J Mol Biol*, 366 (2007) 436-448.
- [10] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S.H. White, G. von Heijne, Recognition of transmembrane helices by the endoplasmic reticulum translocon, *Nature*, 433 (2005) 377-381.
- [11] J.M. Cuthbertson, D.A. Doyle, M.S. Sansom, Transmembrane helix prediction: a comparative evaluation and analysis, *Protein Eng Des Sel*, 18 (2005) 295-308.
- [12] D. Juretic, L. Zoranic, D. Zucic, Basic charge clusters and predictions of membrane protein topology, *J Chem Inf Comput Sci*, 42 (2002) 620-632.

- [13] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol*, 305 (2001) 567-580.
- [14] G.E. Tusnady, I. Simon, Principles governing amino acid composition of integral membrane proteins: application to topology prediction, *J Mol Biol*, 283 (1998) 489-506.
- [15] G.E. Tusnady, I. Simon, The HMMTOP transmembrane topology prediction server, *Bioinformatics*, 17 (2001) 849-850.
- [16] B. Rost, R. Casadio, P. Fariselli, C. Sander, Transmembrane helices predicted at 95% accuracy, *Protein Sci*, 4 (1995) 521-533.
- [17] R.Y. Kahsay, G. Gao, L. Liao, An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes, *Bioinformatics*, 21 (2005) 1853-1858.
- [18] <http://topcons.net/>.
- [19] P.G. Bagos, T.D. Liakopoulos, S.J. Hamodrakas, Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method, *BMC Bioinformatics*, 6 (2005) 7.
- [20] P.L. Martelli, P. Fariselli, A. Krogh, R. Casadio, A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins, *Bioinformatics*, 18 Suppl 1 (2002) S46-53.
- [21] H.R. Bigelow, D.S. Petrey, J. Liu, D. Przybylski, B. Rost, Predicting transmembrane beta-barrels in proteomes, *Nucleic Acids Res*, 32 (2004) 2566-2577.
- [22] H. Bigelow, B. Rost, PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins, *Nucleic Acids Res*, 34 (2006) W186-188.
- [23] <http://pdbtm.enzim.hu/>.
- [24] G.L. Wang, R.L. Dunbrack, PISCES: a protein sequence culling server, *Bioinformatics*, 19 (2003) 1589-1591.
- [25] G. Wang, R.L. Dunbrack, Jr., PISCES: recent improvements to a PDB sequence culling server, *Nucleic Acids Res*, 33 (2005) W94-98.
- [26] G.E. Tusnady, Z. Dosztanyi, I. Simon, TMDet: web server for detecting transmembrane regions of proteins by using their 3D coordinates, *Bioinformatics*, 21 (2005) 1276-1277.

- [27] D. Shortle, Composites of local structure propensities: evidence for local encoding of long-range structure, *Protein Sci*, 11 (2002) 18-26.
- [28] J. Meiler, M. Muller, A. Zeidler, F. Schmaschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, *Journal of Molecular Modeling*, 7 (2001) 360-369.
- [29] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, 25 (1997) 3389-3402.
- [30] A.D. Anastasladis, G.D. Magoulas, M.N. Vrahatis, New globally convergent training scheme based on the resilient propagation algorithm, *Neurocomputing*, 64 (2005) 253-270.
- [31] N. Alexander, M. Bortolus, A. Al-Mestarihi, H. McHaourab, J. Meiler, De novo high-resolution protein structure determination from sparse spin-labeling EPR data, *Structure*, 16 (2008) 181-195.
- [32] E. Dong, J. Smith, S. Heinze, N. Alexander, J. Meiler, BCL::Align-sequence alignment and fold recognition with a custom scoring function online, *Gene*, 422 (2008) 41-46.
- [33] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol*, 292 (1999) 195-202.

CHAPTER 5

Simultaneous prediction of protein secondary structure and trans-membrane spans¹

Introduction

The prediction of secondary structure (SS) and transmembrane (TM) segments is the first step towards structural characterization of proteins. Its importance is emphasized by the fact that alternative experimental methods either yield less information or are much more laborious: CD spectroscopy only yields a percentage of secondary structure types in the protein and CSI data from NMR requires the peak assignments which is a time-consuming task. The outputs of SS and TM prediction tools are a basic requirement for programs performing sequence alignments, fold recognition, and de novo protein structure prediction. Furthermore, it facilitates the design of EPR experiments to find an optimal position for MTSL spin labels [1] or to select detergents to screen for membrane protein NMR experiments based on the hydrophobic thickness of the protein.

Hence, both predictions are typically executed using a variety of SS and TM prediction methods that have been developed in parallel (see below). However, the formation of SS and TM spans is interrelated because the occurrence of secondary structure is greatly increased in the TM region. Peptides or proteins can exist in a disordered state in solution because of their ability to form hydrogen bonds with the surrounding water. When these peptides are inserted into the membrane the hydrophobic environment drives them to form hydrogen bonds to saturate backbone amide protons and carbonyl oxygens. Since water is unavailable in this environment, it

¹ This chapter has been submitted to *Proteins, Structure, Function, and Bioinformatics*.

forms hydrogen bonds with itself therefore forming secondary structure. BCL::Jufo9D leverages this interrelation by simultaneously predicting SS and TM segments.

Machine learning techniques are used for secondary structure prediction

All modern methods for SS prediction use machine learning techniques (see [2]) such as ANNs, HMMs or SVMs. These algorithms are pattern recognition techniques to associate a given input (e.g. the sequence information of a protein) to an output (e.g. the structural information such as SS or TM spans). For supervised learning the output is provided during the training process using structural information of determined protein structures. When training is complete, the algorithms are able to predict the unknown information for a given input – i.e. the secondary structure for a target sequence. The use of machine learning approaches in SS prediction has been pioneered by Rost and co-workers through the development of their PhD program [3-4].

For soluble proteins SS prediction tools usually provide a three-state probability for each residue being either in helix, strand, or coil. Accuracy is often reported as Q_3 value which is the percentage of correctly predicted SS if the state with the highest predicted probability is compared to the experimentally determined SS. Accuracies of up to 80% are achieved [5] with Psipred [6-7] being one of the most accurate SS prediction tools available [5]. PsiPred is a two-stage feed-forward ANN that was trained on a sequence database of soluble proteins with the position-specific scoring matrices (PSSM) from PSI-BLAST [8] as an input. Jufo [9-10] is an ANN that uses dimension-reduced amino acid representations to predict the SS of soluble proteins. It is trained on a database of 430 soluble peptides from the FSSP database [11] using an input window of 31 residues. The SS prediction tool PROFPHD [4, 12-13] as part of the PredictProtein server is also based on ANNs. It is a three-layer feed-forward ANN trained on sequence-to-structure and structure-to-structure context that uses a multiple sequence alignment

and global amino acid composition as inputs. The developers state a three-state accuracy of 76%.

Trans-membrane span prediction methods are specialized to either α -helical proteins or β -barrels

Early attempts to predict the location of TM spans in membrane proteins (MPs) involve averaging hydrophobicity values over a sequence window. Many different hydrophobicity scales have been developed using a variety of experimental [14-20], theoretical [21-25], and consensus approaches [26-28], some of them are reviewed in [25]. Most of the scales consider the two states membrane bilayer and solution. The scales of Wimley & White as well as a recently developed knowledge-based unified hydrophobicity scale (UHS, [25]) take a third interface region into account. Considering an interface region is important since the dielectric environment characterized by the polar lipid head-groups is distinctly different from the aqueous solution as well as from the membrane core region. Aromatic residues like Tyr or Trp as well as amphipathic α -helices usually reside there [16, 25]. Predicting the location of TM-spans using simple averaging schemes for hydrophobicity values achieves accuracies up to 73% in the two-state scenario (membrane bilayer and solution) and ~60% in the three-state scenario (with interface region) [25]. Considerable improvements were achieved by the application of machine learning approaches; however, these methods are specialized to either TM α -helical bundles or β -barrels.

For identification of TM spans in α -helical MPs OCTOPUS [29] is one of the best methods available. It uses four separately trained ANNs to identify one of the four states (membrane, interface, loop, globular) at the residue level and combines the predictions globally using a Hidden Markov Model (HMM). It is designed as a topology predictor and is able to model reentrant/membrane dipping regions and TM hairpins. The prediction

accuracies on an independent benchmark dataset were reported to be as high as 94% to identify the correct topology. Other available methods use mainly HMMs (such as TMHMM [30] and TMMOD [31]), SVMs (such as MEMSAT-SVM [32]) or a consensus of multiple SS prediction servers, such as ConPredII [33-34].

For identification of TM β -barrels, TMbeta-Net [35] is one of few methods available. It consists of an ANN that was trained on 13 outer membrane proteins with a jack-knife approach for cross-validation. Other methods, mostly HMMs, include ProfTMB [36-37] as part of the PredictProtein server [38] and TMBHMM [39].

Another method worth mentioning is the comprehensive protein structure prediction server Proteus2 [40] from Wishart and co-workers. It employs several secondary structure, TM span prediction, and homology modeling tools over 7 residue fragments and maps the output onto the sequence from which a jury-of-expert approach identifies the most optimal output. The difference of Proteus2 to our method is that Proteus2 combines the output of several specialized servers. It uses TMB-Hunt [41] to determine whether a protein is a TM β -barrel, and if it is, then a more specialized approach is used to identify the location of the TM spans. In case TMB-Hunt returns a false identification, or if the specialized servers return contradicting outputs, the prediction accuracy will ultimately suffer. In contrast, our method is trained on a wide variety of sequences being able to identify different regions within a sequence. This alleviates the necessity to combine multiple contradicting outputs into a single prediction. Furthermore, Proteus2 achieves highest accuracies if homologous sequence fragments are found. This is not a requirement for BCL::Jufo9D.

To overcome the specialization of TM span prediction tools for α -helical bundles or β -barrels we developed an ANN that predicts a three-state probability distribution describing residue environment for both types of MPs [42]. BCL::Jufo9D integrates the prediction of three SS states (helix, strand, coil) with the three states for protein

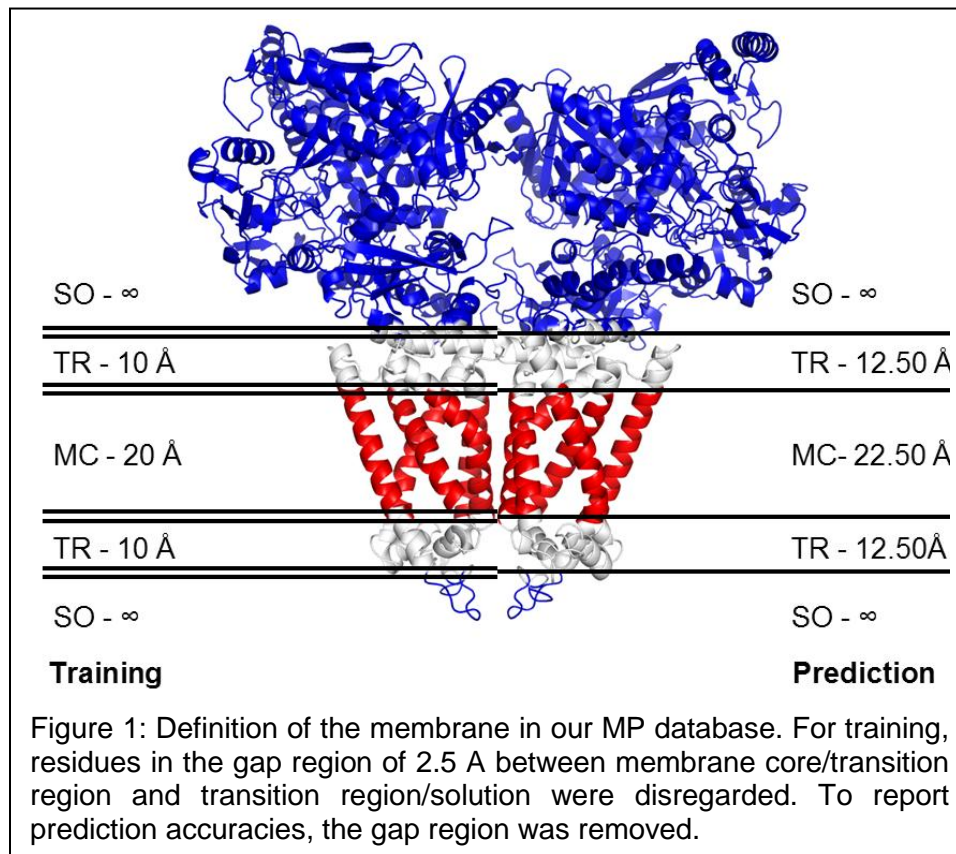
environment (membrane, interface, solution) into a 3x3 probability vector identifying the most likely of these states for each residue in the sequence. The ANN was trained on a database of 226 MP chains in 177 MPs and 6223 soluble protein chains in 6048 soluble proteins. The approach achieves per residue accuracies of 70.3% in a nine state prediction scenario (compared to a random prediction of 11%) for the independent dataset. Furthermore, correlating secondary structure formation and membrane placement not only streamlines the prediction of SS and TM regions in proteins, but it also bears the potential to study conformational switches.

Methods

Establishing the membrane protein database

A list of all membrane protein chains, for which a structure has been determined, were downloaded from the PDBTM [43-44] website (Nov. 2011). Similar sequences were excluded by culling this list with the PISCES protein sequence culling server [45-46] with a percentage sequence identity $\leq 30\%$, resolution 0 – 3 Å, R-factor 0.25, sequence length 40 – 10,000 residues, non X-ray entries as well as CA-only chains were included. EM-structures were excluded. BCL::PDBConvert (Woetzel, N. submitted) was used to convert non-natural amino acids into their natural counterparts and to transform the protein into the membrane coordinate frame using the xml files from the PDBTM website. The membrane was defined by the membrane normal that is specified by the z-coordinate in the PDB file with the membrane center being at $z = 0$. The thicknesses are 20 Å for the membrane core and 10 Å for the transition region on either side of the membrane. Residues in the 2.5 Å gap regions between membrane core and transition region or transition region and solution were disregarded to obtain more distinct regions for the ANN to identify (Figure 1). DSSP [47] (version of 2011) was executed for all PDB structures to obtain a consistent secondary structure identification. Helices below five

residues and strands below three residues were regarded as coil to facilitate the distinction between more distinct secondary structure elements and coil regions. This procedure resulted in a list of 226 chains in 177 membrane proteins.



Establishing the database of soluble proteins

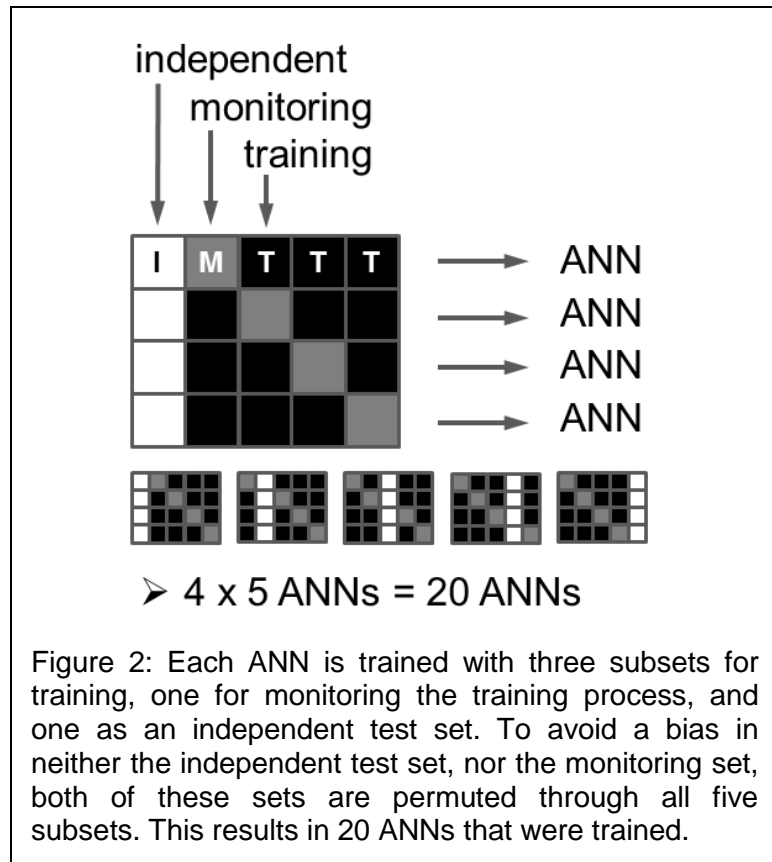
A pre-compiled list of PDB chains that was culled using the PISCES protein sequence culling server [45-46] was downloaded from the PISCES website (date 12/02/2011). The list contained sequences with a percentage sequence identity $\leq 30\%$, resolution 0 – 2 Å, R-factor 0.25, sequence length 40 – 10,000 residues, non X-ray entries as well as CA-only chains were excluded. Trans-membrane sequences were excluded from this list. Similar as for the membrane proteins, BCL::PDBConvert (Woetzel, N. submitted) was used to convert non-natural amino acids into their natural counterparts and DSSP was used to standardize secondary structure identification.

Again, helices shorter than five residues and strands shorter than three residues were regarded as coil. The result was a list of 6,223 chains in 6,048 soluble proteins.

Dataset splitting and cross-validation

The databases were split into five subsets for cross-validation. For the membrane proteins, α -helical bundles as well as β -barrels were distributed as equally as possible. The soluble proteins were distributed randomly.

To train a single ANN, three of the five subsets were used for training (see Figure 2), one subset was used for monitoring the training process to avoid overtraining. The fifth subset was used as an independent test set for computing the prediction accuracies. 20 networks were trained such that the independent as well as the monitoring subset could be permuted through the five datasets (Figure 2).



To report the prediction accuracies for complete protein subsets or single proteins, an average of the network outputs was computed whereas only ANNs were used that contained the subsets or individual proteins in the independent dataset. This ensures that the reported accuracies originate from ANNs that were not trained on that particular subset or protein.

Free energies are used as inputs to the ANNs

Figure 3 shows the input parameters used: (a) several amino acid properties such as steric parameter, volume, polarizability, iso-electric point, solvent-accessible surface area [10]; (b) the free energies for secondary structure type (helix, strand, coil), residue environment (membrane bilayer, interface, solution) [25] and all possible combinations of both; (c) the position-specific scoring matrices from PSIBLAST [8] after six iterations (see [48]). For each residue all of these parameters were collected over a sequence window of 31 residues. The input window size of 31 residues was found by testing all odd window sizes between 15 and 39 residues.

In addition, "global" protein parameters were considered for each residue: (a) the number of residues in the protein chain; (b) the oligomeric state (monomer vs. oligomer); (c) the amino acid parameters, the free energies and the position-specific scoring matrices averaged over the number of residues in the protein chain. This resulted in

(31 residues x (20 numbers from PSSM + 20 amino acid properties)) + (2 parameters: oligomeric state, length) + (40 averages) = 1282 input parameters
to represent the residue at the center of the window.

Balanced training avoids prediction bias towards over-represented states

The datasets (the term "dataset" corresponds to the input and output parameters for each residue in a protein sequence) were randomized and balanced for each protein

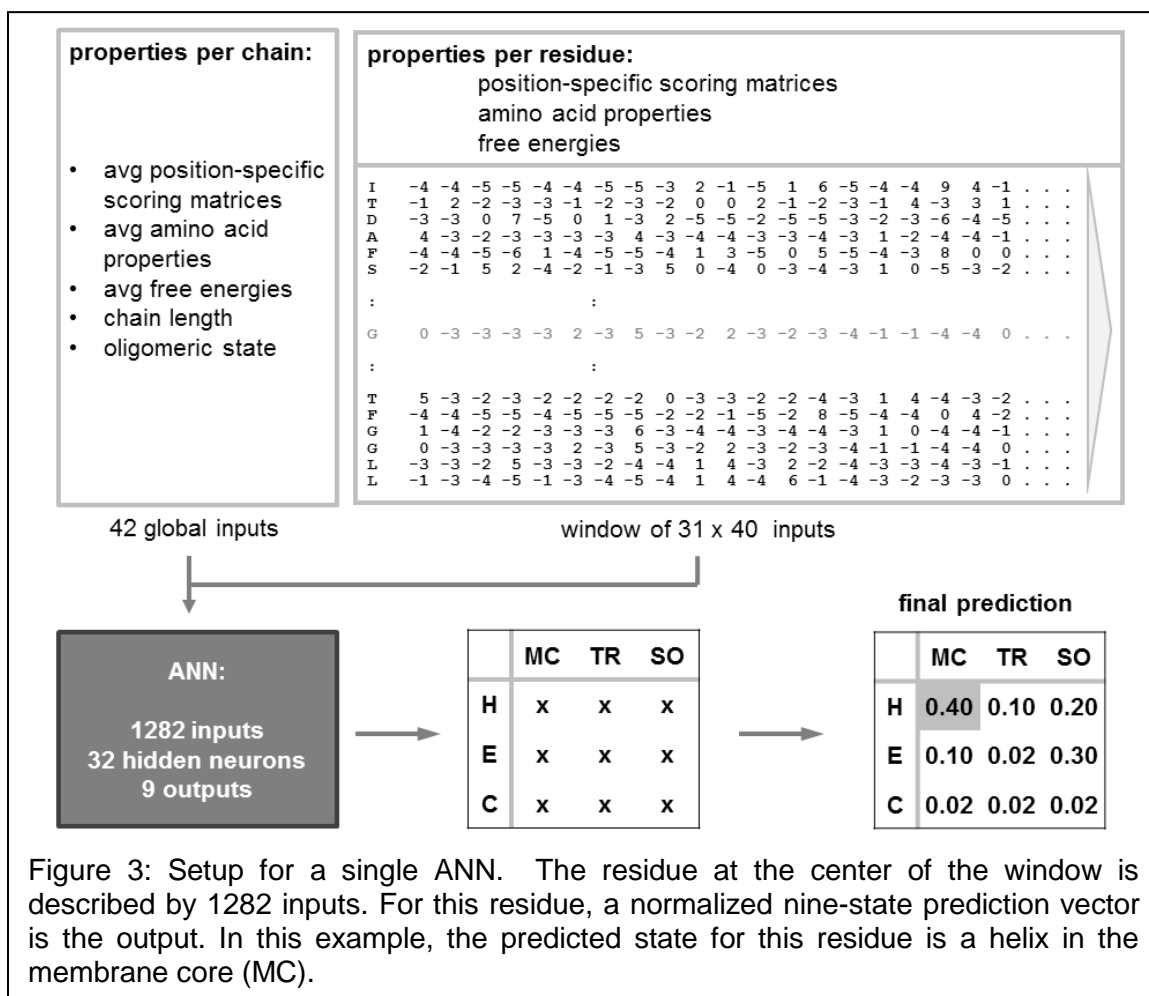


Figure 3: Setup for a single ANN. The residue at the center of the window is described by 1282 inputs. For this residue, a normalized nine-state prediction vector is the output. In this example, the predicted state for this residue is a helix in the membrane core (MC).

subset independently. For balancing, an over-sampling procedure was used to represent each of the nine states equally often to avoid a bias in the predictions towards the more abundant states. This approach also increases the entropy in the input data and maximizes the information content in the ANN.

The ANNs were three-layer feed-forward networks with a bias neuron, a sigmoidal activation function and back-propagation of errors. The hidden layer contained 32 neurons as identified by testing 4, 8, 16, 32, 64, and 128 neurons. The three training subsets combined contained 270,000 datasets. The training protocol consisted of three consecutive steps using a simple propagation algorithm: (1) update after each dataset with momentum $\alpha = 0$ and the learning rate $\eta = 10^{-3}$; (2) batch update with momentum α

= 0.5 and the learning rate $\eta = 5 \cdot 10^{-6}$; (3) update after each step with momentum $\alpha = 1$ and the learning rate $\eta = 5 \cdot 10^{-6}$.

As a post-processing step the output of the four ANNs was averaged that used the same independent subset.

Results

BCL::Juf09D achieves nine-state accuracies of 70.3%

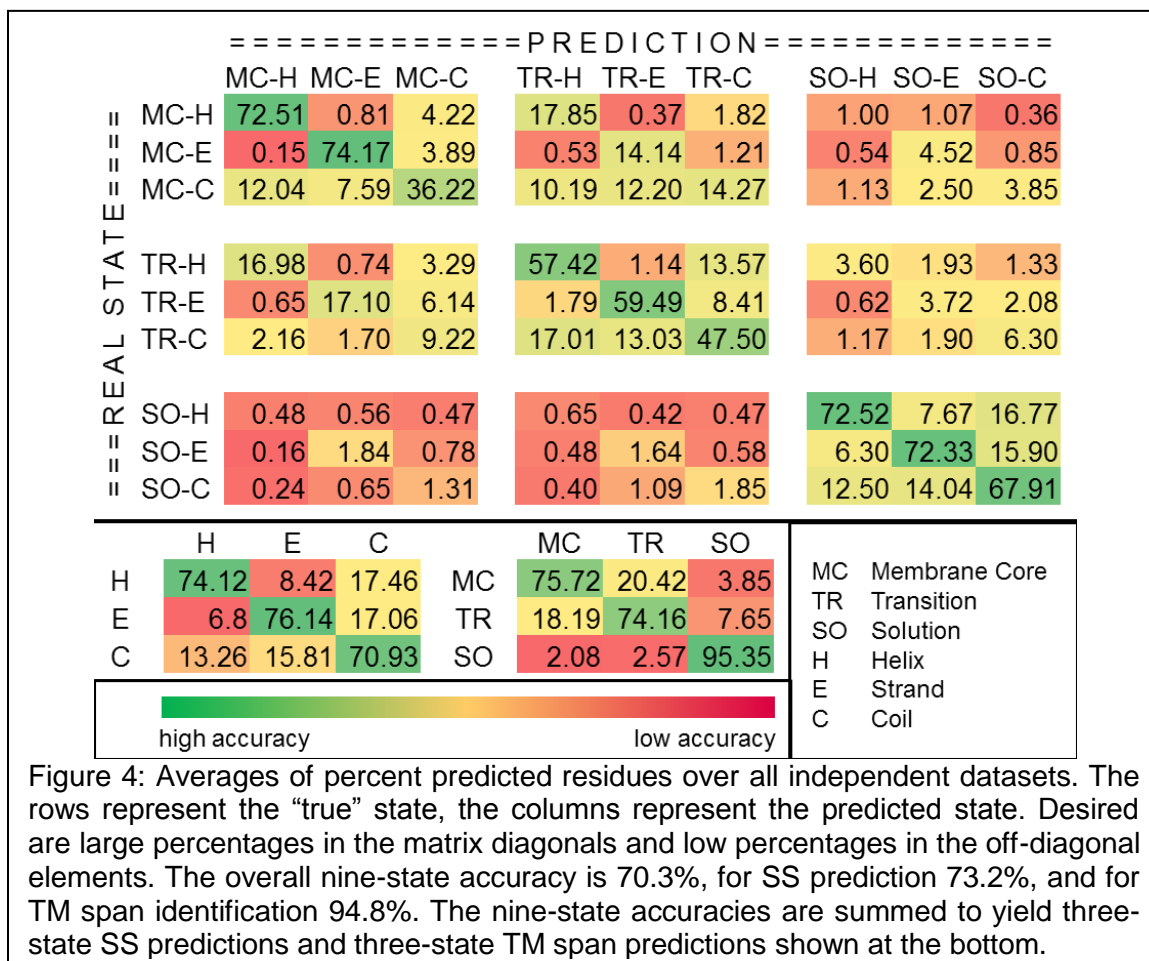
Figure 4 shows the percentage of predicted residues for all nine states whereas the percentages are averages over all independent datasets. The rows correspond to the “true” state as represented in the structure and the columns correspond to the predicted state. Ideally, highest percentages should be seen in the matrix diagonal. As seen from this data, “true” soluble states are identified most accurately since percentages for predicted membrane and transition states range from 0.16 – 1.85%. Helices and strands in solution and the membrane core have highest prediction accuracies ranging from 72.33 – 74.17%. The states in the transition region have lower accuracies ranging from 47.5 – 59.49%. Since membrane proteins have a wide range of hydrophobic thicknesses, the exact location of the transition region is difficult to identify and hence these states have to sacrifice prediction accuracy in favor of both the membrane as well as soluble states. Accuracies of coil states, irrespective of their environment, are always lower than helix or strand prediction accuracies (36.22% in membrane core, 47.5% in transition region, 67.91% in solution).

Three-state secondary structure is identified at 73.2%

When occurrences in the nine states are added together to represent three-state SS prediction, accuracies for helix predictions are at 74.16%, for strand at 76.14%, and for coil at 70.93% (Figure 5). Again, these accuracies are averages over all independent

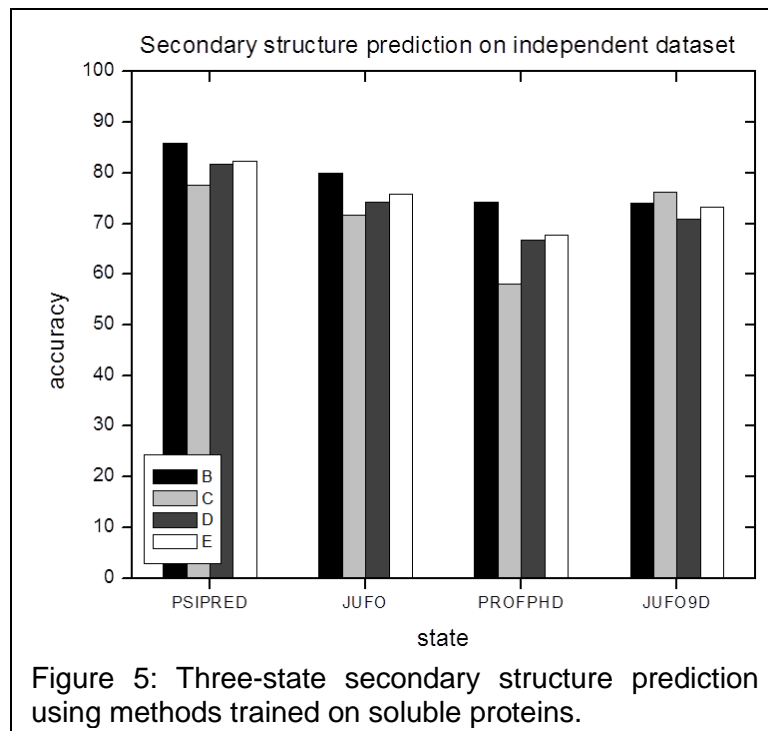
datasets. In the three state scenario, SS prediction identifies on average 73.2% of the residues correctly.

In comparison, PsiPred has highest accuracies (H = 85.91%, E = 77.57%, C = 81.69%, avg = 82.36%), ProfPhD has lowest accuracies (H = 74.26%, E = 58.12%, C = 66.80%, avg = 67.68%), and the earlier version of Jufo (H = 80.05%, E = 71.69%, C = 74.18%, avg = 75.84%) have intermediate accuracies. In general, accuracies in the



Trans-membrane spans are predicted at 94.8% in three states

The nine state occurrences can also be added together to represent three-state TM prediction. Here, 75.72% of the membrane core states are correctly identified, 74.16% of the states in the transition region, and 95.35% of the states in solution.



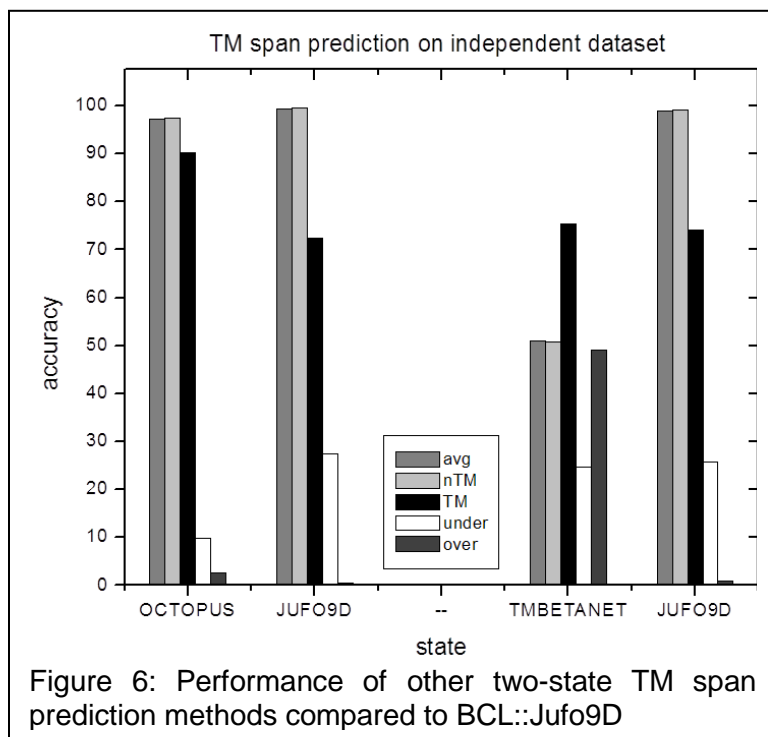
Overall, the environment of 94.8% of the residues in the independent datasets is correctly identified. This highlights that there are many more soluble states in the datasets than there are membrane or transition states. For training, the oversampling procedure guarantees that this bias does not impact the weights in the ANNs.

Two-state trans-membrane span identification yields accuracies of 99%

The nine states can also be summed to represent the two states to directly compare BCL::Jufo9D to other TM prediction methods (Figure 6). Octopus identifies whether a residue is located in a trans-membrane helix or not. For the TM α -helical bundles in our independent datasets, Octopus identifies on average 90.16% of the TM helix states, 97.39% of the “other” states, and correctly predicts the states of 97.34% of the residues. BCL::Jufo9D correctly predicts on average 72.51% of the TM helix states, 99.56% of the “other” states, and overall 99.35% of the residues.

For the TM β -barrels in our independent datasets, TMBeta-Net identifies 75.42% of the TM strand states, 50.84% of the “other” states, and correctly predicts the states of

50.92% of the residues. BCL::Jufo9D correctly predicts on average 74.17% of the TM strand states, 99.09% of the “other” states, and overall 99.00% of the residues.



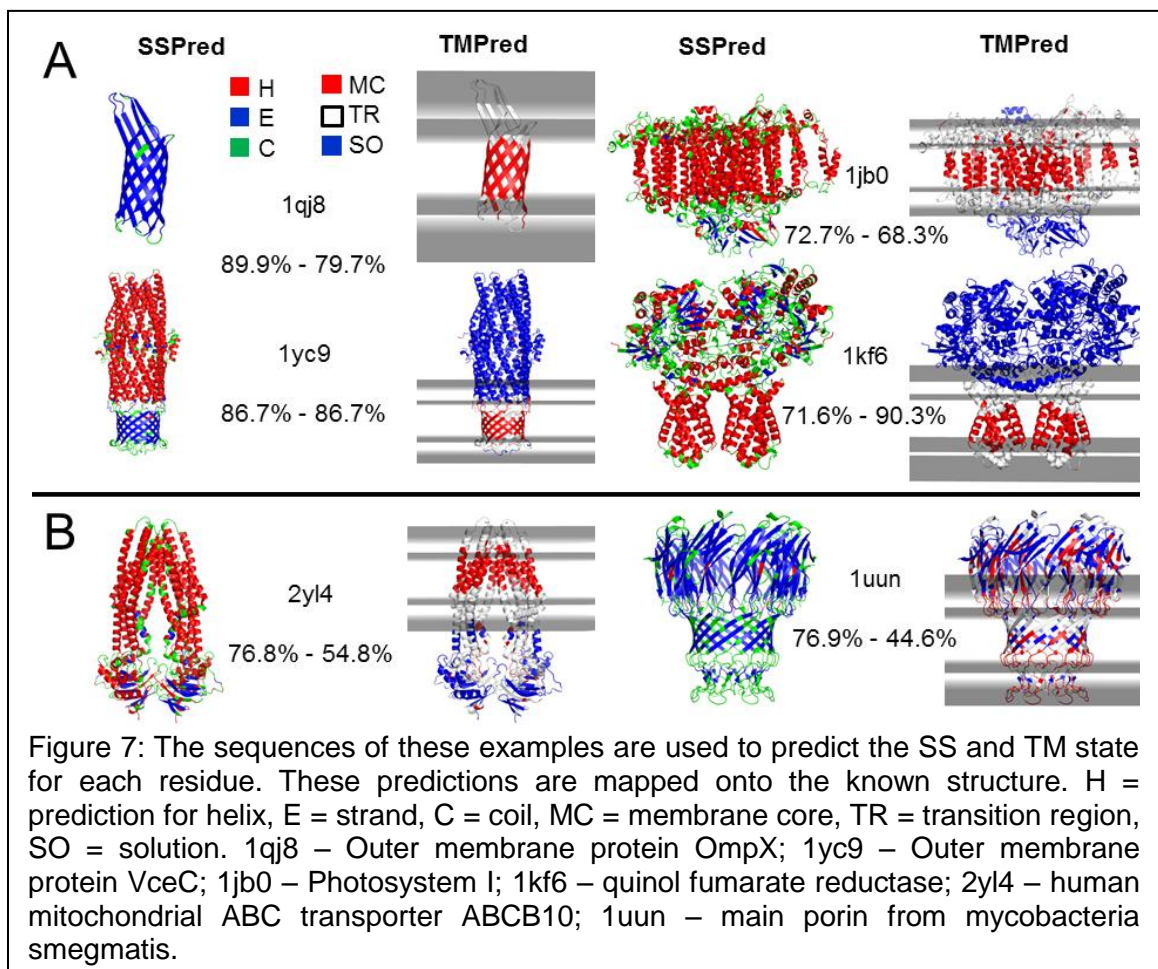
Over- and underpredictions

In addition to the two-state predictions Figure 6 also shows the over- and under-predictions of TM spans for complete datasets. Octopus over-predicts 2.61% of the residues while under-predicting 9.84% of the α -helical trans-membrane residues. In comparison, BCL::Jufo9D over-predicts 0.44% of the residues while under-predicting 27.49% of the α -helical trans-membrane residues. Similar trends are seen for residues in trans-membrane β -strands where BCL::Jufo9D over-predicts only 0.91% while under-predicting 25.83% of trans-membrane β -strand residues. TMBetaNet over-predicts 49.16% (!) of the residues while under-predicting 24.58% of trans-membrane β -strand residues.

Examples demonstrate high prediction accuracies

Figure 7 shows some example cases where the protein sequence was used to predict the SS and TM regions with BCL::Juf09D. These predictions were mapped onto the known protein structures.

The first example shows the outer membrane protein OmpX (PDB: 1qj8) where the SS is correctly identified for 89.9% of the residues and TM regions are correctly predicted for 79.7%. Other examples include the TolC receptor (PDB: 1yc9) with 86.7% in both SS and TM span prediction, the photosynthetic reaction center of cyanobacteria (PDB: 1jb0) with 72.7% in SS prediction and 68.3% in TM span prediction, and the *E.coli* quinol fumarate reductase (PDB: 1kf6) with 71.6% (SS) and 90.3% (TM). Panel B shows challenges where some of the residues are incorrectly identified. The first example is



human mitochondrial ABC transporter (PDB: 2yl4) with 76.8% of the residues correctly identified in terms of SS, and 54.8% for TM span prediction. For the main porin of *mycobacteria smegmatis* (PDB: 1uun) the SS is correctly predicted for 76.9% and the TM spans are correctly identified for 44.6% of the residues.

Discussion

Post-processing reduces noise in the predictions

To compute the final prediction vector from the output of several ANNs, the output of the four ANNs was summed for which a particular protein or dataset was (in) the independent dataset. It was tested whether providing the outputs of all ANNs over a window of 31 residues would further reduce the noise but it showed no significant improvement over the described method (data not shown).

BCL::Jufo9D achieves nine-state accuracies of 70.3%

The averages of the BCL::Jufo9D predictions over the independent datasets yield 70.3% correctly predicted residues. This is a considerable achievement given the fact that for a random prediction in nine states the accuracy would be 11.1%. Furthermore, only a decade ago SS prediction tools obtained three-state accuracies in this range where BCL::Jufo9D provides those accuracies in nine states.

From Figure 4 it can be seen that soluble states are very accurately identified because they are distinctly different from TM states or transition region states. In solution, the SS prediction, however, is not much higher than for TM states. It is easier for the ANN to inversely predict transition region or TM states. One of the reasons is a variety of hydrophobic thicknesses for MPs and this variety is not represented in our method for the following reason: even though PDBTM with its TMDet algorithm is able to identify the hydrophobic thicknesses of MPs, we wanted to circumvent an influence of

these predictions onto our method because an experimental validation is lacking. Furthermore, in early tests it was found that using the hydrophobic thickness from the PDBTM for the development of our algorithm does not substantially increase the prediction accuracies of BCL::Juf09D. Another reason for lower accuracies for the transition region is the fact that it is located between the membrane and the solution. Since varying hydrophobic thicknesses occur in MPs, the exact location of the transition region is more difficult to predict and accuracy is sacrificed in favor of soluble and TM states.

The prediction accuracies of coil states are lower than for helix or strand states, irrespective of their environment. This is expected, since the coil regions lack a defined structure with characteristic properties that enable the identification of a pattern for an accurate prediction.

Inverse predictions rarely happen between helix and strand states, irrespective of their environment, but inverse predictions are seen between helix/coil and strand/coil. This is expected because the properties characteristic for helices with a periodicity of 3.6 are distinctly different than for strands with a periodicity of 2.

The trends for inverse predictions between transition region/TM (but not solution) and helix/coil and strand/coil can be easier noticed by just considering the three-state SS prediction and three-state TM span identification as seen in Figure 4.

Three-state secondary structure is identified at 73.2%

On average, the SS is correctly identified for 73.2% of the residues. Similar accuracies are obtained for helix and strand states for each of the different environments, however the accuracies in the transition region are lower than for TM or solution for reasons already discussed. The TM span prediction identifies 94.8% of the residues correctly. These accurate predictions are obtained because each of the subsets

is largely biased towards soluble proteins which are very accurately identified. For training, however, this bias is eliminated through an oversampling procedure and therefore should not be over-emphasized. Fact is that BCL::Jufo9D recognizes soluble proteins very accurately.

Figure 5 shows the prediction accuracies for SS prediction. PsiPred is the gold-standard method for SS prediction for many years and our results support this fact once more. The developers of PsiPred excluded similar folds in their databases and even though we tested this for developing BCL::Jufo9D, the residue occurrences dramatically decreased for the TM and transition region states and therefore negatively impacted the prediction accuracies.

Two-state trans-membrane span identification and over-and under predictions

It was shown in Figure 6 that Octopus yields extremely high prediction accuracies. It identifies α -helical TM spans to a very high degree but also neither over- nor under-predicts residues substantially.

In comparison, BCL::Jufo9D over-predicts always less than 1% of the residues but under-predicts on average about 25% of the residues both in α -helical or β -strand TM spans. The under prediction is attributed to the existence of the transition region because most of these residues are predicted to be in transition states, as opposed to their actual membrane state. This also suggests that BCL::Jufo9D predicts TM spans too short rather than too long. As an improvement to the method the transition region could be defined with a thickness of less than 10 Å or using a higher membrane thickness. However, this might impact the accuracies in the membrane core region and in solution and different scenarios would have to be tested.

TMBeta-Net identifies only about 75% of the TM β -strand residues while also over-predicting almost half of the total residues in these datasets. This is a rather poor performance considering how specific to TM β -barrels this method is.

Examples demonstrate high prediction accuracies

The examples in Figure 7 demonstrate the high prediction accuracies of BCL::Jufo9D. Given that only sequence information is used to predict the SS and TM location it is astonishing to obtain such high accuracies. It can be seen that the membrane location is accurately predicted for β -barrels, irrespective of the size of the barrel or the number of strands. Even β -barrels where each of the subunits only provides two or four strands to the complete barrel (PDB: 1yc9) are accurately predicted. The examples show that some of the TM helices or strands are predicted too short because the transition region of 10 Å may be too thick.

Challenges and failures

Panel B in Figure 7 highlights challenges that need to be addressed. Whereas the TM region in 2yl4 is accurately identified, a number of residues in solution are predicted to be in the transition region or even the membrane. Interestingly, the SS prediction does not suffer from the wrong identification of membrane regions.

Another example is the oligomeric main porin of *mycobacteria smegmatis* (PDB: 1uun) where stretches of residues in the membrane are predicted to be soluble. In addition, a large number of residues in solution are identified as transition region or membrane states. The difficulty in such cases is that complete stretches of residues are incorrectly identified, hindering an exact classification of the protein or of identification of the number of membrane spanning regions. However, the SS prediction does not suffer from this incorrect identification. The reason for these failures is currently unknown and

to address them would require a considerable amount of effort. We tested the influence of barrel diameter, charged residues in the membrane, we trained ANNs to predict side-chain orientation in the membrane as well as solvent-accessible surface area. Unfortunately, none of these efforts were fruitful. It can be argued that since for these ANNs the databases of proteins contained a vast variety of MPs and soluble proteins in different sizes, shapes, and secondary structures, it could make it difficult for the ANNs to predict all residues to very high accuracies. The examples that were incorrectly identified were very rare and prediction methods will never provide 100% accuracy. Therefore, these examples may be outliers of a generally very accurate method.

Conclusions

We presented the first prediction tool that integrates the prediction of secondary structure with the identification of TM spans. An Artificial Neural Network was trained on a soluble protein and membrane protein database to output the combination of the three secondary structure states helix, strand, coil with the three environment states membrane core, transition region, solution in a nine-state probability vector for each residue in the sequence. It was shown that the per-residue accuracy in nine states is with 70.3% almost as high as some of the secondary structure prediction tools that predict three states. When combined into a three-state prediction, BCL::Jufo9D achieves accuracies for secondary structure prediction of 73.2% and TM span prediction of 94.8%. These results are comparable to current secondary structure and TM span prediction tools, however, BCL::Jufo9D integrates both at the same time.

References

- [1] K. Kazmier, N.S. Alexander, J. Meiler, H.S. McHaourab, Algorithm for selection of optimized EPR distance restraints for de novo protein structure determination, *J Struct Biol*, 173 (2011) 549-557.
- [2] M. Punta, L.R. Forrest, H. Bigelow, A. Kernytsky, J. Liu, B. Rost, Membrane protein prediction methods, *Methods*, 41 (2007) 460-474.
- [3] B. Rost, C. Sander, R. Schneider, Phd - an Automatic Mail Server for Protein Secondary Structure Prediction, *Computer Applications in the Biosciences*, 10 (1994) 53-60.
- [4] B. Rost, PHD: predicting one-dimensional protein structure by profile-based neural networks, *Methods Enzymol*, 266 (1996) 525-539.
- [5] K. Lin, V.A. Simossis, W.R. Taylor, J. Heringa, A simple and fast secondary structure prediction method using hidden neural networks, *Bioinformatics*, 21 (2005) 152-159.
- [6] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol*, 292 (1999) 195-202.
- [7] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, *Bioinformatics*, 16 (2000) 404-405.
- [8] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, *J Mol Biol*, 215 (1990) 403-410.
- [9] J. Meiler, D. Baker, Coupled prediction of protein secondary and tertiary structure, *Proc Natl Acad Sci U S A*, 100 (2003) 12105-12110.
- [10] J. Meiler, M. Muller, A. Zeidler, F. Schmaschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, *Journal of Molecular Modeling*, 7 (2001) 360-369.
- [11] L. Holm, C. Sander, Mapping the protein universe, *Science*, 273 (1996) 595-603.
- [12] B. Rost, C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks, *Proc Natl Acad Sci U S A*, 90 (1993) 7558-7562.
- [13] B. Rost, C. Sander, R. Schneider, PHD--an automatic mail server for protein secondary structure prediction, *Comput Appl Biosci*, 10 (1994) 53-60.
- [14] T.P. Hopp, K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, *Proc Natl Acad Sci U S A*, 78 (1981) 3824-3828.

- [15] D.M. Engelman, T.A. Steitz, A. Goldman, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu Rev Biophys Chem*, 15 (1986) 321-353.
- [16] W.C. Wimley, T.P. Creamer, S.H. White, Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides, *Biochemistry*, 35 (1996) 5109-5124.
- [17] S.H. White, W.C. Wimley, Membrane protein folding and stability: physical principles, *Annu Rev Biophys Biomol Struct*, 28 (1999) 319-365.
- [18] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S.H. White, G. von Heijne, Recognition of transmembrane helices by the endoplasmic reticulum translocon, *Nature*, 433 (2005) 377-381.
- [19] T. Hessa, N.M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, G. von Heijne, Molecular code for transmembrane-helix recognition by the Sec61 translocon, *Nature*, 450 (2007) 1026-U1022.
- [20] C.P. Moon, K.G. Fleming, Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers, *Proc Natl Acad Sci U S A*, 108 (2011) 10174-10177.
- [21] J. Janin, Surface and inside volumes in globular proteins, *Nature*, 277 (1979) 491-492.
- [22] M. Punta, A. Maritan, A knowledge-based scale for amino acid membrane propensity, *Proteins*, 50 (2003) 114-121.
- [23] T. Beuming, H. Weinstein, A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins, *Bioinformatics*, 20 (2004) 1822-1835.
- [24] A. Senes, D.C. Chadi, P.B. Law, R.F. Walters, V. Nanda, W.F. Degradó, E(z), a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices, *J Mol Biol*, 366 (2007) 436-448.
- [25] J. Koehler, N. Woetzel, R. Staritzbichler, C.R. Sanders, J. Meiler, A unified hydrophobicity scale for multispan membrane proteins, *Proteins*, (2008).
- [26] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J Mol Biol*, 157 (1982) 105-132.

- [27] D. Eisenberg, R.M. Weiss, T.C. Terwilliger, W. Wilcox, Hydrophobic Moments and Protein-Structure, Faraday Symposia of the Chemical Society, (1982) 109-120.
- [28] H.R. Guy, Amino acid side-chain partition energies and distribution of residues in soluble proteins, *Biophys J*, 47 (1985) 61-70.
- [29] H. Viklund, A. Elofsson, OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar, *Bioinformatics*, 24 (2008) 1662-1668.
- [30] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol*, 305 (2001) 567-580.
- [31] R.Y. Kahsay, G. Gao, L. Liao, An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes, *Bioinformatics*, 21 (2005) 1853-1858.
- [32] T. Nugent, D.T. Jones, Transmembrane protein topology prediction using support vector machines, *BMC Bioinformatics*, 10 (2009).
- [33] M. Arai, H. Mitsuke, M. Ikeda, J.X. Xia, T. Kikuchi, M. Satake, T. Shimizu, ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability, *Nucleic Acids Res*, 32 (2004) W390-393.
- [34] J.X. Xia, M. Ikeda, T. Shimizu, ConPred_elite: a highly reliable approach to transmembrane topology prediction, *Comput Biol Chem*, 28 (2004) 51-60.
- [35] M.M. Gromiha, S. Ahmad, M. Suwa, Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins, *J Comput Chem*, 25 (2004) 762-767.
- [36] H.R. Bigelow, D.S. Petrey, J. Liu, D. Przybylski, B. Rost, Predicting transmembrane beta-barrels in proteomes, *Nucleic Acids Res*, 32 (2004) 2566-2577.
- [37] H. Bigelow, B. Rost, PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins, *Nucleic Acids Res*, 34 (2006) W186-188.
- [38] B. Rost, J. Liu, The PredictProtein server, *Nucleic Acids Res*, 31 (2003) 3300-3304.
- [39] N.K. Singh, A. Goodman, P. Walter, V. Helms, S. Hayat, TMBHMM: a frequency profile based HMM for predicting the topology of transmembrane beta barrel

proteins and the exposure status of transmembrane residues, *Biochim Biophys Acta*, 1814 (2011) 664-670.

- [40] S. Montgomerie, J.A. Cruz, S. Shrivastava, D. Arndt, M. Berjanskii, D.S. Wishart, PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation, *Nucleic Acids Res*, 36 (2008) W202-209.
- [41] A.G. Garrow, A. Agnew, D.R. Westhead, TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins, *BMC Bioinformatics*, 6 (2005) 56.
- [42] J. Koehler, R. Mueller, J. Meiler, Improved prediction of trans-membrane spans in proteins using an Artificial Neural Network, *IEEE Comp. Intel. Bioinf. Comp. Biol.*, (2009) 68-74.
- [43] G.E. Tusnady, Z. Dosztanyi, I. Simon, Transmembrane proteins in the Protein Data Bank: identification and classification, *Bioinformatics*, 20 (2004) 2964-2972.
- [44] G.E. Tusnady, Z. Dosztanyi, I. Simon, PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank, *Nucleic Acids Res*, 33 (2005) D275-278.
- [45] G.L. Wang, R.L. Dunbrack, PISCES: a protein sequence culling server, *Bioinformatics*, 19 (2003) 1589-1591.
- [46] G. Wang, R.L. Dunbrack, Jr., PISCES: recent improvements to a PDB sequence culling server, *Nucleic Acids Res*, 33 (2005) W94-98.
- [47] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22 (1983) 2577-2637.
- [48] K. Ginalski, J. Pas, L.S. Wyrwicz, M. von Grotthuss, J.M. Bujnicki, L. Rychlewski, ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure, *Nucleic Acids Res*, 31 (2003) 3804-3807.

CONCLUSIONS

Membrane proteins are very important drug targets and the determination of membrane protein structures is lagging far behind the determination of soluble protein structures. The basis for this dissertation work is the development of methods that facilitate membrane protein structure determination both from an experimental as well as computational standpoint.

It has been shown numerous times on soluble proteins that paramagnetic restraints can be very useful to obtain long-range and orientational restraints to facilitate structure determination. Moreover, these restraints possess the ability to replace long-range NOEs where they are unavailable, for instance for proteins yielding poorly resolved or overlapped resonances in NMR spectra. Since this problem occurs very often for large α -helical proteins, such as membrane proteins, the goal of my dissertation work was to establish paramagnetic tagging on membrane proteins and to possibly use these restraints for structure calculations. The theory of paramagnetic NMR restraints is introduced in the first chapter, which has been reproduced from a review published in *Progress in NMR Spectroscopy*.

Early approaches describe the replacement of the metal ion in metal-binding proteins with a lanthanide ion to measure paramagnetic restraints (Lee & Sykes, 1980). Later work included the determination of local structure of the two trans-membrane spanning protein F_1F_0 ATP synthase using a PROXYL-label (Girvin & Fillingame, 1995). Despite of this work several decades ago, tagging membrane proteins with non-nitroxide paramagnetic tags has only scarcely been described the literature. In 2000, Ma & Opella attached an EF-hand to the N-terminus of the 81-residue single trans-membrane span

protein Vpu. Instead of calcium which is typically bound to the EF-hand, they provided a lanthanide ion that partially oriented the micelle-bound protein in the magnetic field and allowed the measurement of RDCs. In 2007, Kamen, Cahill & Girvin attached an EDTA-based tag onto the two trans-membrane span protein F_1F_0 ATP synthase via a disulfide linkage. The authors measured RDCs and PCSs but did not report the use of these restraints for structure calculations.

To push the measurements of paramagnetic restraints in conjunction with structure determination, we initially started tagging Diacylglycerolkinase (DAGK), a homotrimeric α -helical membrane protein with a total of nine trans-membrane spans. This 40kDa protein creates a 110 kDa complex with the detergent and is currently close to the limit of feasibility in terms of complex size for NMR purposes. Unfortunately, after analyzing the NMR data closely, it was noticed that the highest quality restraints originated from residues at the flexible N-terminus and were therefore unsuitable for structure calculations. Additionally, the large PRE effects, i.e. line-broadening, prohibited the collection of a large number of restraints. These large PRE effects originated from DAGK being a homotrimer that therefore contained three paramagnetic metal ions contributing to the line-broadening. If a protein contains three metal ions, all three of them contribute to the restraints creating a mathematically complex problem if these restraints were to be interpreted. To circumvent this, an asymmetric tagging strategy was proposed, where only one of the three subunits would contain a paramagnetic metal ion, whereas the other two subunits would be untagged. This approach, even though interesting in thought, is definitely a challenging one to carry out practically.

Subsequently, we decided to use a different model system, namely the single trans-membrane span protein KCNE3. The second chapter is dedicated to the experimental methods for sample preparation, verification experiments, and NMR spectroscopy on KCNE3. It has been shown that paramagnetic tagging of KCNE3 is

feasible and yields RDCs, PREs, and PCSs. However, establishing the protocol requires considerable effort since the detergent in the sample interacts with all the components present and it is difficult to predict the behavior of the protein, the small molecule tags, and the lanthanides on a molecular level and in conjunction with the media used for sample preparation. Even though KCNE3 is certainly a “real-world” test case for paramagnetic tagging in terms of sample preparation, the size of KCNE3 with its 12 kDa is not in the ballpark of a challenging test case for solution-state NMR. Also, KCNE3 has a single trans-membrane span where the full potential of the paramagnetic restraints with their long-range information is hardly recognized. A more optimal vehicle to recognize the power of paramagnetic tagging would be a multi-span α -helical membrane protein, such as DAGK, where the fold of the protein could be elucidated solely by paramagnetic restraints.

Gaining structural information about KCNE3 is of high importance to elucidate the mechanism by which KCNE3 modulates the potassium channel KCNQ1. Mutations in both the channel as well the modulatory KCNE family members lead to a variety of diseases, among them Long QT syndrome leading to atrial fibrillation, and possibly sudden death. Since different KCNE family members modulate KCNQ1 in different manners (as outlined in Chapter 2) high-resolution structures of KCNE family members would allow docking of these structures into the homology model of KCNQ1. These docking models could be used to generate testable hypothesis of the mechanism of channel function and differing modulation by KCNE family members. This, in turn, would allow for the development of small molecule drugs that alter this interaction to mitigate disease symptoms.

It is suspected that tagging MPs with small molecule tags will be carried out more often and more efficiently in the future. Lanthanide-binding tags are constantly being optimized as shown by numerous publications from the Griesinger group, amongst

others. Currently, efficient tagging of MPs is hampered by difficulties in sample preparation which has to be optimized for each protein individually since a single working protocol that allows tagging of various MPs does not yet exist. Once this hurdle is overcome, lanthanide tagging of MPs will prove efficient to obtain paramagnetic restraints for structure calculations, since separate sample preparations for the measurement of RDCs and PREs will be avoided. Combining this sparse NMR data with adequate computational tools optimized for the use of sparse restraints, such as BCL::Fold developed in the Meiler group, will facilitate MP structure determination and substantially decrease the time required for this effort.

Chapter 3 describes the derivation of a knowledge-based potential that statistically describes the energetics of amino acids in environments of different hydrophobicity. Very polar amino acids, such as aspartate, glutamate, lysine, and arginine are rarely found in very hydrophobic environments such as the membrane bilayer. The number of amino acid occurrences in the ProteinDataBank was converted into a transfer free energy in three regions: membrane bilayer, transition region, and solution. This scale (termed the *Unified Hydrophobicity Scale* - UHS) was the first hydrophobicity scale derived from both α -helical proteins as well as β -barrels and complements the many hydrophobicity scales available that are derived in different manners and are therefore optimized for different uses (see Chapter 3). Since various folds, even ones with aqueous interior in the membrane, were used for derivation of the scale, the UHS likely under-estimates the penalty required to transfer polar or charged residues into the membrane when they are in contact with the lipid bilayer. However, from a computational point of view, such a statistical approach is useful for trans-membrane span identification if assumptions about secondary structure or solvent accessibility are disregarded.

The UHS was used to create a prediction tool that is able to predict trans-membrane spans from a protein sequence. The novelty of this approach described in Chapter 4 was that both α -helical trans-membrane spans, as well as β -strands could be predicted being superior to existing methods that are specialized to *either* α -helical proteins *or* β -barrels. Artificial Neural Networks were used as the underlying method because they are well suited to recognize patterns in the residue characteristics of the protein sequence.

The concept of using Artificial Neural Networks for a sequence-based prediction of trans-membrane spans was advanced in Chapter 5 to combine it with secondary structure prediction into a prediction tool that can simultaneously predict both. The hypothesis for developing this prediction method was the notion that the hydrophobic environment is a key influence on the formation of secondary structure: when a polypeptide chain is transferred from solution into the membrane, it tries to saturate its backbone carbonyl and amino groups by the formation of hydrogen bonds and therefore secondary structure. On the contrary, in a hydrophilic environment the backbone functional groups can undergo hydrogen bonds with the surrounding water.

The prediction method BCL::Jufo9D achieves accuracies that are comparable or higher than for the best prediction tools available. However, BCL::Jufo9D does not always outperform highest quality secondary structure or trans-membrane span predictors. For some rare examples complete stretches of residues are incorrectly identified. These errors occur mostly for TM span prediction where the secondary structure prediction generally is still correct. It could be argued that it becomes difficult for a prediction method to predict a very broad range of features that include different secondary structure types in different environments. Usually, prediction methods achieve very high accuracies because they are specialized to recognize a certain pattern along the sequence, for instance for trans-membrane β -strands. Since a broad range of

features in a database is difficult to generalize and capture in detail, it is not surprising that BCL::Jufo9D is not substantially better than other methods to predict these features. However, the strength of BCL::Jufo9D is the ability to predict different secondary structure types in different environments simultaneously which does not require creating a consensus prediction of several prediction tool outputs that possibly even contradict each other.

Proteus2 developed in the Wishart lab is a predictor that combines several high-accuracy predictors in a single prediction tool: it identifies β -barrel proteins by a binary predictor that is not specialized on predicting particular regions of β -strands but rather whether β -strands are contained in the sequence, and subsequently, a high-quality β -barrel predictor is used on the sequences that were predicted to contain β -strands. α -helical proteins are subjected to a different predictor specialized on prediction for α -helical proteins. Even though both Proteus2 and BCL::Jufo9D have the ability to predict different types of secondary structure in different environments, the advantage of BCL::Jufo9D is that it should yield higher accuracies in identifying secondary structures in mixed α/β proteins that are neither pure α -helical bundles nor pure β -barrels and that Proteus' sub-servers are specialized for. Unfortunately, Proteus2 could not be tested on our databases due to technical difficulties.

The development of BCL::Jufo9D is hoped to have impact on the experimental community. During its nine years of existence the previous version of Jufo, developed in 2003, was used almost 100,000 times by experimentalists all over the world. BCL::Jufo9D has an improved setup with additional output information. The output is also converted into its original Jufo output format which is additionally available on the webserver, making the transition to the new BCL::Jufo9D seamless for experimentalists.

Compared to other trans-membrane span prediction methods BCL::Jufo9D does not assume a fixed length of TM spans (as many other methods do) and is also able to

predict short secondary structure elements. This results in higher resolution in the predictions containing more information. At last, the setup of BCL::Jufo9D bears the potential to predict conformational switches. We are not claiming that BCL::Jufo9D is able to predict conformational switches, since it is not specifically designed to do so. However, an identical setup can be used to train a tool for sequence-based conformational switch prediction, a method that is much sought-after since efforts to do so have been unsuccessful to date.

APPENDIX TO CHAPTER 1

A1. Definition of tensors

In the literature many different interdependent tensors are defined. These tensors will be briefly explained,

Assumed is a coordinate system that is fixed to the molecule. In this coordinate system the orientation of the external magnetic field can be described by a probability tensor \mathbf{P} which is a real, symmetric tensor

$$\mathbf{P} = \begin{pmatrix} P_{xx} & P_{xy} & P_{xz} \\ P_{yx} & P_{yy} & P_{yz} \\ P_{zx} & P_{zy} & P_{zz} \end{pmatrix} \quad (\text{A1})$$

with a trace of 1:

$$P_{xx} + P_{yy} + P_{zz} = 1. \quad (\text{A2})$$

Its principal values describe the probability of the external magnetic field pointing along its principal axes (x, y, z) . Under isotropic re-orientation the principal components of this tensor will all be equally $\frac{1}{3}$. As a symmetric tensor, it can be described by five independent values. The probability tensor is not used in the literature but is introduced here to give the derived tensors physical meaning. All related tensors \mathbf{T} (Sections A1.(2) to A1.(5) below) can be decomposed into an isotropic and an anisotropic tensor

$$\mathbf{T} = \mathbf{T}^{iso} + \mathbf{t}^{aniso} \quad (\text{A3})$$

corresponding to

$$\begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{xy} & T_{yy} & T_{yz} \\ T_{xz} & T_{yz} & T_{zz} \end{pmatrix} = \bar{T}^{iso} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} -t_{yy}-t_{zz} & t_{xy} & t_{xz} \\ t_{xy} & t_{yy} & t_{yz} \\ t_{xz} & t_{yz} & t_{zz} \end{pmatrix}. \quad (\text{A4})$$

The isotropic tensor \mathbf{T}^{iso} has the same trace as the overall tensor \mathbf{T}

$$\bar{T}^{iso} = \frac{1}{3} \text{Tr}(\mathbf{T}) = \frac{1}{3} (T_{xx} + T_{yy} + T_{zz}) \quad (\text{A5})$$

which means that the anisotropic tensor \mathbf{t}^{aniso} is traceless (i.e. has a trace of zero). In our work the isotropic tensors are not considered unless otherwise noted, as only the anisotropic part contributes to molecular alignment and the resulting effects.

- (1) The magnetic susceptibility tensor χ is a real, symmetric, and traceless tensor that is described above (Eq.1 and Eq.2).
- (2) The probability tensor can be decomposed into an isotropic and an anisotropic part where the alignment tensor \mathbf{A} represents the anisotropic part of the probability tensor:

$$\mathbf{A} = \mathbf{P} - \frac{1}{3} \mathbf{I} \quad (\text{A6})$$

$$aniso = tot - iso$$

| | | | |
|-------|---|---|---|
| Trace | 0 | 1 | 1 |
|-------|---|---|---|

where I is the identity matrix and the alignment tensor \mathbf{A} (sometimes also denoted as \mathbf{D}) is a real, symmetric tensor. The alignment tensor has the same orientation as the probability tensor. If the molecular alignment originates in magnetic susceptibility anisotropy the alignment tensor is related to the magnetic susceptibility tensor (see above) by [1]

$$\mathbf{A} = \frac{B_0^2}{15\mu_0 kT} \chi. \quad (\text{A7})$$

This shows that the degree of alignment increases with the magnetic field strength [1] and with the magnetic susceptibility. The notation \mathbf{A} of the alignment tensor should not be confused with the hyperfine coupling constant A .

- (3) The Saupe order matrix \mathbf{S} is a real, symmetric and traceless tensor that can be calculated from the alignment tensor by $\mathbf{S} = \frac{3}{2}\mathbf{A}$. For calculations mostly either the alignment tensor, the susceptibility tensor or the Saupe order tensor are used. It has to be noted, that the letter S in the theory of paramagnetic NMR can have three different meanings: it describes the Saupe order tensor, the spin quantum number, and the order parameter. In this review, \mathbf{S} will denote the Saupe order matrix, and S will denote the spin quantum number unless noted otherwise.
- (4) The g-factor (the electron g-factor or the Landé-g-factor) is a dimensionless proportionality constant relating the magnetic moment of a particle to its quantum numbers. It can be calculated from the spin quantum number S , the angular momentum quantum number L and the total angular momentum quantum number J by

$$g_J = 1 + \frac{J(J+1) - L(L+1) + S(S+1)}{2J(J+1)}. \quad (\text{A8})$$

The g-tensor results when the g-factor is orientation dependent. It can be related to the elements of the susceptibility tensor [2] by

$$g_{aa}^2 = \frac{3kT}{\mu_0 \mu_B^2 S(S+1)} \chi_{aa}. \quad (\text{A9})$$

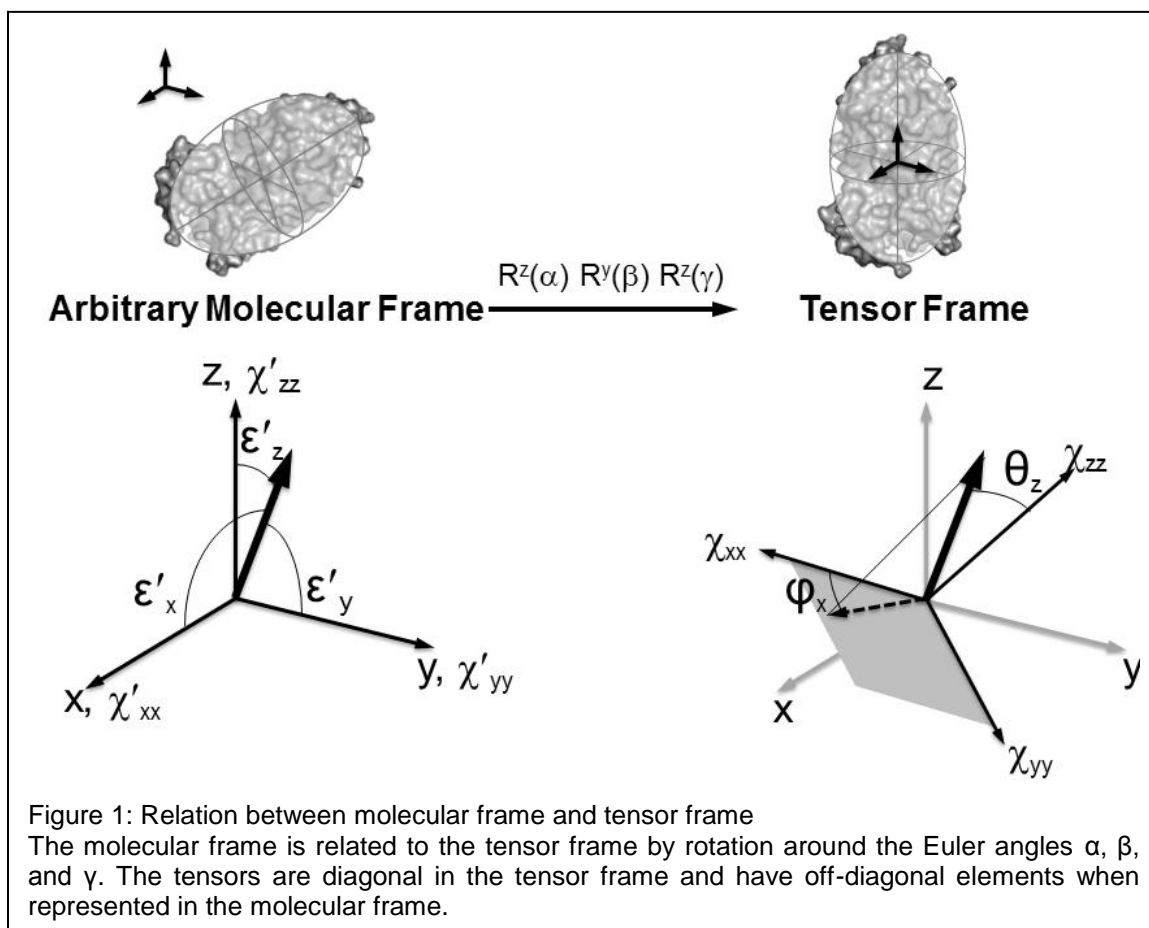
The g-tensor is a real, symmetric, traceless tensor. For spin-1/2 nuclei g-values for various metals can be measured by EPR spectroscopy and the tensor values can be obtained by single-crystal EPR measurements [2].

The principal axes of the tensor are defined such that $|\chi_{zz}| > |\chi_{yy}| > |\chi_{xx}|$. When the protein alignment originates solely in the MSA, these tensors have the same orientation and are therefore diagonal in the same frame [3] (see below). However, this is not generally the case [3]. Since for the current review the tensors will have identical orientation, the terms alignment tensor, susceptibility tensor, and Saupe order matrix will be used interchangeably.

A2. Definition of coordinate frames

There are three different coordinate frames (Fig.A1): (a) the lab frame in which the magnetic field is considered to be aligned with the z-coordinate; (b) the molecular frame that is fixed to the molecule. It can be arbitrarily defined, for instance depending on the shape of the molecule or as the frame of the protein in the ProteinDataBank file; (c) tensor frame which defines the principal axes of the magnetic susceptibility tensor

associated with the unpaired electron. In the current review the mathematical descriptions will be restricted to the tensor frame with the variables (χ, θ, φ) and the molecular frame with the variables $(\chi', \theta', \varphi')$.



The orientation of the molecular frame or the tensor frame with respect to the lab frame is usually unknown at the beginning of a study and is determined during the calculations. Even if partial alignment is imposed, there is still residual tumbling that makes it impossible to determine the rotation angles between these coordinate frames. Under the assumption that there is no or negligible internal mobility the orientation of the molecular frame with respect to the tensor frame is often assumed to be fixed. Therefore each internuclear vector has a fixed orientation with respect to the tensor frame. The tensor frame depends on the shape and the charge distribution within the molecule.

In the molecular frame the tensors can be described by five unknown parameters due to the symmetry property and the trace of the matrices:

molecular frame:

$$\chi' = \begin{pmatrix} \chi'_{xx} & \chi'_{xy} & \chi'_{xz} \\ \chi'_{yx} & \chi'_{yy} & \chi'_{yz} \\ \chi'_{zx} & \chi'_{zy} & \chi'_{zz} \end{pmatrix} = \begin{pmatrix} -\chi'_{yy} - \chi'_{zz} & \chi'_{xy} & \chi'_{xz} \\ \chi'_{xy} & \chi'_{yy} & \chi'_{yz} \\ \chi'_{xz} & \chi'_{yz} & \chi'_{zz} \end{pmatrix} \quad (\text{A10})$$

When this matrix is rotated into the tensor frame, it adopts a diagonal form and all off-diagonal elements are equal to zero:

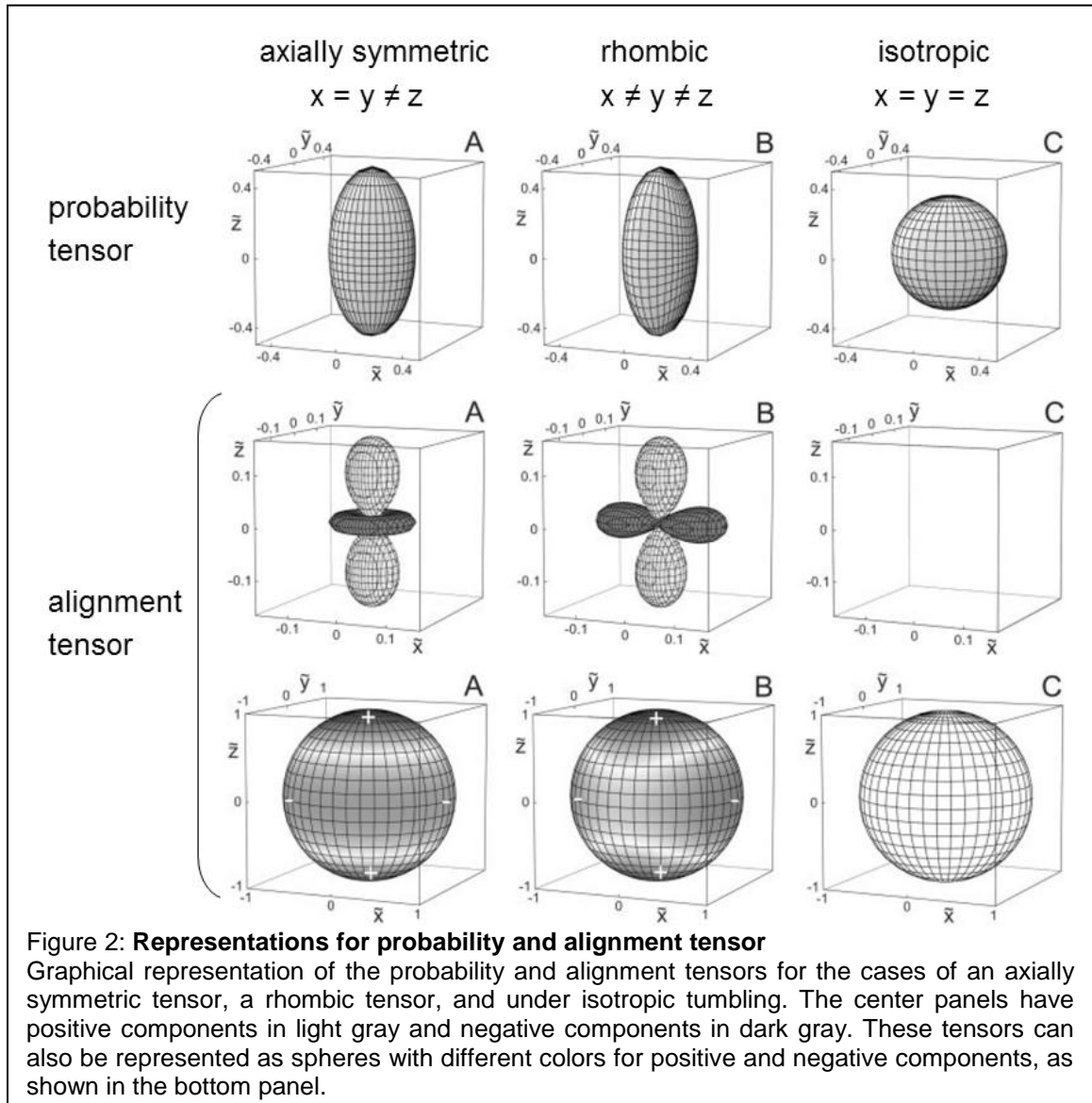
tensor frame:

$$\chi = [R^z(\alpha)R^y(\beta)R^z(\gamma)]^T \times \chi' \times [R^z(\alpha)R^y(\beta)R^z(\gamma)]$$

$$\chi = \begin{pmatrix} -\chi_{yy} - \chi_{zz} & 0 & 0 \\ 0 & \chi_{yy} & 0 \\ 0 & 0 & \chi_{zz} \end{pmatrix} \quad (\text{A11})$$

It should be noted that the trace of a matrix is invariant under rotation which means that it is independent of the coordinate frame. The number of unknowns remains five since the rotation angles α , β , and γ are unknown. The eigenvalues of the diagonal matrix are the principal components of the tensor. They can also be described by the axial and rhombic components of the tensor (Eq.4).

Tensors can be depicted by ellipsoids or shapes that look like atomic orbitals (Fig.A2). Since the probability tensor contains only positive elements in its diagonal, it can be illustrated by cigar shaped ellipsoids or a sphere, depending on the rhombicity and axially. A rhombic tensor is the most general case having different components in the x, y, and z direction. The rhombicity of a tensor describes how much the x and y-



components deviate from each other (Eq.4a). An axially symmetric tensor is symmetric around the z-axis. It has identical elements in the x and y dimensions so that these elements can be described as parallel and perpendicular components. The axially of a

tensor describes how much the z-component deviates from the average of the x and y-components (Eq.4b). For an axially symmetric probability tensor the following equations hold:

$$\mathbf{P} = \begin{pmatrix} P_{\perp} & 0 & 0 \\ 0 & P_{\perp} & 0 \\ 0 & 0 & P_{\parallel} \end{pmatrix} \text{ with its trace } P_{\parallel} + 2P_{\perp} = 1 \quad (\text{A12})$$

so that the axial and rhombic components can be described as

$$P_{rh} = P_{xx} - P_{yy} = 0$$

$$P_{ax} = P_{zz} - \frac{P_{xx} + P_{yy}}{2} = P_{\parallel} - P_{\perp}. \quad (\text{A13})$$

For an axially symmetric alignment tensor

$$\mathbf{A} = \begin{pmatrix} A_{\perp} - \frac{1}{3} & 0 & 0 \\ 0 & A_{\perp} - \frac{1}{3} & 0 \\ 0 & 0 & A_{\parallel} - \frac{1}{3} \end{pmatrix} \text{ with its trace } A_{\parallel} + 2A_{\perp} - 1 = 0 \quad (\text{A14})$$

such that the axial and rhombic components can be calculated the same way as for the probability tensor (Eq.4).

Traceless tensors, such as the alignment tensor, have negative elements in their diagonal and can be described either by a sphere with differently colored regions (for positive and negative contributions) or by the orbital-like shapes. The shapes results if

surfaces of constant PCS or RDC values (isosurfaces) are plotted. Examples for the graphical representations are shown in Fig.A2.

The difficulty in using RDCs and PCSs for protein structure elucidation is that the orientation of the molecular frame with respect to the tensor frame (which is defined by the three Euler rotation angles α , β , and γ) as well as the Saupe order tensor (which is defined by two independent variables in the tensor frame) are not known a priori and have to be determined in an iterative fashion as described above.

A3. Determination of the correlation times

The use of all relaxation equations requires the knowledge of the correlation times. The overall correlation time can generally be calculated by

$$\frac{1}{\tau_C} = \frac{1}{\tau_e} + \frac{1}{\tau_r} + \frac{1}{\tau_M} \quad (\text{A15})$$

however, not all correlation terms influence all relaxation terms equally. In these sections, the overall correlation time is defined in a separate equation. The electron spin correlation time does not apply to the Curie mechanism because there it is already averaged over all the electron density. The Curie relaxation is therefore only modulated by the rotation of the molecule [2]. The ranges for the correlation times are $10^{-13} - 10^{-7}$ s for the electron spin correlation time, $10^{-11} - 10^{-6}$ s for the rotational correlation time, and $10^{-10} -$ several seconds or minutes for the exchange correlation time [2]. For simplicity exchange relaxation will be neglected.

The total correlation time is determined by the shortest of the correlation times. For spin-labeled proteins the lower limit for the total correlation time is ~ 10 ns [4]. τ_C can be calculated from T_1 and T_2 measurements using [5]

$$\tau_c = \left(\frac{6 \left(\frac{\Delta R_2}{\Delta R_1} \right) - 7}{4\omega_H^2} \right)^{1/2} \quad (\text{A16})$$

and Eq.A20 and Eq.A21. Even though τ_c can vary about an order of magnitude the error in the distance remains small due to the $1/r^6$ dependence [4].

The rotational correlation time can be estimated from the model-free analysis [6], light-scattering experiments [7] or by measuring T_1 of the diamagnetic molecule at different magnetic fields [8]:

$$\tau_r^2 = \frac{T_1(B_1) - T_1(B_2)}{T_1(B_2)\omega(B_2)^2 - T_1(B_1)\omega(B_1)^2}. \quad (\text{A17})$$

τ_r can also be estimated using the Stokes-Einstein relationship [2]

$$\tau_r = \frac{4\pi\eta r_{eff}^3}{3kT} = \frac{\eta M}{\rho N_A kT} \quad (\text{A18})$$

with the viscosity of the solvent η (kg/sm), the effective radius of the molecule r_{eff} , the molecular weight M (kg/mol = 1 kDa), and the density of the molecule ρ (typically taken as 10^3 kg/m³). For elliptical molecules with the same volume the relaxation rates can be an order of magnitude larger than for spherical molecules [9].

The electron spin correlation times depend on the atomic number and the occupancy of the atomic orbitals [10]. They can be determined by NMR dispersion measurements [11] as was done for lanthanide aqua-complexes [12]. Short electron spin

correlation times are due to low-lying excited energy levels [2] with Orbach or Raman relaxation mechanisms [13]. $S=1/2$ ions like Cu^{2+} have excited states far above the ground state, therefore the electron spin correlation time is long [2].

A4. Additional notes on lifting the angular degeneracy in RDCs

Since the alignment tensor depends on the alignment medium for each alignment medium there are five unknowns (the tensor elements). If one of the tensor frames is considered as an anchor frame and the other tensor frames are expressed with respect to that frame, the number of variables is $5n - 3$ with n being the number of alignment media. Therefore twelve parameters are needed to describe the tensors in three alignment media [14].

Using this approach of describing the tensors as relative order tensors the system of equations is overdetermined and a solution exists as long as the number of datapoints $nk \geq 5n - 3 + 2k$. Here k is the number of internuclear vectors. The factor of $2k$ arises because there are two degrees of freedom to describe the orientation of an internuclear unit vector in the tensor frame [14].

Al-Hashimi et al. presented an order tensor analysis that completely removes the degeneracy using only two independent alignment media [15]. In this approach the protein is arbitrarily cut into two fragments and the order tensors of these fragments are separately determined using the RDCs. There are four possibilities to orient these tensors with respect to one another. When the tensor frame of one of the alignment media is taken as a reference frame the existence of the second alignment medium can lift this degeneracy when the tensors of the two fragments are superimposed. This works only if the alignment is external and if the alignment tensors of the two fragments are identical. In the case of motion that condition might not hold. In the case of internal

alignment this approach is not valid because the alignment is due to anisotropic magnetic susceptibility. This depends on the shape of the molecule and the charge distribution and is therefore not identical for both fragments. Hence, the order tensors are different and cannot be superimposed.

A5. Hyperfine shifts for lanthanides

For the lanthanides the hyperfine shift of the donor site is predominantly contact in origin which does not vary much along the lanthanide series [16]. The second half of the lanthanide series has the largest PCS/contact shift ratio but also exhibits larger line-broadening [17]. The ratio PCS/contact shift follows the pattern Yb > Tm > Dy > Tb > Er > Ho > Nd > Eu [18].

If, in the case of axial symmetry, the ratios of the shifts of different nuclei to a specified nucleus are independent of the lanthanide, then the shifts are PCSs and not contact shifts [19]. To separate contact and PCS most methods require that the lanthanide-ligand complex possesses axial symmetry which is not true for Ca-binding proteins [17]. In addition to the axial symmetry it is usually assumed that the hyperfine coupling constant A is constant for different ions, that the complexes are isostructural and that the crystal field parameters are independent of the paramagnetic ion [20].

A6. PREs: Alternative ways used to extract distance restraints

As discussed above, PREs can be converted into distance restraints. The spectral peak intensities decay exponentially depending on the evolution time:

$$\frac{V(t)}{V(0)} = \exp\left(-\frac{t}{T_1}\right). \quad (\text{A19})$$

The longitudinal PREs can be measured as

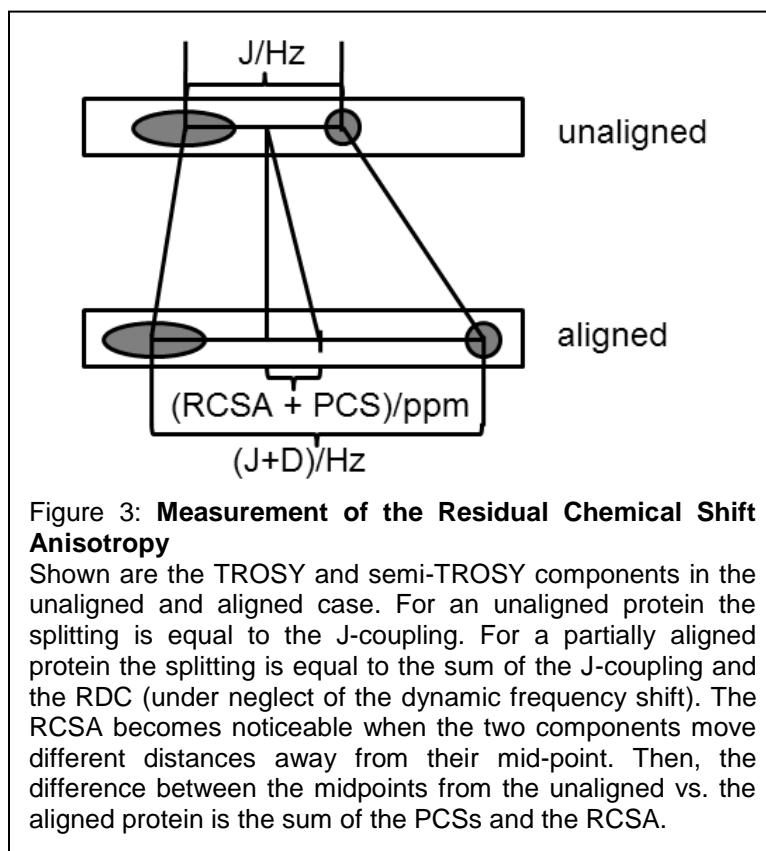
$$\Delta R_1 = \left(\frac{1}{T_1^{para}} \right) - \left(\frac{1}{T_1^{dia}} \right) \quad (A20)$$

where the T_1 can be determined by inversion-recovery experiments. Transverse PREs can be measured as the ratio of peak intensities or peak volumes or as the differences in linewidths [21]:

$$\Delta R_2 = \left(\frac{1}{T_1^{para}} \right) - \left(\frac{1}{T_1^{dia}} \right) = \frac{1}{t} \ln \left(\frac{I^{dia}}{I^{para}} \right) = \frac{1}{t} \ln \left(\frac{V^{dia}}{V^{para}} \right) \quad (A21a)$$

$$\Delta R_2 = \left(\frac{1}{T_1^{para}} \right) - \left(\frac{1}{T_1^{dia}} \right) = \pi(lw^{dia} - lw^{para}). \quad (A21b)$$

A7. Measurement of the Residual Chemical Shift Anisotropy



References

- [1] G. Otting, Prospects for lanthanides in structural biology by NMR, *J Biomol NMR*, 42 (2008) 1-9.
- [2] I. Bertini, C. Luchinat, S. Aime, NMR of paramagnetic substances, *Coordination Chemistry Reviews*, 150 (1996) R7-&.
- [3] I. Bertini, C. Luchinat, G. Parigi, Hyperfine shifts in low-spin iron(III) hemes: A ligand field analysis, *Eur J Inorg Chem*, (2000) 2473-2480.
- [4] P.A. Kosen, Spin Labeling of Proteins, *Methods in Enzymology*, 177 (1989) 86-121.
- [5] N.U. Jain, A. Venot, K. Umamoto, H. Leffler, J.H. Prestegard, Distance mapping of protein-binding sites using spin-labeled oligosaccharide ligands, *Protein Sci*, 10 (2001) 2393-2400.
- [6] I. Bertini, G. Cavallaro, M. Cosenza, R. Kummerle, C. Luchinat, M. Piccioli, L. Poggi, Cross correlation rates between Curie spin and dipole-dipole relaxation in

paramagnetic proteins: the case of cerium substituted calbindin D9k, *J Biomol NMR*, 23 (2002) 115-125.

- [7] L. Lee, B.D. Sykes, Nuclear magnetic resonance determination of metal-proton distances in the EF site of carp parvalbumin using the susceptibility contribution to the line broadening of lanthanide-shifted resonances, *Biochemistry*, 19 (1980) 3208-3214.
- [8] V. Gaponenko, J.W. Howarth, L. Columbus, G. Gasmi-Seabrook, J. Yuan, W.L. Hubbell, P.R. Rosevear, Protein global fold determination using site-directed spin and isotope labeling, *Protein Sci*, 9 (2000) 302-309.
- [9] D.E. Woessner, Nuclear spin relaxation in ellipsoids undergoing rotational Brownian motion, *Journal of Chemical Physics*, 37 (1962) 647-654.
- [10] I. Bertini, C. Luchinat, G. Parigi, R. Pierattelli, NMR spectroscopy of paramagnetic metalloproteins, *Chembiochem*, 6 (2005) 1536-1549.
- [11] I. Bertini, L. Banci, C. Luchinat, Proton magnetic resonance of paramagnetic metalloproteins, *Methods Enzymol*, 177 (1989) 246-263.
- [12] I. Bertini, F. Capozzi, C. Luchinat, G. Nicastro, Z.C. Xia, Water Proton Relaxation for Some Lanthanide Aqua Ions in Solution, *Journal of Physical Chemistry*, 97 (1993) 6351-6354.
- [13] I.I. Bertini, O. Galas, C. Luchinat, G. Parigi, G. Spina, Nuclear and Electron Relaxation in Magnetic Exchange Coupled Dimers: Implications for NMR Spectroscopy, *J Magn Reson*, 130 (1998) 33-44.
- [14] X. Miao, R. Mukhopadhyay, H. Valafar, Estimation of relative order tensors, and reconstruction of vectors in space using unassigned RDC data and its application, *J Magn Reson*, 194 (2008) 202-211.
- [15] H.M. Al-Hashimi, H. Valafar, M. Terrell, E.R. Zartler, M.K. Eidsness, J.H. Prestegard, Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings, *J Magn Reson*, 143 (2000) 402-406.
- [16] R.M. Golding, M.P. Halton, Theoretical Study of N-14 and O-17 Nmr Shifts in Lanthanide Complexes, *Australian Journal of Chemistry*, 25 (1972) 2577-2581.
- [17] J.G. Shelling, M.E. Bjornson, R.S. Hodges, A.K. Taneja, B.D. Sykes, Contact and Dipolar Contributions to Lanthanide-Induced Nmr Shifts of Amino-Acid and Peptide Models for Calcium-Binding Sites in Proteins, *Journal of Magnetic Resonance*, 57 (1984) 99-114.

- [18] J. Reuben, Origin of Chemical-Shifts in Lanthanide Complexes and Some Implications Thereof, *Journal of Magnetic Resonance*, 11 (1973) 103-104.
- [19] C.D. Barry, A.C. North, J.A. Glasel, R.J. Williams, A.V. Xavier, Quantitative determination of mononucleotide conformations in solution using lanthanide ion shift and broadenine NMR probes, *Nature*, 232 (1971) 236-245.
- [20] M.D. Kemple, B.D. Ray, K.B. Lipkowitz, F.G. Prendergast, B.D.N. Rao, The Use of Lanthanides for Solution Structure Determination of Biomolecules by Nmr - Evaluation of the Methodology with Edta Derivatives as Model Systems, *Journal of the American Chemical Society*, 110 (1988) 8275-8287.
- [21] P.E. Johnson, E. Brun, L.F. MacKenzie, S.G. Withers, L.P. McIntosh, The cellulose-binding domains from *Cellulomonas fimi* beta-1, 4-glucanase CenC bind nitroxide spin-labeled cellooligosaccharides in multiple orientations, *J Mol Biol*, 287 (1999) 609-625.

APPENDIX TO CHAPTER 2

Lanthanide tagging experiments on DAGK

Tagging protocol on DAGK

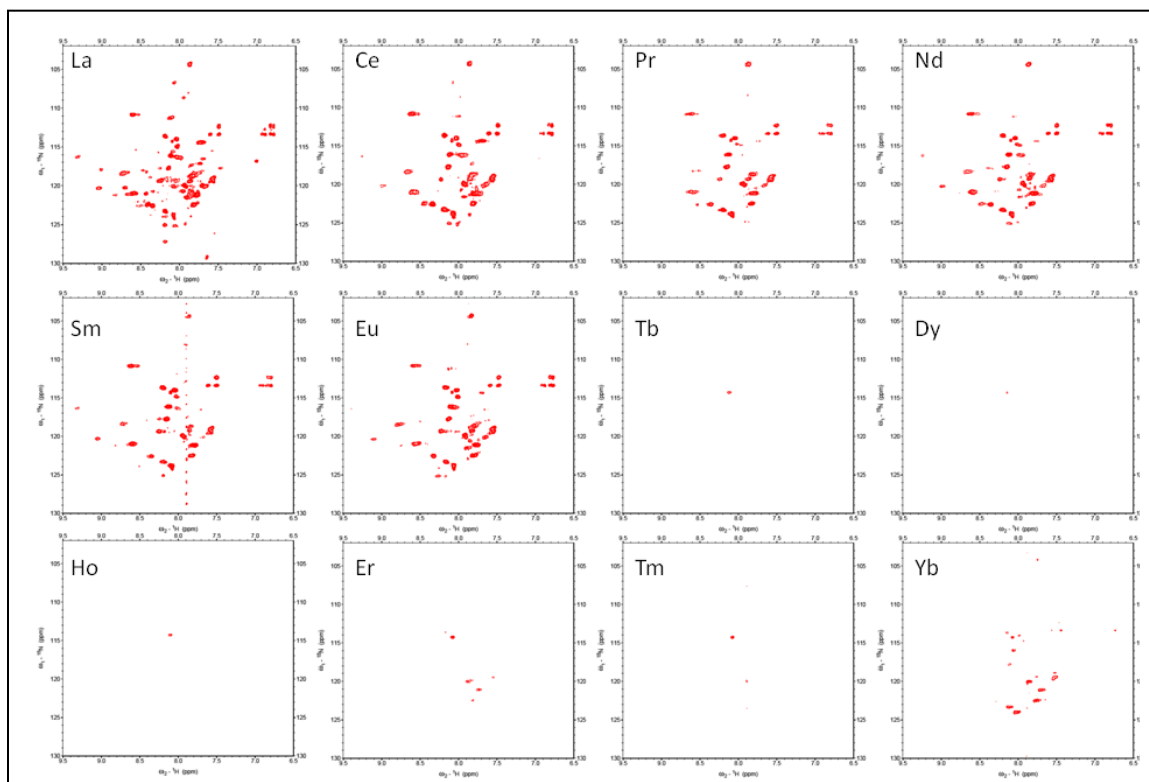
This protocol specifies how the tagging experiments were carried out on DAGK. For continued experiments it should be adapted according to the latest protocol on KCNE3, however, dimerization is not a problem for DAGK.

- solubilize and purify as standard protocol, only take most concentrated fraction during purification
- **DAGK eluted protein solution contained:**
 - 0.5% DPC
 - 250 mM imidazole
- **MTS-EDTA stock solution:**
 - 10 mg MTS-EDTA
 - 100 mM imidazole
 - 10% D2O
 - pH 6.5
 - in 1 ml water
 - gives 23 mM MTSL stock solution (stored at 4C)
- **for Griesinger tag (MTS-CA-EDTA):**
 - 10 mg tag
 - 100 mM imidazole
 - 10% D2O
 - pH 6.5
 - in 1 ml water

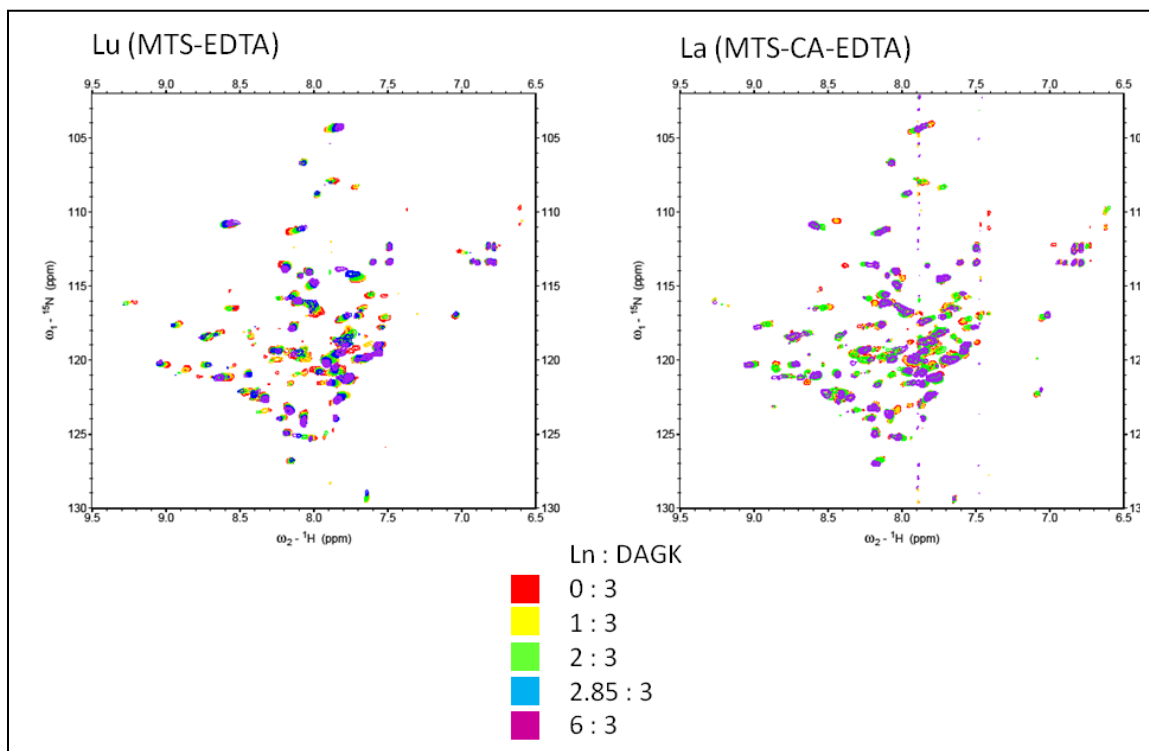
- gives 20.29 mM stock solution (stored at 4C)
- mix MTS-EDTA:DAGK monomer in 1.1:1 ratio
- mix overnight at room temperature
- desalting using PD10 column to get rid of excess tag, impurities and metal ions
- **desalting buffer:**
 - 0.5% DPC
 - 100 mM imidazole
 - (10% D2O)
 - pH 6.5
- desalting:
 - equilibrate with 25 ml buffer, discard flowthrough
 - load sample of 2.5 ml, discard flowthrough
 - elute with 3.5 ml buffer, collect
- **final NMR sample was:**
 - pH 6.5
 - add 10% D2O
 - NO EDTA!!!
 - titration points: Ln:DAGK monomer: 0:1; 1:3 ; 2:3 ; 0.95:1; 2:1
- **Ln stock solutions** (~10 ml or more of 100 mM):
 - 100 mM imidazole
 - 10% D2O
 - pH 6.5

NMR spectra of MTS-EDTA tagged DAGK

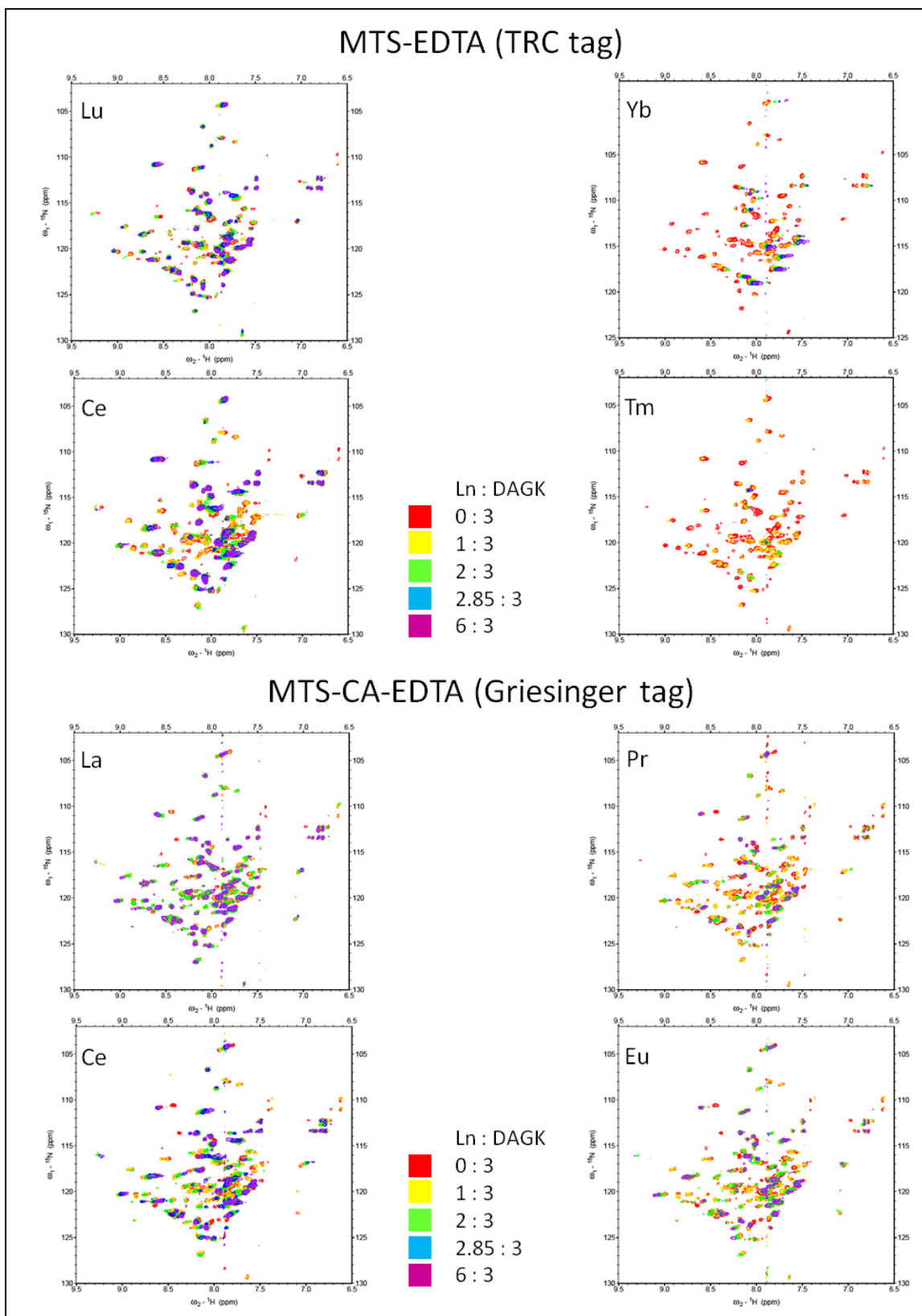
The ratio of lanthanide to DAGK is 0.95:1. The measurements were carried out at 800 MHz at 318 K.



Titration spectra of tagged DAGK with diamagnetic analogs

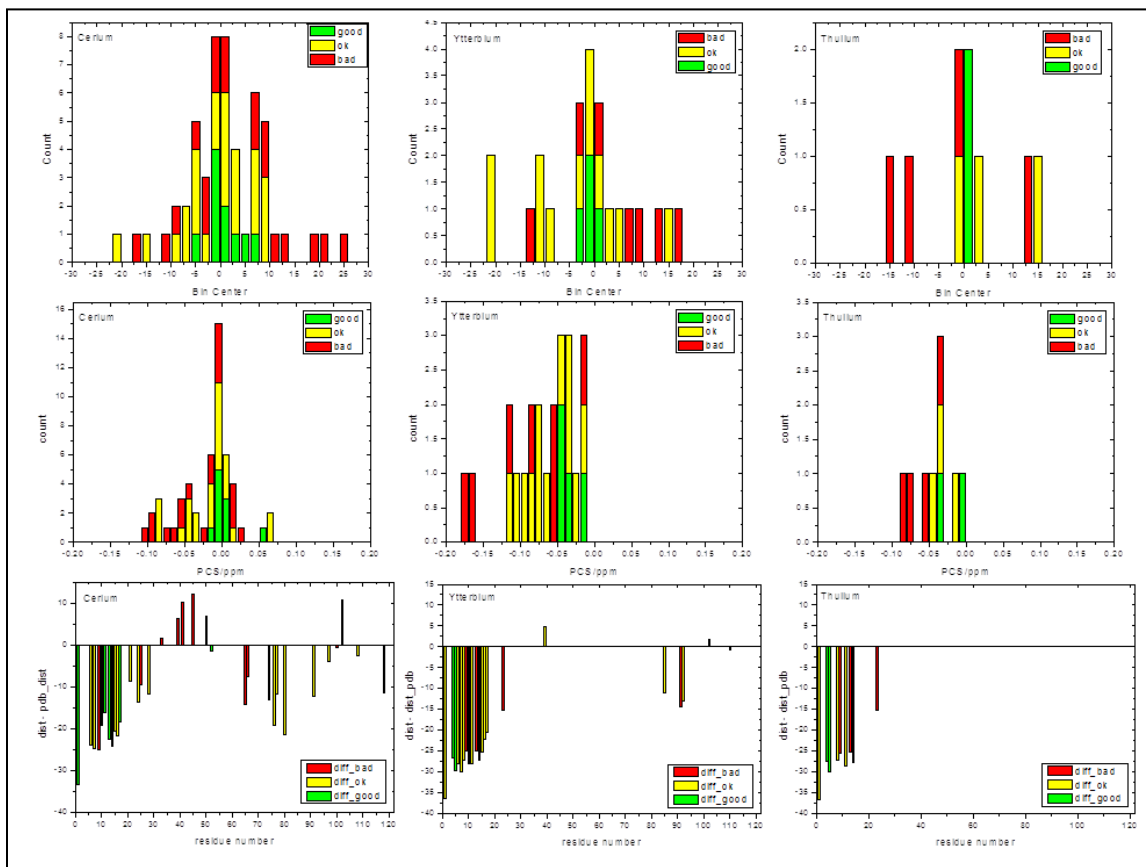


Comparing titration spectra of MTS-EDTA tag with MTS-CA-EDTA tag



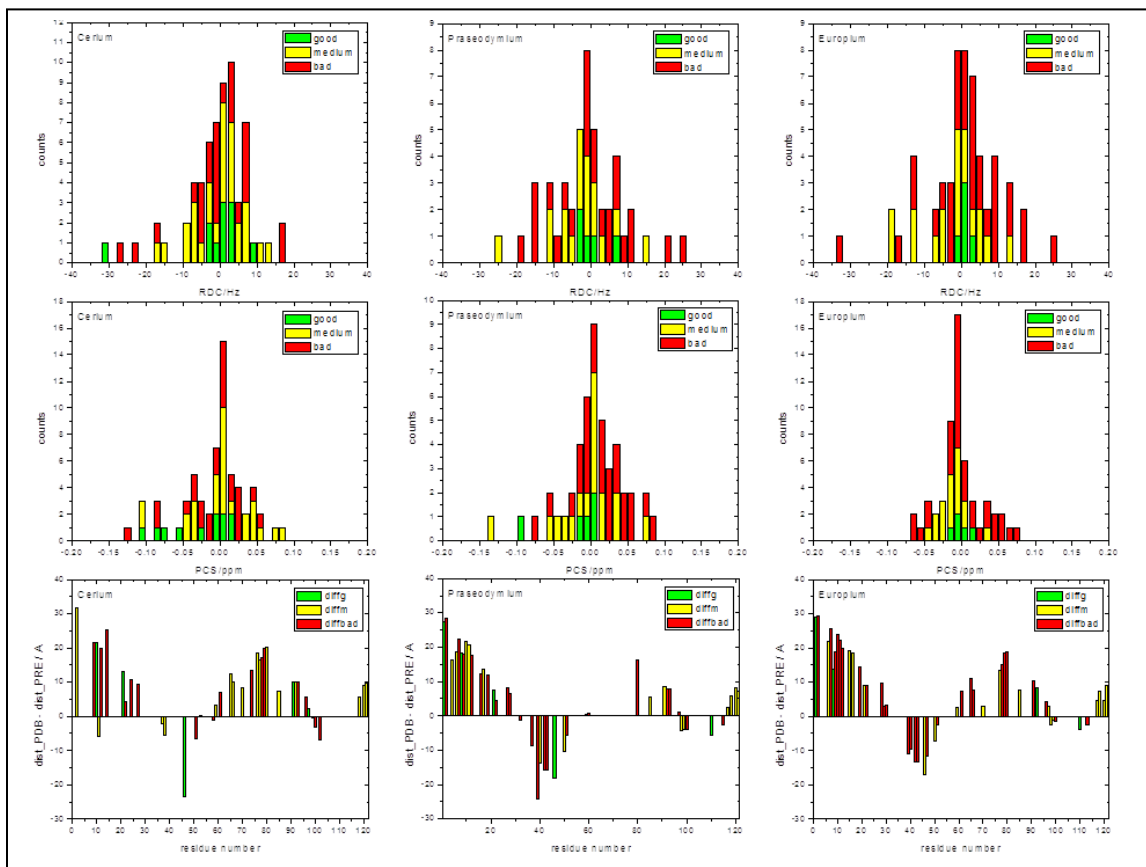
Paramagnetic restraints on DAGK using MTS-EDTA tag

All measurements were carried out with Lutetium as a reference. The restraints are color coded according to their quality in terms of correct peak assignment, signal-to-noise ratio, and peak splitting: green = high quality; yellow = intermediate quality; red = bad quality.



Paramagnetic restraints on DAGK using MTS-CA-EDTA tag

All measurements were carried out with Lanthanum as a reference. The restraints are color coded according to their quality in terms of correct peak assignment, signal-to-noise ratio, and peak splitting: green = high quality; yellow = intermediate quality; red = bad quality.



APPENDIX TO CHAPTER 3

Supplementary Table (I)

Composition of the membrane protein database used for the derivation of the
UHS

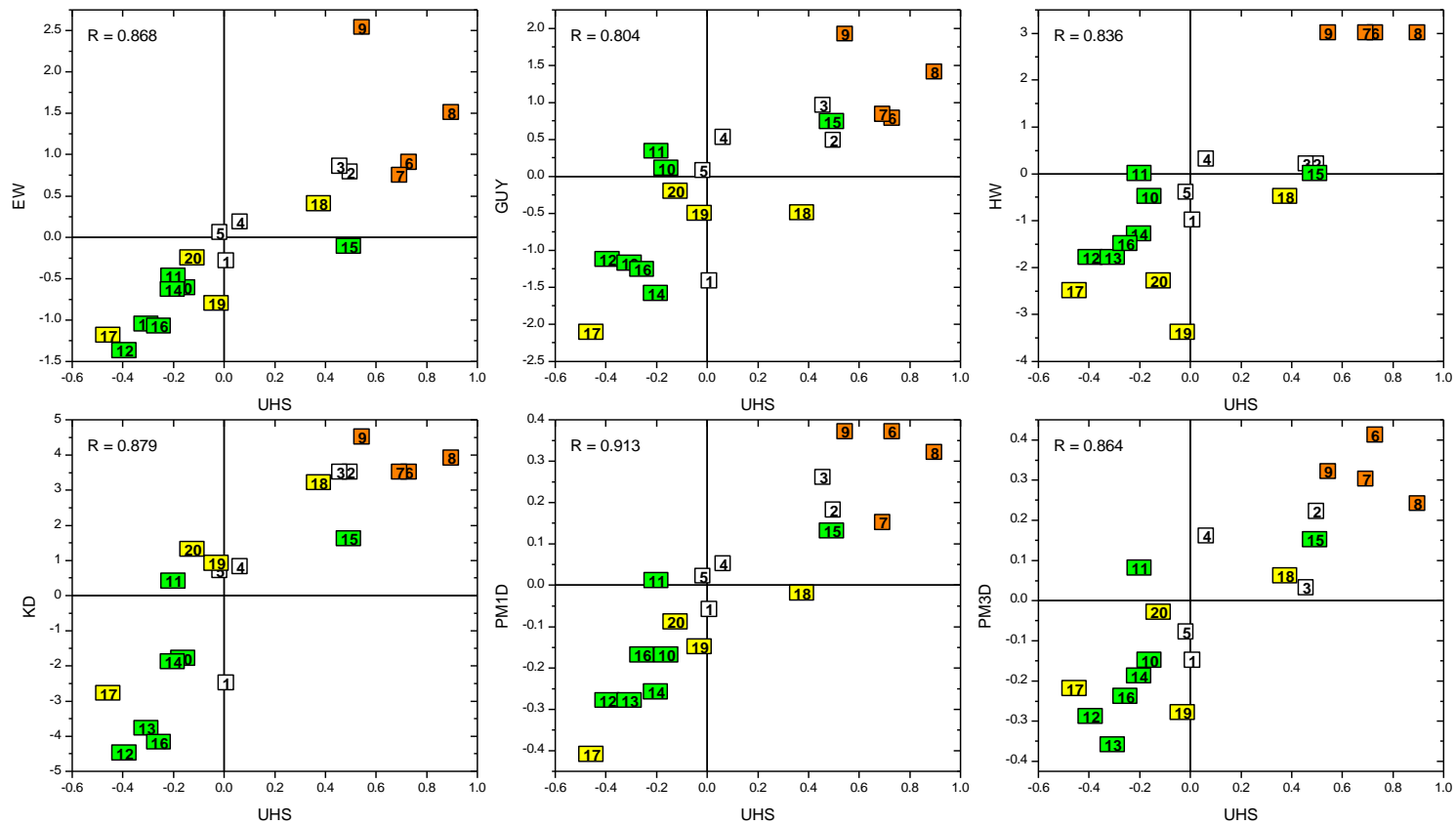
The MP database used for the derivation of the UHS was constructed by culling a complete list of multi-span MPs from the PDBTM with the PISCES server. The resulting MP database consists of 60 MPs and was divided into five parts for cross-validation. Each part contained approximately the same number of α -helices and β -sheets. Since the MPs in the database had very different sizes the number of proteins in the different datasets vary. The table shows the datasets that were used for cross-validation, where the columns represent the number of the dataset, the number of proteins in the dataset, the PDB code of the proteins, the number of α -helices and β -strands and the number of proteins in the database that were purely α -helical, β -barrels and which contained both secondary structure elements (from left to right). For cross-validation the free energies were derived five times for four of the datasets and tested on the remaining one.

| dataset | #proteins | PDB code of proteins | #α-helices | #β-strand | α | β | $\alpha+\beta$ |
|----------------|------------------|--|-------------------------------------|-----------------------------------|----------------------------|---------------------------|----------------------------------|
| 1 | 7 | 1I78, 1KMO, 1QFG, 1R3J, 1V54, 2BL2, 7AHL | 196 | 209 | 2 | 1 | 4 |
| 2 | 11 | 1PPJ, 1S3E, 1U7G, 1XRD, 1YMG, 1ZLL, 2BG9, 2CFQ, 2ERV, 2FGQ, 2MPR | 196 | 213 | 4 | 1 | 6 |
| 3 | 9 | 1C17, 1M0K, 1OKC, 1QJP, 1UUN, 1WAZ, 1YC9, 1YCE, 1YEW | 192 | 208 | 4 | 1 | 4 |
| 4 | 16 | 1EK9, 1EQ8, 1HXX, 1K24, 1KPL, 1P49, 1QD6, 1QJ8, 1T16, 1THQ, 1UYN, 1WP1, 1XME, 2A65, 2F2B, 2FBW | 196 | 217 | 4 | 1 | 11 |
| 5 | 17 | 1AFO, 1BA4, 1BZK, 1FDM, 1KQF, 1NKZ, 1NQE, 1P4T, 1RWT, 1RZH, 1U19, 1WPG, 1XKW, 1Y4Z, 1ZZA, 2F1V, 2POR | 197 | 209 | 7 | 1 | 9 |

Supplementary Figure 1

Correlation plots of the UHS with other scales

Plots showing the correlation of the hydrophobicity values in kcal/mol between the UHS and the scales from EW, Guy, HW, KD, PM1D and PM3D. The correlation coefficients are shown in the upper left corner of the plots. The amino acids are numbered according to the scheme on the right and colored according to their class: white = polar, red = charged, green = apolar, yellow = aromatic.



- | | | |
|----|---|---|
| 1 | = | C |
| 2 | = | N |
| 3 | = | Q |
| 4 | = | S |
| 5 | = | T |
| 6 | = | D |
| 7 | = | E |
| 8 | = | K |
| 9 | = | R |
| 10 | = | A |
| 11 | = | G |
| 12 | = | I |
| 13 | = | L |
| 14 | = | M |
| 15 | = | P |
| 16 | = | V |
| 17 | = | F |
| 18 | = | H |
| 19 | = | W |
| 20 | = | Y |

Supplementary Figure 2

Prediction of trans-membrane spans using a window for averaging

The figure shows the sliding-window approach for averaging the free energies for the prediction of trans-membrane spans from a protein sequence. The free energy is calculated as an average of the free energies of the amino acids located in the window where the middle residue has the highest weight. The result of the free energy is assigned to the central residue of the window.



Supplementary Table (II)

Over-prediction of amino acids in the soluble region as being in the membrane

To assess the over-prediction of amino acids in solution as being in the trans-membrane region the scales were tested on a dataset of non-redundant soluble proteins. The set was created by culling the PDB with the PISCES server as described in the Methods section. The set consisted of 2569 proteins with 3538 chains and 526,422 residues. The agreements are given in %.

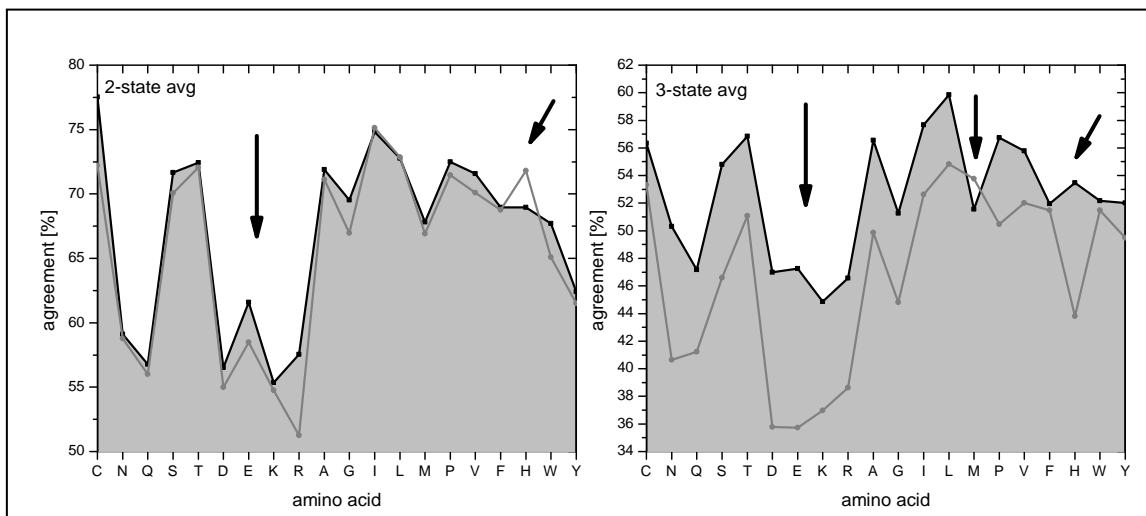
| | predicted SOL | predicted TM |
|-------|---------------|--------------|
| HWvH | 100 | 0 |
| WW | 95.6 | 4.4 |
| GES | 86.3 | 13.7 |
| UHS | 85.7 | 14.3 |
| Janin | 74.5 | 25.2 |
| KD | 63.2 | 36.8 |
| PM3D | 53.2 | 46.7 |
| Guy | 51.7 | 48.3 |
| PM1D | 50.2 | 49.7 |
| HW | 49.4 | 50.5 |
| EW | 44.3 | 55.6 |

Supplementary Figure 3

Performance of the UHS as seen for the individual amino acid averages

The figure shows agreements between the predicted and actual locations for the individual amino acids. Figure (a) shows the performance of the UHS (black) and the GES (gray) in two-state scenario (TM and SOL) where the averages of the diagonal matrix elements (compare Table (IV)) are plotted against the amino acids. Figure (b) shows the performance of the UHS (black) and the WW (gray) in the three-state scenario with the averages of the diagonal matrix elements (compare Table (V)). For both scenarios a window length of 15 residues was used for averaging. The details are given in the Results and Discussion section:

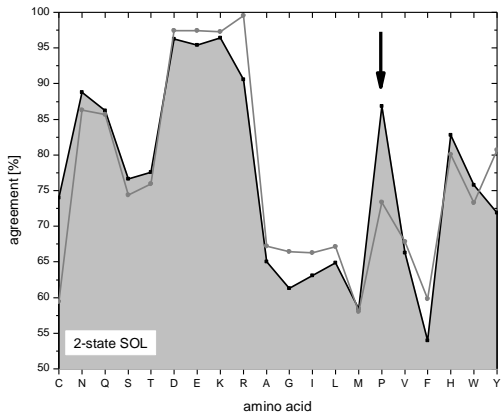
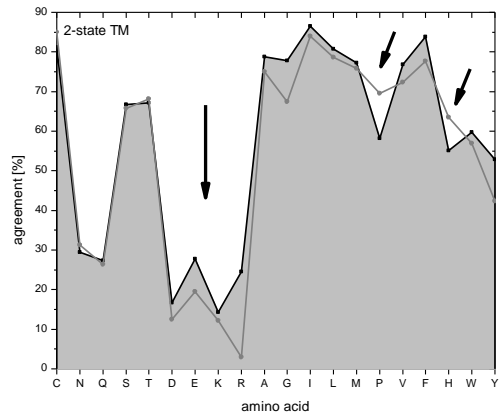
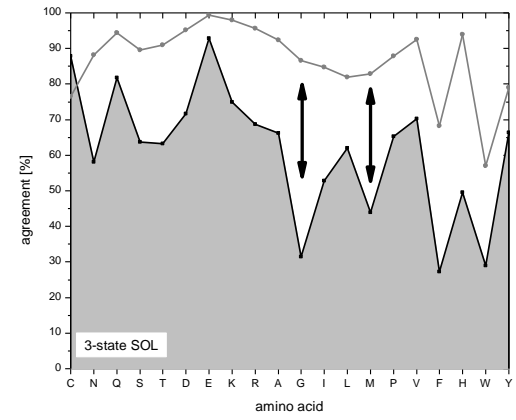
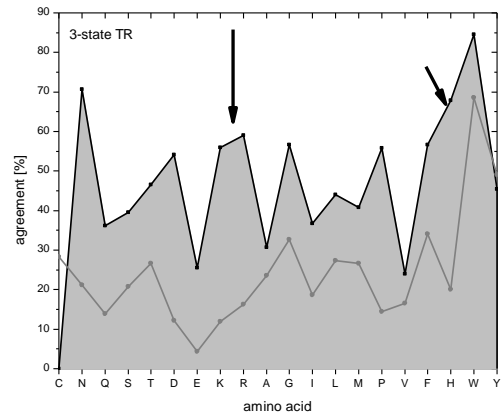
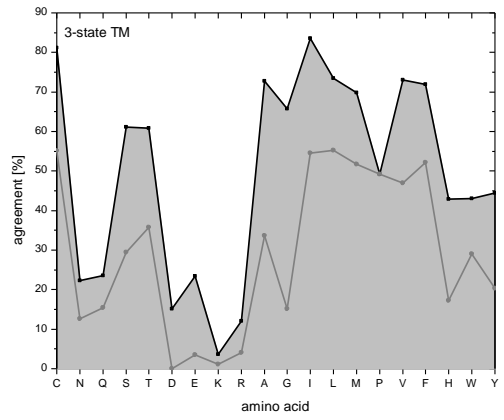
"Comparing the GES scale with the UHS, the average agreements have increased most for Arg (51% to 58%), Cys (72% to 78%), and Glu (58% to 62%). Note that the average agreement in the UHS is lower than in the GES scale only for His (72% to 69%). This indicates a slightly better representation of polar residues in the present UHS."



Supplementary Figure 4

Performance of the UHS as seen for the individual amino acids in the different regions

The figure shows the individual amino acid agreements in the three-state (a-c) and two-state (d & e) scenario at a 15 residue window length for the UHS (black line) and the WW (gray in the upper panel) or the GES (gray in the lower panel). (a) 3-state TM agreement; (b) 3-state TR agreement; (c) 3-state SOL agreement; (d) 2-state TM agreement; (e) 2-state SOL agreement. It can be seen that in the three-state scenario "the polar residues Arg, Asn, Asp, Glu, Gln, His, Lys, and Ser are predicted in a more balanced manner in the UHS than in the WW scale. When comparing the overall prediction accuracies, all amino acids either display an improvement or at least a similar accuracy for the UHS. Highest changes are observed for Asp and Glu (from 36% to 47%), Asn (from 41% to 50%), and His (from 44% to 53%)." (see Results and Discussion).



Supplementary Information

The UHS is largely independent of the protein fold

We systematically excluded folds when deriving the UHS to address the question whether or not our scale is biased towards protein folds represented in the PDB. The following five folds were excluded one by one: aquaporins, outer membrane proteins, porins, bacteriorhodopsin, and the potassium channel (see Supplementary Table (III)).


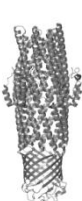

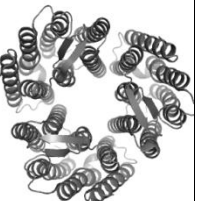
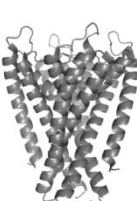
The hydrophobicity values that were derived without these different folds deviate on average 0.6 standard deviations from the UHS with a maximal deviation of three standard deviations for Glu in class 3. The largest deviations occur for classes 2 and 3. These changes are small in actual numbers given the range of hydrophobicity values. This indicates that the hydrophobicity value derived here is mostly an amino acid centered property largely independent of the fold of the protein. Further, the five resulting "leave-one-fold-out" UHS scales were used to predict TM and SOL regions within the "left-out" folds. The results of this experiment are summarized in Supplementary Table (IV).

The performance of these "leave-one-fold-out" UHS scales agrees on average to within 2.4% accuracy compared to the performance of the UHS scale. The largest deviations are 3% (SOL) for class 1 (1.51% for the average), 3.2% (SOL) for class 2 (1.92% for the average), 8.8% (TM) for class 3 (2.95% for the average), 1.4% (SOL) for class 4 (0.71% for the average), and 3.7% (SOL) for class 5 (1.84% for the average). This supports our argument that the UHS scale is largely fold independent.

Supplementary Table (III)

The UHS is largely independent of the protein fold (continued)

Classes of proteins that were excluded from the derivation to assess the performance of the scale on novel folds. The UHS was derived when the following folds were excluded one by one and then tested on the excluded folds.

| | | | | | |
|-----------|---|---|---|---|---|
| class # | 1 | 2 | 3 | 4 | 5 |
| class | aquaporins | outer membrane proteins | porins | bacterio-rhodopsin | potassium channel |
| #proteins | 2 | 3 | 3 | 1 | 1 |
| PDB ID | 1YMG 2F2B | 1EK9 1WP1 1YC9 | 1HXX 2FGQ 2MPR | 1M0K | 1R3J |
| #AAs | 1912 | 3885 | 3273 | 666 | 412 |
| fold |  |  |  |  |  |

Supplementary Table (IV)

The UHS is largely independent of the protein fold (continued)

The table shows the performance of the UHS for folds that have not been used for the derivation of the scale. #1 to #5 are the class numbers from Supplementary Table (III).

| | | <i>PDB</i> | | | <i>PDB</i> | | |
|-------------|------------|------------|------------|------------|------------|------------|------------|
| | | <i>TM</i> | <i>SOL</i> | <i>avg</i> | <i>TM</i> | <i>SOL</i> | <i>avg</i> |
| <i>pred</i> | <i>TM</i> | #1 | | | #2 | | |
| | <i>SOL</i> | | 85.0 | 60.4 | | 47.0 | 16.3 |
| | | | 14.9 | 38.9 | | 53.0 | 83.5 |
| | | | | 61.95 | | | 65.26 |
| <i>pred</i> | <i>TM</i> | #3 | | | #4 | | |
| | <i>SOL</i> | | 18.1 | 9.2 | | 92.4 | 54.7 |
| | | | 81.5 | 90.7 | | 7.6 | 43.3 |
| | | | | 54.42 | | | 67.84 |
| <i>pred</i> | <i>TM</i> | #5 | | | | | |
| | <i>SOL</i> | | 92.9 | 46.3 | | | |
| | | | 7.1 | 50.8 | | | |
| | | | | 71.84 | | | |

Supplementary Table (V)

The performance of the MHS in the two-state scenario

The prediction quality of the MHS was assessed by cross-validation and by testing the scale on the bacterial part of the MP database (*bact* in this table). The agreements of the MHS from cross-validation are very high for SOL (89.0%) and somewhat lower for the TM region (77.2%). The average agreement is therefore 83.1% which is the highest agreement of a hydrophobicity scale in this paper. When the MHS is tested on a bacterial dataset, the agreement in SOL decreases to 51.0%, leaving an average agreement of 67.74%. These results are somewhat expected considering that the database used for the MHS only consists of α -helical proteins that are easier to predict than β -barrels (see below). In contrast, the bacterial database includes β -barrel proteins explaining the lower agreement on this set.

| | | <i>PDB</i> | | |
|-------------|-------------------------|-------------|------------|------------|
| | | <i>TM</i> | <i>SOL</i> | <i>avg</i> |
| <i>pred</i> | <i>TM</i> <i>SOL</i> | <i>MHS</i> | | |
| | | | 77.2 | 10.9 |
| | | | 22.8 | 89.0 |
| | | | | 83.08 |
| <i>pred</i> | <i>TM</i> <i>SOL</i> | <i>bact</i> | | |
| | | | 51.0 | 15.5 |
| | | | 49.0 | 84.5 |
| | | | | 67.74 |

APPENDIX TO CHAPTER 5

BCL::Jufo9D

A note on dataset creation

It was noticed that after using PISCES to exclude similar sequences at a sequence similarity < 30% there were still sequences in the databases that had a higher sequence similarity, even over 90%. To mitigate that problem all sequences in both the MP and soluble protein database were pairwise aligned using BCL::Align, the sequence identity was calculated, and the sequences were clustered according to sequence identity using BCL::Cluster with a cutoff of 30%. Only the cluster center was retained and the other protein chains were discarded.

Commandlines for BCL::Jufo9D

All scripts, executables, and files are provided on the DVD. The commandlines are given for the example 1a0t.

Dataset creation

First, unnatural amino acids are converted into natural counterparts using

```
bcl2011-12-14.exe PDBConvert database_MPs/1a0t.pdb -fasta -  
bcl_pdb -output_prefix database_MPs/1a0t_1 -  
convert_to_natural_aa_type
```

To create the secondary structure predictions and position-specific scoring matrices after six iterations of PsiBlast

```
runss 1a0t.fasta 6
```

is run over fasta files. DSSP is run over all PDBs using the script 001_run_dssp_over_db.pl which first removes the HELIX and SHEET lines, runs

DSSP with the -35 mode that includes 3_{10} and π -helices, and creates new PDB files from the DSSP output. The biomolecules are created from the PDB files using

```
bcl2011-12-14.exe PDBConvert database_MPs/1a0t_2.pdb -bcl_pdb -  
output_prefix database_MPs/1a0tbio -pdbtm_xml  
database_MPs/1a0t.xml -biomolecule 1 -helix_classes 1 5
```

To remove sequences that are mistakenly included by the PISCES server and have a sequence identity higher than 30%, a pairwise sequence alignment is carried out using `005_sequence_alignment_calc_seqid.pl`. This script makes a pairwise sequence alignment using `BCL::Align` and calculates the sequence identity. The script `008_compute_seqid_matrix.pl` is used to create the input matrix for clustering. The clusters are analyzed with `010_analyze_clusters.pl` and can be visualized individually in PyMol using `011_visualize_clusters.pl` which creates a PyMol script as output file. The script `012_create_oligomeric_state_dictionary.pl` creates the oligomeric state dictionary file that contains the oligomeric state for each protein. 0 represent monomer and 1 represents a multimer.

For membrane proteins the proteins are classified as α -helical or β -barrel in the membrane region and all helical proteins after clustering are randomly divided into five subsets. The β -barrels are also distributed into five subsets. For each of the subsets, for example `dataset1.ls`, the ANN input file is created using

```
bcl-apps-static2012-02-17.exe GenerateJufoDescription -pdb_list  
dataset1.ls -path database_all/ -oligo-dict  
oligomeric_state_dictionary.txt -membrane_orientation_path  
database_all/ -output_prefix descriptors1_PDB_9D_10000_w31.dat -  
convert_to_natural_aa_type -nr_entries_per_state 10000 -  
creating_input_for_first_layer -window_radius 15
```

and then converted into .bin format using

```
bclWill.exe GenerateDataset -source  
'File(filename=descriptors_PDB_9D_10000_w31.dat, number chunks=1,  
chunks=[0])' -output descriptors_PDB_9D_10000_w31.bin
```

The for the training runs the following three commandlines are used:

```
bcl-apps-static_03192012.exe TrainModel `NeuralNetwork(transfer  
function = Sigmoid, weight update = Simple(eta=0.001, alpha = 0),  
objective function = RMSD, steps per update = 1, hidden  
architecture(32))' -training `Subset(filename =  
descriptors_PDB_9D_10000_w31.bin, number chunks = 5, chunks  
="[0,5]-[0]-[1]")' -monitoring `Subset(filename =  
descriptors_PDB_9D_10000_w31.bin, number chunks = 5, chunks  
="[1]")' -independent `Subset(filename =  
descriptors_PDB_9D_10000_w31.bin, number chunks = 5, chunks  
="[0]")' -print_training_predictions  
descriptors_PDB_9D_10000_w31_2012-03-16.train0 -  
print_monitoring_predictions descriptors_PDB_9D_10000_w31_2012-  
03-16.mon0 -print_independent_predictions  
descriptors_PDB_9D_10000_w31_2012-03-16.ind0 -feature_labels  
features_code_1282.object -result_labels results_code_9D.object -  
scheduler Serial -final_objective_function RMSD -storage_model  
'File(directory = .)' -max_iterations 50
```

```
bcl-apps-static_03192012.exe TrainModel `NeuralNetwork(initial  
network file = 000000.model, transfer function = Sigmoid, weight  
update = Simple(eta=0.000005, alpha = 0.5), objective function =  
RMSD, steps per update = 1, hidden architecture(32))' -training
```

```

`Subset(filename = descriptors_PDB_9D_10000_w31.bin, number
chunks = 5, chunks = "[0,5]-[0]-[1]")' -monitoring
`Subset(filename = descriptors_PDB_9D_10000_w31.bin, number
chunks = 5, chunks = "[1]")' -independent `Subset(filename =
descriptors_PDB_9D_10000_w31.bin, number chunks = 5, chunks
="[0]")' -print_training_predictions
descriptors_PDB_9D_10000_w31_2012-03-16.train1 -
print_monitoring_predictions descriptors_PDB_9D_10000_w31_2012-
03-16.mon1 -print_independent_predictions
descriptors_PDB_9D_10000_w31_2012-03-16.ind1 -feature_labels
features_code_1282.object -result_labels results_code_9D.object -
scheduler Serial -final_objective_function RMSD -storage_model
`File(directory = .)' -max_iterations 10

bcl-apps-static_03192012.exe TrainModel `NeuralNetwork(initial
network file = 000001.model, transfer function = Sigmoid, weight
update = Simple(eta=0.000005, alpha = 1), objective function =
RMSD, steps per update = 1, hidden architecture(32))' -training
`Subset(filename = descriptors_PDB_9D_10000_w31.bin, number
chunks = 5, chunks = "[0,5]-[0]-[1]")' -monitoring
`Subset(filename = descriptors_PDB_9D_10000_w31.bin, number
chunks = 5, chunks = "[1]")' -independent `Subset(filename =
descriptors_PDB_9D_10000_w31.bin, number chunks = 5, chunks
="[0]")' -print_training_predictions
descriptors_PDB_9D_10000_w31_2012-03-16.train2 -
print_monitoring_predictions descriptors_PDB_9D_10000_w31_2012-

```

```
03-16.mon2 -print_independent_predictions
descriptors_PDB_9D_10000_w31_2012-03-16.ind2 -feature_labels
features_code_1282.object -result_labels results_code_9D.object -
scheduler Serial -final_objective_function RMSD -storage_model
'File(directory = .)' -max_iterations 100
```

The training, independent, or monitoring files can be analyzed using
019_analyze_overprediction_9D_newformat.pl.
020_avg_ANN_outputs_to_1_prediction.pl averages the output of four ANNs
and compares it to the desired output.

First version of BCL::Jufo9D

Methods

The methods for the first version of BCL::Jufo9D were identical to the ones presented in Chapter 5, except for four things:

1. The sequence similarity cutoff was 25%.
2. The datasets were split into 10 subsets where the MP database was split up visually such that each of the subsets would contain similar representatives of folds. This means each of the subsets contained single TM helix proteins, small β -barrels, large β -barrels, and so on. The soluble protein database was still split up randomly.
3. For cross-validation, one subset was used as independent set, one was used for monitoring the training process, and eight subsets were used for training. One of the important differences was that for cross-validation, only the monitoring subset was permuted, but not the independent subset.
4. At that time, it was not noticed that PISCES would not always properly exclude similar folds. Therefore, the sequence alignments and clustering procedure described above were omitted.

Results

The following results are per-residue accuracies on whole chains of the independent test set. Nine-state accuracies:

| 9-state | | PREDICTION | | | | | | | | |
|---------|-------|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | | tm_C | tr_C | sol_C | tm_E | tr_E | sol_E | tm_H | tr_H | sol_H |
| PDB | tm_C | 0.9 | 13.0 | 13.0 | 2.8 | 4.6 | 5.6 | 34.3 | 10.2 | 15.7 |
| | tr_C | 0.4 | 38.8 | 28.0 | 0.2 | 6.3 | 5.2 | 0.4 | 16.4 | 4.3 |
| | sol_C | 0.6 | 3.3 | 66.1 | 0.4 | 0.5 | 14.5 | 0.3 | 0.7 | 13.7 |
| | tm_E | 0.3 | 1.1 | 5.7 | 51.4 | 9.2 | 29.6 | 0.0 | 0.3 | 2.4 |
| | tr_E | 1.6 | 9.8 | 8.9 | 16.3 | 56.9 | 4.9 | 0.0 | 0.0 | 1.6 |
| | sol_E | 0.4 | 0.8 | 15.9 | 1.1 | 0.9 | 75.9 | 0.3 | 0.6 | 4.1 |
| | tm_H | 1.2 | 2.0 | 3.6 | 0.0 | 0.1 | 3.6 | 75.6 | 11.9 | 2.0 |
| | tr_H | 0.3 | 13.1 | 10.4 | 0.3 | 0.1 | 5.1 | 13.0 | 51.9 | 5.9 |
| | sol_H | 0.3 | 0.8 | 16.1 | 0.2 | 0.2 | 4.4 | 0.8 | 1.9 | 75.2 |

Secondary structure prediction:

| | | PREDICTION | | | | |
|-----|---|----------------|--------------|--------------|--------------|--------------|
| | | SSPred | H | E | C | avg |
| PDB | H | | 78.11 | 4.83 | 17.06 | 75.04 |
| | E | | 4.87 | 78.15 | 16.98 | |
| | C | | 14.86 | 15.31 | 69.83 | |
| | | | | | | |
| | | PsiPred | H | E | C | avg |
| PDB | H | | 77.61 | 2.06 | 20.33 | 78.06 |
| | E | | 2.88 | 79.43 | 17.69 | |
| | C | | 10.92 | 11.26 | 77.82 | |
| | | | | | | |
| | | Jufo | H | E | C | avg |
| PDB | H | | 75.25 | 3.01 | 21.74 | 74.21 |
| | E | | 7.33 | 73.55 | 19.12 | |
| | C | | 11.51 | 15.12 | 73.37 | |
| | | | | | | |
| | | ProfPhd | H | E | C | avg |
| PDB | H | | 70.07 | 8.15 | 21.77 | 66.68 |
| | E | | 16.61 | 60.77 | 22.62 | |
| | C | | 17.6 | 16.34 | 66.06 | |
| | | | | | | |

TM span prediction in three states:

| | | PREDICTION | | | |
|-----|--------|--------------|--------------|--------------|--------------|
| | TMPred | sol | tr | tm | avg |
| PDB | sol | 95.26 | 3.33 | 1.41 | |
| | tr | 26.8 | 63.95 | 9.25 | |
| | tm | 17.97 | 14.15 | 67.87 | |
| | | | | | 94.25 |
| | UHS | sol | tr | tm | avg |
| PDB | sol | 60.75 | 29.26 | 9.98 | |
| | tr | 27.07 | 47.29 | 25.63 | |
| | tm | 10.93 | 21.07 | 68.01 | |
| | | | | | 60.65 |

TM span prediction in two states for a subset of the independent dataset:

TM helix prediction:

| | | PREDICTION | | |
|-----|----------|------------|---------|-----|
| | TM-helix | TM-H | no TM-H | avg |
| PDB | TM-H | 83.28 | 16.72 | |
| | no TM-H | 3.40 | 96.60 | |
| | | | | |
| | CONPRED | TM-H | no TM-H | avg |
| PDB | TM-H | 91.1 | 8.9 | |
| | no TM-H | 9.7 | 90.3 | |
| | | | | |
| | TMMOD | TM-H | no TM-H | avg |
| PDB | TM-H | 91.3 | 8.7 | |
| | no TM-H | 9.9 | 90.1 | |
| | | | | |

TM strand prediction

| | | PREDICTION | | |
|-----|-----------|--------------|--------------|-----|
| | TM-strand | TM-S | no TM-S | avg |
| PDB | TM-S | 49.39 | 50.61 | |
| | no TM-S | 0.91 | 99.09 | |
| | | | | |
| | HMM | TM-S | no TM-S | avg |
| PDB | TM-S | 95.5 | 4.5 | |
| | no TM-S | 10.7 | 89.3 | |
| | | | | |

Discussion

These results indicate that the independent test set is a poor choice and that accuracies could be increased if an average over several independent test sets would be computed. Therefore, in later studies, both monitoring and independent test set were permuted.

Since the earlier version of BCL::Jufo9D did not outperform PsiPred, we hypothesized that by excluding similar folds (as was done for training PsiPred) the prediction accuracies would increase because a bias in fold families would be removed.

Furthermore, some of the β -barrel proteins were poorly identified, either TM β -barrels were predicted to be water soluble or soluble fibrils were predicted to be in the membrane. To address these issues, we hypothesized that inclusion of side-chain information in membrane and transition states would account for a more accurate representation of residue environment.

BCL::Jufo16D

Methods

For the membrane protein database, all chains from the PDBTM were culled by PISCES with a sequence similarity cutoff of 25%. For the soluble protein database, all chains from the PDB were culled using PISCES and MP chains were removed. Additionally to culling by sequence similarity, similar folds were removed by a pairwise structure-structure alignment using MAMMOTH, clustering by MAMMOTH Z-score using BCL::Cluster, and retaining only the chain at the cluster center.

The resulting MP chains were split into five subsets, as in the earlier version of BCL::Jufo9D they were split up visually such that each of the subsets would contain similar representatives of “folds”. For the soluble protein database the resulting chains were randomly split up into five subsets.

For cross-validation, both the monitoring and the independent dataset were permuted through all of the five subsets.

To address the incorrect identification of some of the β -strands two modifications were made to the algorithm: (1) for each residue the solvent-accessible surface area (SASA) was included as a prediction output; (2) for each residue the side-chain orientation (water vs. lipid) in the membrane core and transition region was included as a prediction output. Whereas the SASA could easily be computed by a BCL application that used the overlapping sphere algorithm, the side-chain orientation in membrane and transition states had to be visually estimated using PyMol. Ultimately, the 9-dimensional output vector was replaced by a 16-dimensional output vector:

| | | | | | | | |
|-----------|---|---|---|------------|-------|---|---|
| | H | S | C | | H | S | C |
| SOL | 1 | 0 | 0 | SOL | 1 | 0 | 0 |
| TR | 0 | 0 | 0 | TR, lipid | 0 | 0 | 0 |
| TM | 0 | 0 | 0 | TR, h2o | 0 | 0 | 0 |
| | | | | TM, lipid | 0 | 0 | 0 |
| | | | | TM, h2o | 0 | 0 | 0 |
| | | | | exposure | 0...1 | | |
| 9 outputs | | | | 16 outputs | | | |

Results

After many training iterations with different parameters, the prediction accuracies did not exceed 67% for three-state secondary structure prediction, 75% for three-state TM span prediction, 67% for side-chain orientation, and had an RMSD of 0.4 (between 0 and 1) for the SASA as an exposure measure. Similar or lower accuracies were achieved by using the nine dimensional output vector.

It was also found that using 5-fold or 10-fold cross-validation does not have a significant influence on the accuracy.

Discussion

The results led us to the conclusion that the residue occurrences in each of the 15 states were too low to yield high-accuracy predictions. Especially excluding similar folds, which resulted in removal of proteins while splitting up the residues into 15 bins, was not a meaningful approach.

BCL::Jufo9D with 90% sequence similarity cutoff

Motivation

The conclusions from the previous experience were that the residue occurrences in each of the 15 states were too low. To account for that, we returned to our previous approach to use a nine-state output and also included as much MP information as possible.

Methods

All MP chains from the PDBTM were culled using the PISCES server with a sequence similarity cutoff of 80%. The RMSDs of these structures were ≤ 5 Å where EM structures with a resolution above this threshold were removed. The sequences were pairwise aligned using BCL::Align, the sequence identities were calculated, and BCL::Cluster was used to cluster the sequences according to sequence identity. The 62 clusters present at 30% sequence similarity were equally distributed into five subsets for cross-validation. The MPs were transformed into the membrane coordinate frame, and DSSP was used to standardize secondary structure representation.

For the soluble protein database, a pre-compiled list with a sequence similarity cutoff of 30% from the PISCES website was used, MP chains were removed, and chains shorter than 40 residues were removed.

Results

The following results are for the averages of prediction accuracies over complete sequences in the independent datasets. The outputs over four ANNs are also averaged which have the same independent dataset. The prediction accuracies over all nine states are 71.19%, for SS prediction 74.98%, and for TM prediction 93.59%. The rows represent “true” states, whereas the columns represent predicted states.

| | MC-H | MC-E | MC-C | TR-H | TR-E | TR-C | SO-H | SO-E | SO-C |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| MC-H | 76.65 | 0.67 | 4.69 | 13.09 | 0.30 | 1.46 | 1.45 | 1.12 | 0.57 |
| MC-E | 0.15 | 78.77 | 3.74 | 0.24 | 12.73 | 0.68 | 0.39 | 2.84 | 0.45 |
| MC-C | 19.90 | 8.32 | 30.53 | 9.64 | 12.88 | 13.12 | 0.72 | 2.15 | 2.72 |
| TR-H | 15.48 | 0.55 | 2.92 | 60.07 | 0.79 | 10.49 | 6.14 | 2.00 | 1.55 |
| TR-E | 0.10 | 11.84 | 5.33 | 3.51 | 62.78 | 9.31 | 0.79 | 4.09 | 2.24 |
| TR-C | 2.31 | 1.74 | 8.18 | 21.22 | 14.20 | 38.82 | 1.59 | 2.85 | 9.08 |
| SO-H | 0.67 | 0.54 | 0.51 | 1.23 | 0.35 | 0.55 | 77.23 | 6.79 | 12.14 |
| SO-E | 0.19 | 1.69 | 1.08 | 0.78 | 1.87 | 0.63 | 6.56 | 73.12 | 14.08 |
| SO-C | 0.17 | 0.69 | 1.67 | 0.70 | 1.27 | 2.57 | 12.15 | 14.82 | 65.98 |

Discussion

Even though the per-residue accuracies are in the ballpark of what is expected for these predictions, it was noticed that the outputs are very noisy when plotted over the sequence.

Furthermore, we wanted to test whether including MP sequence information where structures are not yet know, would improve the prediction accuracies.

BCL::Jufo9D with additional MP sequence information

Motivation

To include even more MP data, we wanted to test whether the inclusion of additional sequence information, for which no structural information is yet obtained, would increase prediction accuracies.

Methods

All MP chains from the PDBTM were culled using the PISCES server with a sequence similarity cutoff of 90%. The RMSDs of these structures were ≤ 5 Å where EM structures with a resolution above this threshold were removed. The sequences were pairwise aligned using BCL::Align, the sequence identities were calculated, and BCL::Cluster was used to cluster the sequences according to sequence identity. The 62 clusters present at 30% sequence similarity were equally distributed into five subsets for cross-validation. The MPs were transformed into the membrane coordinate frame, and DSSP was used to standardize secondary structure representation.

Additionally, PSI-Blast was used on all MP sequences (templates) to identify similar sequences. Of these ~210,000 sequences, ~112,000 of them were unique whereas the others were repetitions. To retain feasibility, for each original MP template sequence the top 50 hits below an E-value = 0.01 were used, adding up to ~16,000 sequences, 5,144 of which contained TM spans. To reduce the possibility that these additional sequences would be too divergent from the original template sequence as to possess a different fold, only sequences above 50% sequence similarity were retained. Since none of these protein sequences had their structure resolved, structural information (SS and TM region) was obtained from the corresponding residue in the template structure according to the sequence alignment. These “dummy” proteins were distributed into the same subset as their template sequence.

For the soluble protein database, a pre-compiled list with a sequence similarity cutoff of 30% from the PISCES website was used, MP chains were removed, and chains shorter than 40 residues were removed.

Results

The following results are for the averages of prediction accuracies over complete sequences in the independent datasets. The outputs over four ANNs are also averaged which have the same independent dataset. The prediction accuracies over all nine states are 67.39%, for SS prediction 71.98%, and for TM prediction 91.77%. The rows represent “true” states, whereas the columns represent predicted states.

| | MC-H | MC-E | MC-C | TR-H | TR-E | TR-C | SO-H | SO-E | SO-C |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| MC-H | 80.05 | 0.68 | 1.10 | 13.74 | 0.29 | 0.92 | 1.43 | 1.16 | 0.61 |
| MC-E | 0.34 | 80.64 | 3.54 | 0.41 | 9.73 | 1.18 | 0.30 | 3.41 | 0.45 |
| MC-C | 26.64 | 10.48 | 13.67 | 14.49 | 12.34 | 12.55 | 0.64 | 3.88 | 5.31 |
| TR-H | 19.44 | 0.73 | 0.92 | 60.45 | 0.75 | 5.27 | 7.40 | 2.91 | 2.12 |
| TR-E | 0.27 | 15.76 | 5.20 | 4.79 | 53.86 | 11.32 | 0.00 | 5.29 | 2.94 |
| TR-C | 4.71 | 2.45 | 4.19 | 30.54 | 14.13 | 27.10 | 2.02 | 4.09 | 10.77 |
| SO-H | 1.12 | 0.94 | 0.62 | 2.17 | 0.46 | 0.65 | 73.43 | 8.71 | 11.90 |
| SO-E | 0.45 | 2.49 | 1.17 | 1.22 | 1.65 | 0.82 | 7.57 | 70.28 | 14.35 |
| SO-C | 0.43 | 0.93 | 1.75 | 1.19 | 1.50 | 2.90 | 13.43 | 16.28 | 61.59 |

Discussion

Even though the per-residue accuracies are in the ballpark of what is expected for these predictions, it was noticed that the outputs are very noisy when plotted over the sequence. This led us to conclude that the clustering approach might actually induce noise because it is more difficult for the ANN to generalize if the proteins in the clusters are very similar internally, but very different to the proteins in the other clusters. As a

next step, we returned to a 30% sequence similarity cutoff for membrane proteins with a random distribution of sequences into the subsets for cross-validation, which is described in Chapter 5. Overall, the noise is much reduced for a random distribution of sequences into the subsets because it is much easier for the ANN to describe and abstract from a hyper-dimensional space when the datapoints for training are evenly distributed. In contrast, when the datapoints are clustered in this descriptor space, it is more difficult for the ANN to recognize a certain pattern and especially abstract it into the hyper-space that was not described in the training process but is described by an independent test set.

SUPPLEMENT

Lyso-Phospholipid Micelles Sustain the Stability and Catalytic Activity of Diacylglycerol Kinase in the Absence of Lipids¹

Introduction

Solution NMR and X-ray crystallographic structural studies of purified integral membrane proteins are often carried out in detergent micelle solutions, an imperfect medium given that protein-lipid interactions are sometimes both specific and important to integral membrane protein structure and function (4-7). Moreover, it is now clear that some high resolution structures of membrane proteins include micelle-generated distortions (8-11) and also that the energetics of membrane protein folding and intermolecular interactions can be altered in micelles relative to native-like membrane bilayers (12-14). This has led to increased use of lipid-containing mixed micelles, bicelles, nanodiscs, and other model membranes to better-approximate lipid bilayers than detergent-only micelles (15-20). In this paper, we explore the alternative approach of finding improved detergents for sustaining the native-like stability and function of membrane proteins, without resorting to lipid-containing media.

E. coli diacylglycerol kinase (DAGK) is well-suited for studies designed to identify optimal detergents. DAGK is a homotrimeric membrane enzyme with 9 transmembrane helices and three active sites per trimer that catalyzes direct phosphoryl transfer from MgATP to diacylglycerol to produce phosphatidic acid. In pioneering early work, the labs of Kennedy, Bell, and Sandermann showed that DAGK does not exhibit significant catalytic activity in micelles formed by common detergents unless lipid is added (21-25).

¹ This supplement has been published in: Koehler, J., et al., *Lyso-phospholipid Micelles Sustain the Stability and Catalytic Activity of Diacylglycerol Kinase in the Absence of Lipids*. *Biochemistry*, 2010. **49**(33): p. 7089-7099.

These early studies suggested that lipids play a cofactor role in support of DAGK catalysis. However, the range of commercially available detergents has dramatically expanded since those studies were carried out. The structure of DAGK was recently determined in DPC micelles using NMR spectroscopy (26), conditions in which DAGK retains considerable catalytic activity, but only at very high substrate concentrations as a consequence of dramatically elevated substrate K_m . This latter fact prevents structural studies of DAGK in DPC micelles under conditions in which it is saturated with its substrates or products. Here we re-explore detergent space to see if surfactants are now available that can sustain native-like DAGK structure, stability, and catalysis. It is shown that certain C_{14} chain detergents are able to do so, with the lyso-phospholipids proving especially effective.

Materials and Methods

Detergents and lipids used in this study were purchased from Anatrace (Maumee, OH), Avanti (Alabaster, AL), Sigma (St. Louis, MO), or Calbiochem (San Diego, CA). The diacylglycerols dibutyrylglycerol (DBG) and dihexanoylglycerol (DHG) were synthesized in-house as described previously (3).

Expression and Purification of DAGK

The gene that encodes N-terminal His₆-tagged wild type *E. coli* DAGK was ligated into the pSD005 plasmid (3;27), which was then transformed into *E. coli* WH1061 cells. WH1061 is a leucine auxotroph strain that does not express endogenous DAGK (28). DAGK was expressed in isotopically labeled form and then purified to the point where it is a pure protein attached to Ni(II)-chelate resin bathed in a buffer containing 1.5% (v/v) Empigen BB detergent (Sigma, St. Louis, MO), 40 mM HEPES, 300 mM

NaCl, and 40 mM imidazole pH 7.5 essentially as described elsewhere(26;29;30). Empigen BB was then exchanged out for the detergent of interest (e.g., LMPC) by passing 10 column volumes of 25 mM sodium phosphate buffer (pH 7.2) containing the test detergent through the column. Finally, DAGK was eluted from the column with 250 mM imidazole solution (pH 7.8) containing the same test detergent. The amount of DAGK in the elution fractions was determined spectrophotometrically based on an extinction coefficient of $2.18 \text{ (mg/ml)}^{-1} \text{ cm}^{-1}$ at 280 nm.

Measurement of DAGK Activity.

The activity assay is derived from protocols that have been described previously (3;31) whereby DAGK-catalyzed phosphoryl transfer from MgATP to DAG is coupled to NADH oxidation by pyruvate kinase (PK) and lactate dehydrogenase (LDH, Sigma) at 30 °C. The relatively short-chained dibutyrylglycerol (DBG) and dihexanoylglycerol (DHG) were the forms of diacylglycerol used in these studies because of their reasonably high solubility in detergent solutions (3;32). The pH 6.9 activity assay mix was composed of 75 mM PIPES, 50 mM LiCl, 0.1 mM EGTA, 0.1 mM EDTA 1 mM phosphoenolpyruvate (Sigma), 3 mM MgATP (Sigma), 0.25 mM NADH (Sigma), 12 mM magnesium acetate and 7.8 mM DBG. For the standard mixed micellar assay, this mixture also contained the detergent DM (at 21 mM—19 mM of which is micellar) and the lipid cardiolipin (CL, from beef heart, at 0.66 mM—which corresponds to 3 mol%). For other assays, DM and CL were replaced with the detergent of interest. DAGK stocks were prepared by diluting the purified protein to a concentration 0.15 mg/ml using detergent-containing elution buffer. Aliquots of this stock were added to the activity assay mix that had been equilibrated with PK and LDH (14 units and 20 units, respectively, per ml of mix). The decrease in absorbance at 340 nm resulting from NADH oxidation (as coupled to the DAGK reaction) was monitored spectrophotometrically, with the slope being converted to

units of DAGK activity (1 U = 1 micromole of DAG phosphorylated per minute) using the extinction coefficient for NADH of $6110 \text{ M}^{-1} \text{ cm}^{-1}$.

Activity data for determination of steady-state kinetic parameters V_{max} and K_m were collected using the same methods described above, with the exception that in each analysis the concentration of one substrate was varied (0-8 mM MgATP or 0-25 mM DBG) while the other substrate was held constant at a near-saturating level (20 mM for DBG and 3 mM for MgATP). The measured rates were plotted as a function of variable substrate concentrations and fit by the Michaelis-Menten equation (with a Hill coefficient being applied to the variable substrate concentration) using the Solver module in Microsoft Excel.

Thermal Stability of DAGK.

Purified DAGK was diluted to a concentration of 0.1 mg/ml using elution buffer plus detergent at either pH 7.8 or pH 6.5. Samples were incubated at 45 and 70 °C, and aliquots were withdrawn at various time points, rapidly frozen in liquid N₂, and then stored at -80 °C. Samples were later thawed and subjected to the standard DM/CL/DHG mixed micellar DAGK activity assay to determine the levels of remaining DAGK activity.

Circular Dichroism Spectroscopy.

Samples for CD spectroscopy were prepared by removing imidazole from purified DAGK using a PD-10 desalting column (GE Healthcare) equilibrated with buffer containing 100 mM sodium chloride, 20 mM sodium phosphate pH 6.5, and the test detergent of interest. For acquisition of CD spectra, DAGK was diluted using desalting buffer to 50-60 micromolar for near-UV CD spectroscopy or to 10-12 micromolar for far-UV CD spectroscopy.

CD experiments were carried out using a Jasco J-810 instrument equipped with a Peltier temperature control, and the sample were placed in either 1 cm (near-UV) or 0.1 cm (far-UV) path length quartz cuvettes. CD spectra were acquired at 5 °C increments between 20-80 °C, with 1 min of equilibration prior to each acquisition. The far-UV CD spectra were acquired between 190-260 nm with 1 nm bandwidth, while the near-UV spectra were acquired from 250-350 nm with 1 nm bandwidth. Baseline spectra were acquired for the protein-free desalting buffers, and subtracted from the spectra of protein-containing samples. For all acquisitions three spectra were collected and averaged to give the final trace.

The K2D algorithm (<http://www.embl.de/~andrade/k2d.html>) was used to calculate secondary structure from far-UV CD spectra.

NMR Spectroscopy of DAGK.

¹⁵N-labeled DAGK was purified using the protocol described above. When the protein was purified into LMPC and DPC, D₂O and EDTA were added to 10% and 0.5 mM, respectively, and the sample was concentrated using centrifugal ultrafiltration (Millipore Ultracel, 10 ml, 10kDa cutoff) and the sample was transferred to an NMR tube. For the protein in TDPC and LMPG the pH 7.8 purification buffer containing 250 mM imidazole (pH 7.8) was exchanged for a pH 6.5 10 mM Bis-Tris buffer by repeated centrifugal ultrafiltration/re-dilution cycles. The completeness of the exchange was monitored by checking the pH of the filtrate. EDTA and D₂O were added to all samples to final concentration of 0.5 mM and 10% (v/v). For DAGK in TDPC and LMPG magnesium chloride was also added to 2 mM.

2D ¹H,¹⁵N-TROSY NMR spectra (33) were acquired at 45°C using a Bruker 800 MHz Avance spectrometer equipped with a triple resonance cryoprobe. The Weigelt version of the TROSY experiment was used(34). Data were processed using

NMRPipe/NMRDraw software (35) and analyzed using SPARKY 3 (T.D. Goddard and D. N. Kneller, University of California, San Francisco). 3-D ^1H - ^{15}N NOESY-TROSY(36-38) spectra were acquired at 800 MHz and 45°C. For DAGK dissolved in LMPC micelles, the spectrum was acquired with 99 complex points and an acquisition time of 8.87 msec in the indirect ^1H dimension, 24 complex points and an acquisition time of 9.25 msec in the ^{15}N dimension, and 1024 complex points and an acquisition time of 91.8 msec in the ^1H observe dimension. 24 scans were acquired for each increment. The mixing time and the delay for relaxation between scans were 150 msec and 1.1 sec, respectively. For DAGK in DPC micelles, the spectrum was acquired using a slightly different version of the same pulse program (based on a different version of the TROSY(34)) with 128 complex points and an acquisition time of 16.13 msec in the indirect ^1H dimension, 64 complex points and an acquisition time of 26.88 msec in the ^{15}N dimension, and 1024 complex points and an acquisition time of 91.8 msec in the ^1H observe dimension with 8 scans. The mixing time and the delay for relaxation between scans were 100 msec and 1.3 sec, respectively.

Results

C14-Based Detergents Show Promise for Biochemical Studies of DAGK.

While it has been shown that the activity of purified DAGK is generally low in a variety of lipid-free micelles (21;22;25;32;39), some data has suggested that DAGK is more active in longer chain detergents relative to shorter chain detergents (40). We therefore screened for detergents that are able to sustain DAGK's activity even in the absence of added lipids, with a particular emphasis on detergents that are lipid-like in terms of having relatively long C_{14} alkyl chains. Detergents tested included nonionic, ionic, zwitterionic, lyso-phospholipids, and sterol-based detergents, each of which was first verified not to hinder the DAGK assay reaction coupling system. These assays

were initially carried out by adding small aliquots of DAGK stock solutions prepared in DM micelles (to far below the DM's CMC) into assay mixtures containing the test detergent at concentrations well above the test detergent's CMC. Results for this screen are given in Table 1. DM/CL detergent/lipid mixed micelles that are known to sustain native-like DAGK activity (3) were used as a positive control for this screen. It was observed that when solubilized in the C₁₄-based TDPC and lyso-phospholipids (LMPG and LMPC), DAGK exhibited activity that matched or exceeded its DM/CL control activity, whereas in all other cases the activity was much lower reflecting either a very low V_{max} for catalysis and/or of grossly elevated substrate K_m. The C₁₄ chain detergents ASB-14 and Z3-14 failed to support catalysis, indicated that having a C14 chain is not the only factor that determines detergent efficacy.

The fact that TDPC, LMPC, and LMPG support considerable DAGK activity appears to depend in part on their C₁₄ chains. Activities were measured for DAGK prepared in the corresponding C₁₂- and C₁₆-based compounds as summarized in the final column of Table 1. For these tests, DAGK was directly purified into each of the detergents and then assayed in a mixture containing the same detergent. For all three classes of detergents, the C₁₄ compound yields the highest activity within each class, with the lyso-PC compounds exhibiting higher activities than the corresponding alkyl-PC or lyso-PG compounds. Together these data suggest that significant DAGK activity can be supported by detergents that have both C₁₄ chains and suitable headgroups.

Table 1. Activity levels of DAGK assayed in various detergent conditions

| Detergent | Detergent Class | Concentrations (% w/v) | DAGK activity when small aliquots of DM/DAGK stock | DAGK activity when assays were initiated with DAGK stocks in the same detergent as used in assay. (U/mg) |
|-----------|-----------------|------------------------|--|--|
|-----------|-----------------|------------------------|--|--|

| | | | solutions were used to initiate assay. (U/mg) | |
|--------------------|---------------------|-----|--|---------|
| LLPC | lyso-PC | 0.5 | ND | 18 ± 6 |
| LMPC | lyso-PC | 0.2 | 66 | 83 ± 7 |
| LPPC | lyso-PC | 0.2 | ND | 44 ± 2 |
| LMPG | lyso-PG | 0.2 | 22 | 19 ± 3 |
| LPPG | lyso-PG | 0.2 | ND | 6.6 ± 5 |
| DPC | alkylphosphocholine | 0.5 | 0.1 | 0.3 |
| TDPC | alkylphosphocholine | 0.2 | 28 | 10 ± 1 |
| CYF7 | alkylphosphocholine | 0.5 | ~0 | ND |
| Z3-14 | zwitterionic | 0.2 | ~0 | ND |
| ASB-14 | zwitterionic | 0.2 | 0.6 | ND |
| LS | anionic | 2.0 | ~0 | ND |
| DTAB | cationic | 0.5 | ~0 | ND |
| DM | non-ionic | 0.5 | 0.3 | ND |
| GRA | saponin | 2.0 | ~0 | ND |
| DM/CL ^b | mixed micelles | | 16 | ND |
| DMPC ^c | lipid vesicles | | 63 | ND |
| POPC ^c | lipid vesicles | | 52 | ND |

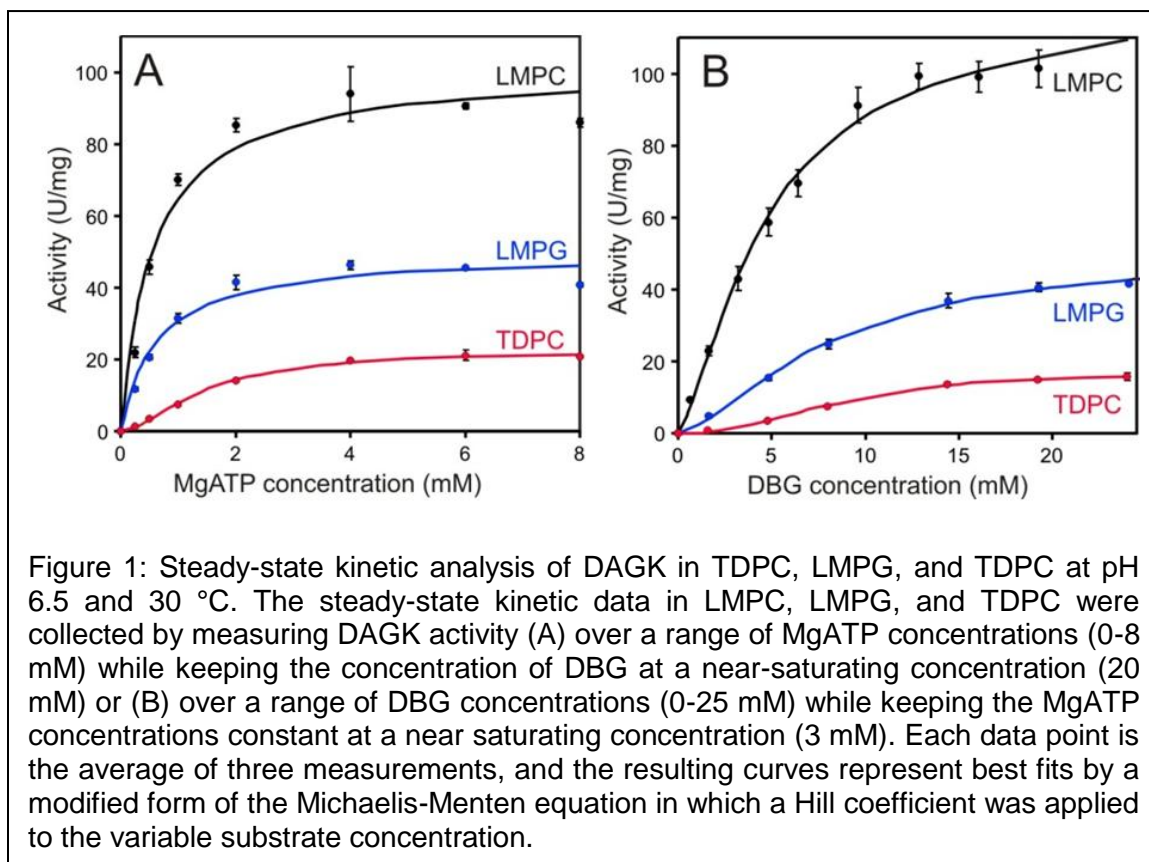
^a ND: not determined

^b While the activity of DAGK in DM/CL (ideal mixed micelles for DAGK) can reach 100 U/mg when saturating DHG is used as the lipid (diacylglycerol) substrate, the conditions used for obtaining the data of this table involved the use of DBG as the lipid substrate at a concentration that is sub-saturating even for DAGK in DM/CL. This was to avoid problems with solubility and stability for some detergents and assay components that can result from the high concentrations of diacylglycerol. Thus, the observed <100 U/mg activity under DM/CL conditions reflects the fact that DAGK is not saturated with its DAG substrate under these conditions.

LMPC Micelles Yielded the Most Favorable Steady-State Kinetic Parameters.

We determined V_{\max} and K_m for DAGK and its substrates diacylglycerol and MgATP in LMPC, LMPG, and TDPC. Rates were measured at varying concentrations of one substrate while the other substrate was maintained at a near-saturating concentration. The C4-chain DBG was used as the diacylglycerol substrate because it can be employed at high (saturating) concentrations unlike longer-chain forms of DAG, which tend either to form oil droplets or to induce precipitation of assay components before saturating concentrations can be reached.

Figure 1 shows the kinetic data for DAGK in LMPC, LMPG, and TDPC micelles. In each case, the data is slightly sigmoidal, exhibiting a slight lag phase at low substrate concentrations suggesting a modest deviation of DAGK from ideal Michaelis-Menten behavior. The origin of this phenomenon could be related to the fact that DAGK is



homotrimeric, with each of its 3 active sites being shared between subunits, although this is not the only possible explanation. Because of this modest apparent cooperativity, we applied a Hill coefficient to fit the Michaelis-Menten equation to the data, with results given in Table 2. For both substrates, the Hill coefficients determined in all cases indicate positive cooperativity. For LMPG and LMPC the V_{\max} determined when MgATP was varied (at fixed DBG concentration) was slightly less than when the concentration of DBG was varied (with fixed MgATP). This reflects the fact that the fixed concentration of MgATP was closer to saturation than was fixed DBG. This was not the case for TDPC because K_m for MgATP is elevated 3-fold in that detergent.

Table 2. Steady-state kinetic parameters for DAGK catalysis in various detergents.

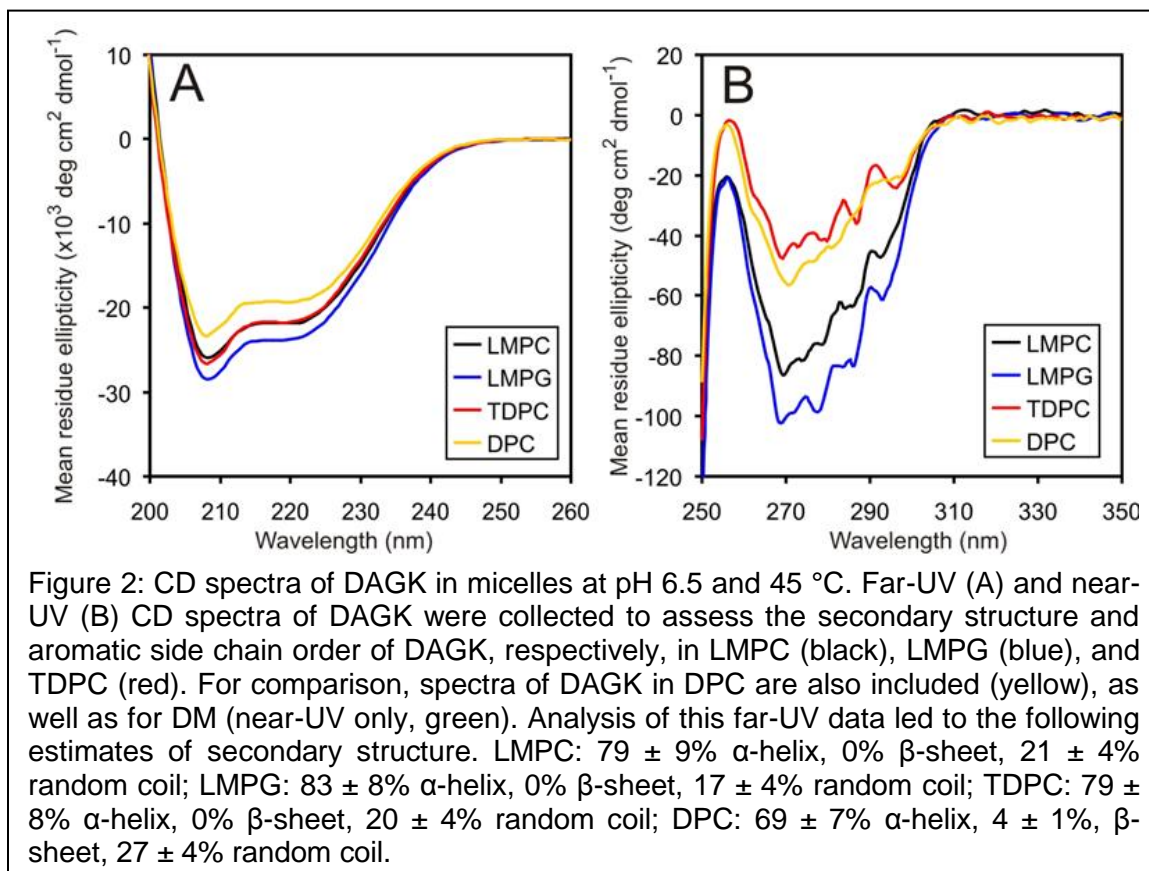
| Detergent | DBG varied | | | MgATP varied | | |
|--------------------|--------------------------|---------------------|---------------------|--------------------------|---------------------|---------------------|
| | $V_{\max,DBG}$ (U/mg) | $K_{m,DBG}$ (mM) | Hill _{DBG} | $V_{\max,ATP}$ (U/mg) | $K_{m,ATP}$ (mM) | Hill _{ATP} |
| LMPC | 119 ± 6 | 4.7 ± 0.2 | 1.4 | 91 ± 5 | 0.5 ± 0.02 | 1.8 |
| LMPG | 50 ± 3 | 7.9 ± 0.4 | 1.6 | 46 ± 3 | 0.5 ± 0.02 | 1.5 |
| TDPC | 17 ± 1 | 8.8 ± 0.4 | 2.4 | 23 ± 2 | 1.4 ± 0.07 | 1.7 |
| DM/CL ^a | ND | ND | ND | >61 ± 8 ^a | 0.6 ± 0.1 | 1.0 |

^aFrom (39). This data was collected at fixed DBG = 10 mM, which most likely was not saturating. Therefore this V_{\max} be regarded as a lower limit to the true V_{\max} when both substrate concentrations are saturating.

The kinetic parameters of Table 2 confirm that DAGK is most catalytically robust in LMPC micelles. The values for the apparent V_{\max} and K_m are comparable to those observed for DAGK under ideal mixed micellar conditions (27;31;41), which are similar to those seen for DAGK in vesicles(3;42;43). DAGK's catalytic properties in both LMPG and TDPC are less ideal than in LMPC, but are still impressive compared.

Circular Dichroism of DAGK in LMPC, LMPG, and TDPC micelles.

The secondary structure and aromatic side chain order of DAGK in LMPC, LMPG, and TDPC was probed using far- and near-UV CD spectroscopy, respectively. Data was collected at 45 °C and pH 6.5, which match the conditions used for NMR-based structural determination of DAGK (26). The far-UV CD spectra of DAGK in LMPC, LMPG, and TDPC show the strong negative bands at 208 and 222 nm characteristic of alpha-helical proteins and are very similar to the spectrum of DAGK in DPC (Figure 2A). From these spectra the percent alpha-helical content of DAGK in each of these detergents was calculated to be 79% (LMPC), 83% (LMPG), 79% (TDPC), and 69% (DPC). These values are all close to or within error of the 78% alpha-helicity observed in the NMR-determined structure of DAGK(26).



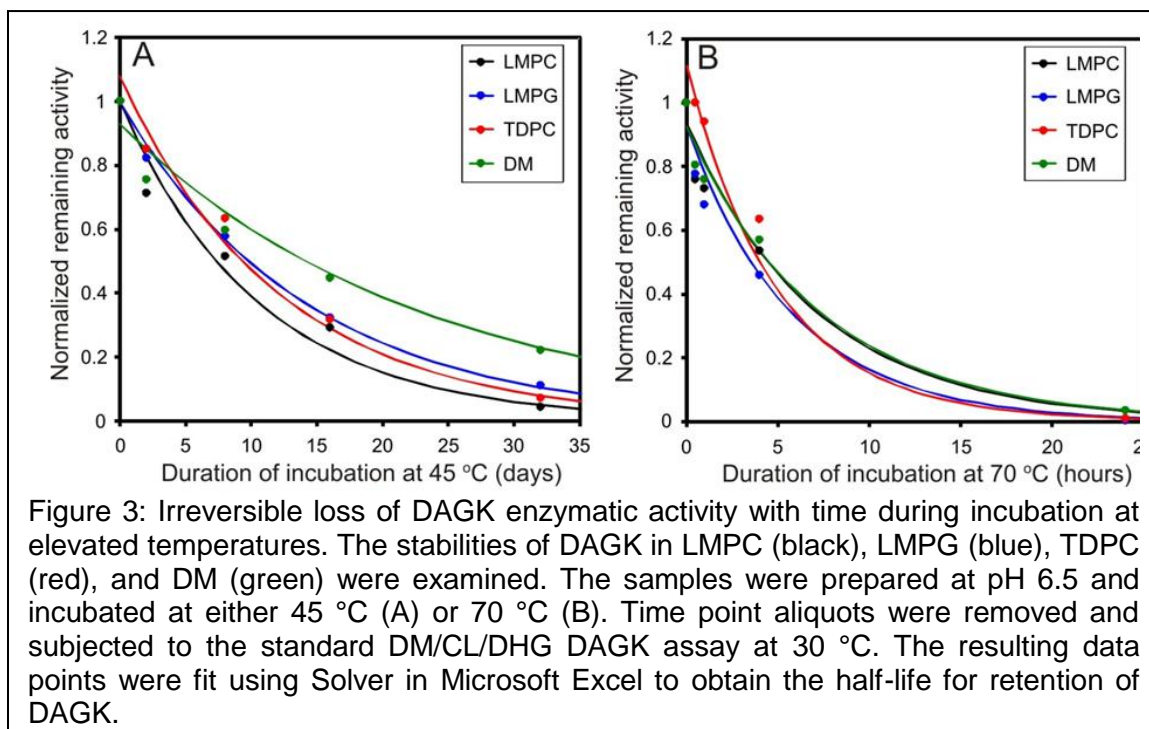
Near-UV CD spectroscopy provides information on the degree of structural order of aromatic side chains (44). It has previously been shown that when DAGK is unfolded, it exhibits no near-UV CD signal (45). As indicated in Figure 2B and in previous work (45), near-UV CD spectra of folded DAGK exhibit negative intensities indicating significant side chain order at least some of its 5 Trp, 3 Phe, and 2 Tyr residues. Since most of DAGK's aromatic residues are believed to be located at or near the water-micelle interface rather than being either deeply buried or fully water-exposed, the near-UV CD spectra mostly report on side chain structural order at or near the water-micelle interface.

While the shapes of the spectra in Figure 2B are similar from detergent to detergent, the intensities vary dramatically, being much more intense for DAGK in lysophospholipids than in the alkylphosphocholine detergents. A reasonable interpretation of the data of Figure 2B is that the aromatic side chains of DAGK in the alkylphosphocholines generally exhibit a lower degree of structural order than in lysophospholipids, an observation that correlates with the relatively high activities observed for DAGK in the latter class of detergents. However, this correlation only holds within the structurally similar alkylphosphocholine and lyso-phospholipid series, as DAGK's spectrum in DM is similar in intensity to that in LMPC, even though DAGK's activity in DM is low (<1 U/mg).

Thermal Stability of DAGK in LMPC, LMPG, and TDPC Micelles.

Wild type DAGK's thermal stability was assessed by measuring its half-life for irreversible inactivation at elevated temperatures, a process previously examined in detail by Bowie and coworkers (27;41;46). Samples at pH 6.5 and 7.8 were incubated at 45 °C and at 70 °C. The data of Figure 3 led to the reported $t_{1/2}$ for activity loss

presented in Table 3, which show that DAGK is much more stable at pH 6.5 than at pH 7.8.



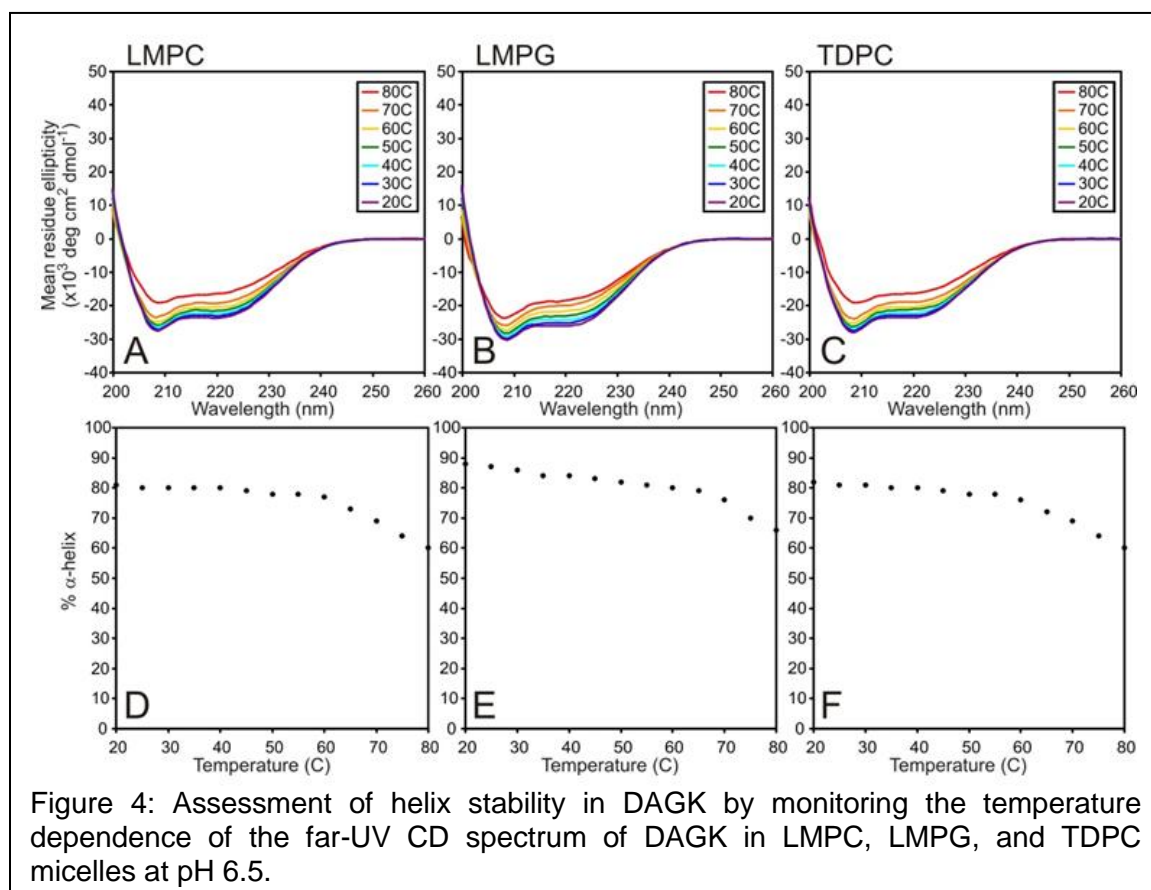
DM micelles have been reported to maintain DAGK in a highly stable state (27;47) even though the enzyme exhibits only very low activity in this detergent. At pH 6.5 DAGK's stability in LMPG, LMPC, and TDPC is comparable to its stability in DM, with $time_{1/2}$ ranging from several hours at 70 °C to 8-13 days at 45 °C (Table 3). That DAGK is not significantly more resistant to heat inactivation in the new detergent systems than in DM despite being more active in the new systems shows that there is no correlation within this set of detergents between enzyme activity and stability.

Table 3. Thermal stability of DAGK in different micelles.

| Detergent | Time _{1/2} for irreversible loss of DAGK activity. | | | |
|-----------|---|------------------|-----------------|------------------|
| | pH 7.8 | | pH 6.5 | |
| | 45 °C (days) | 70 °C (hours) | 45 °C (days) | 70 °C (hours) |
| LMPC | 1.3 ± 0.1 | 0.3 ± 0.02 | 8.0 ± 0.4 | 4.6 ± 0.3 |

| | | | | |
|------|----------------|----------------|----------------|---------------|
| LMPG | 0.1 ± 0.01 | 0.1 ± 0.01 | 9.8 ± 0.5 | 3.5 ± 0.2 |
| TDPC | 0.7 ± 0.1 | 0.4 ± 0.02 | 10.1 ± 0.5 | 5.0 ± 0.3 |
| DM | 7.7 ± 0.4 | 0.7 ± 0.04 | 12.6 ± 0.7 | 5.1 ± 0.3 |

The temperature dependency of DAGK's far-UV CD spectra is shown in Figure 4, where only a minor loss of helicity is observed for LMPC, LMPG, and TDPC as the



temperature is raised from 20 to 60°C. Only above 60°C does the helicity begin to drop steeply. In each case, when the temperature reaches 80 °C, DAGK has lost 18-23% of its alpha-helical content (Table 4). Most likely, this loss is due to melting of the N-terminal amphipathic helix, which encompasses about 20% of DAGK's helical content at 45°C (26).

For LMPC and TDPC the loss of helicity observed in DAGK upon elevating the temperature to 80 °C was partially reversible, whereas in the case of LMPG reversibility is nearly complete (Table 4) although this detergent was not superior to the others in protecting against thermal inactivation (Table 3).

Table 4. Percent alpha-helix of DAGK at pH 6.5 determined from far-UV CD spectra acquired successively at 20 °C, 80 °C, and after return to 20 °C

| Detergents | 20 °C | 80 °C | 20 °C (after heating) |
|------------|--------|--------|--------------------------|
| LMPC | 81 ± 8 | 60 ± 6 | 74 ± 7 |
| LMPG | 88 ± 8 | 66 ± 6 | 85 ± 9 |
| TDPC | 82 ± 8 | 59 ± 7 | 72 ± 7 |

Near-UV CD spectra were also acquired as a function of temperature, which provides insight into aromatic side chain order (Figure 5). In the case of TDPC the data resembles the corresponding far-UV CD data in that there is little change until ca. 60 °C, above which signal intensity is steeply reduced until it reaches baseline near 80 °C. For LMPG and LMPC there are significant reductions in the far-UV CD signal intensity as the temperature is raised, with baseline in both cases being reached by 80 °C. However, for the LMPC case reductions in signal intensity become much steeper when 60°C is exceeded. We speculate that DAGK in lyso-phospholipids below 60 °C possesses a class of side chain conformational order that is absent in TDPC even at low temperatures, and that this “missing order” is likely to be related to the lower catalytic activity of DAGK observed in TDPC relative to the lyso-phospholipids. At 60 °C, DAGK has lost this class of side chain order in all three detergents tested, but still retains a second class of side chain order. For all three detergents this remaining side chain

order is lost by the point 80 °C is reached, which likely corresponds with complete loss of stable tertiary structure.

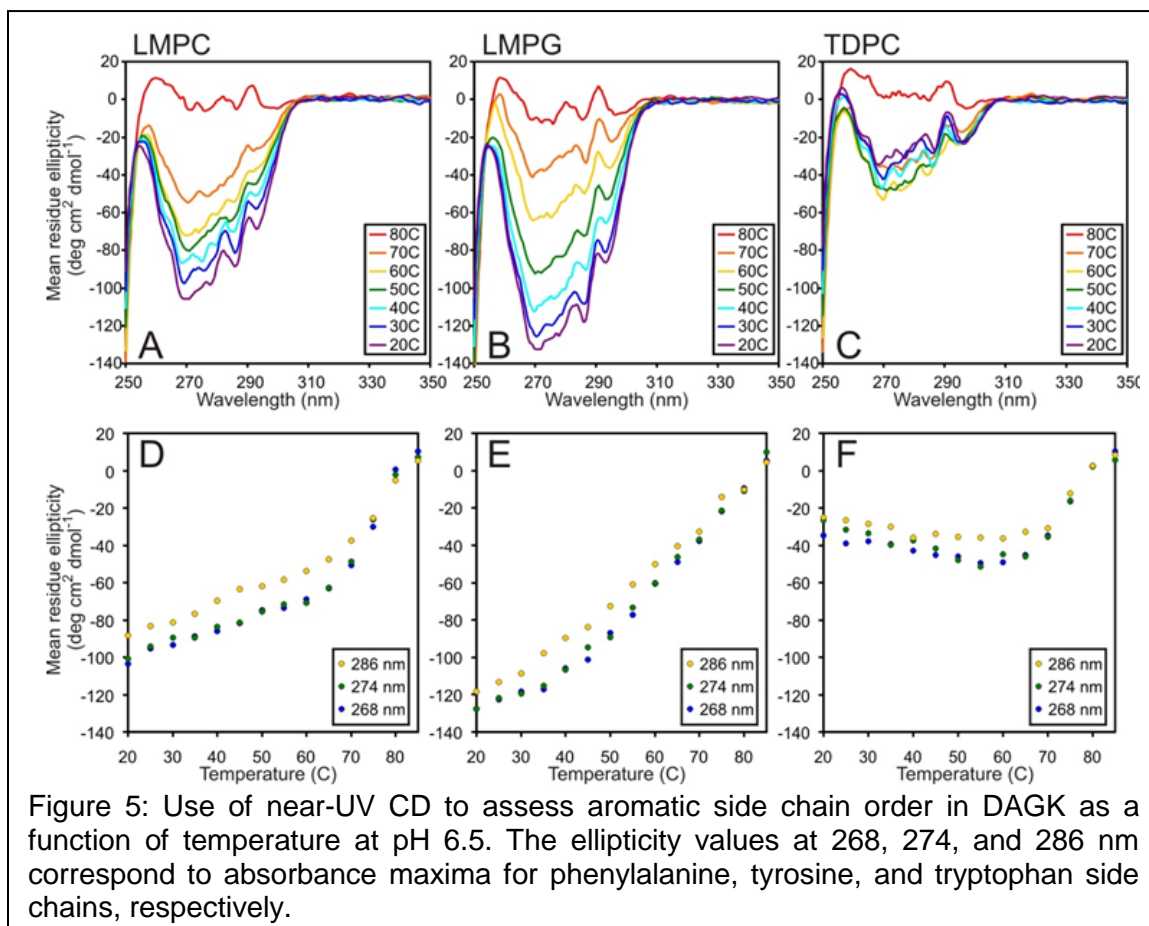


Figure 5: Use of near-UV CD to assess aromatic side chain order in DAGK as a function of temperature at pH 6.5. The ellipticity values at 268, 274, and 286 nm correspond to absorbance maxima for phenylalanine, tyrosine, and tryptophan side chains, respectively.

Unlike the case for the far-UV CD data, we found that returning the temperature to 20 °C in no case allowed DAGK to recapitulate its original near-UV CD spectrum (data not shown), indicating that the loss of tertiary structural order is not reversible. This is consistent with the previous observations of Bowie and co-workers (46).

TROSY NMR Spectra of DAGK in LMPC, LMPG, and TDPC Micelles

¹⁵N-TROSY-HSQC spectra were acquired at 45 °C for pH 6.5 samples of WT DAGK in LMPC, LMPG, TDPC, and DPC micelles (Figure 6). Overall, the spectra are generally similar and exhibit the modest spectral dispersion that is usually associated with helical membrane proteins. However, while the quality of the spectra in LMPG and

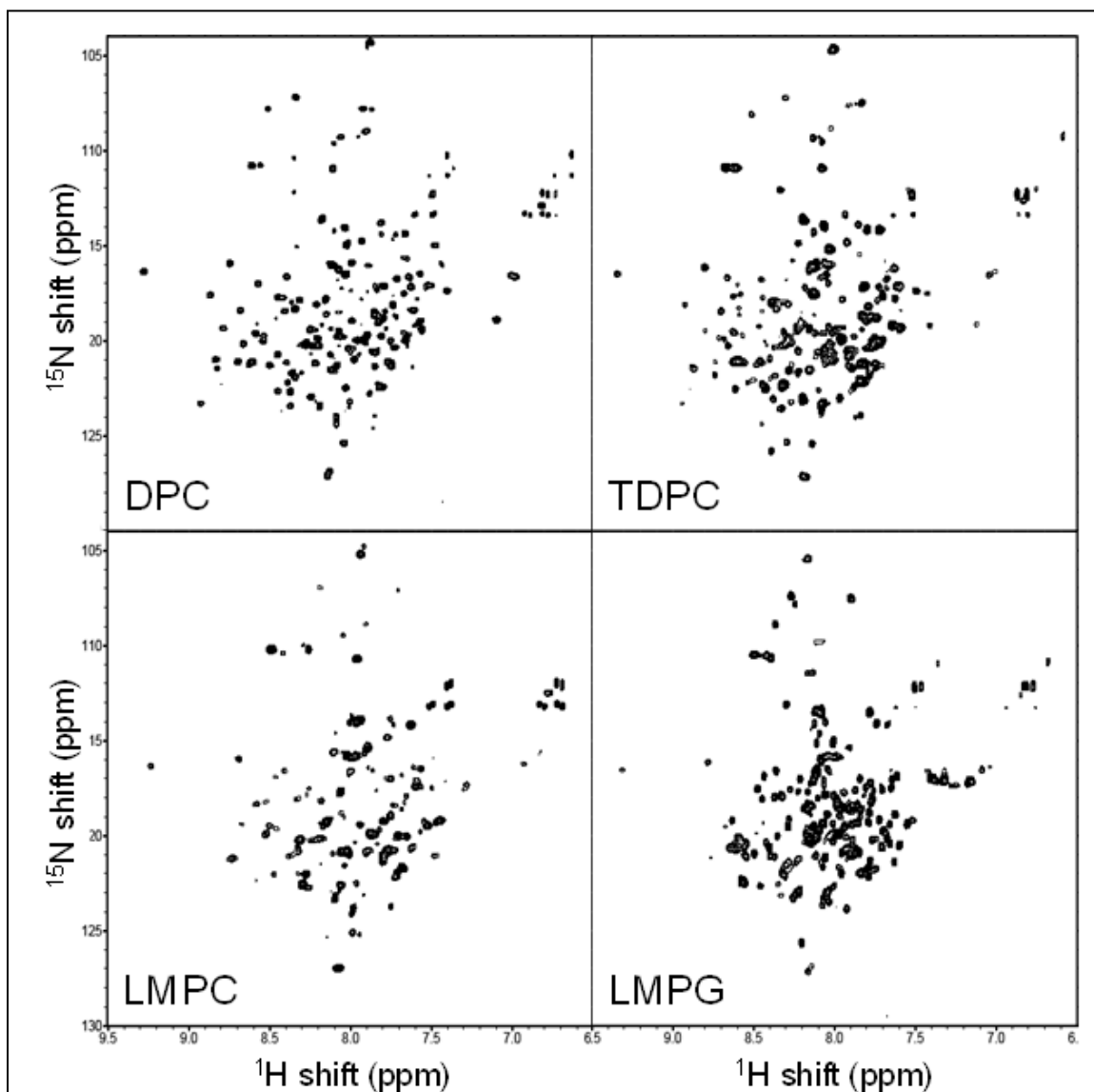


Figure 6: 800 MHz ^{15}N -TROSY spectra of DAGK in LMPC, LMPG, DPC, and TDPC micelles at 45 °C. The samples in TDPC and LMPG contained 10 mM Bis-Tris, 2 mM magnesium chloride, 0.5 mM EDTA, and 10% (v/v) D_2O , pH 6.5. The samples in DPC and LMPC contained 250 mM imidazole, 0.5 mM EDTA, and 10% (v/v) D_2O , pH 6.5.

TDPC is high—approaching the case of DPC, the LMPC spectrum exhibits fewer peaks.

This is not because there are many missing peaks; rather, it is because at the peak plotting level used (which is comparable for all 4 spectra) many LMPC peaks are broadened to the point where their maxima fall below the plotting level threshold. There are several possible sources for the linebroadening. One possibility is that DAGK-LMPC mixed micelles are larger than the corresponding DAGK-TDPC and DAGK-LMPG

micelles. Another possible contribution to line broadening includes the presence of internal conformational motions for DAGK in LMPC micelles that are intermediate on the NMR time scale, leading to exchange broadening. A final possible contributing factor is conformational microheterogeneity that results in many similar but non-identical/non-exchanging superimposed peaks. Additional experiments would be required to determine which of the above phenomena are the actual contributing factors.

The quality of the DAGK spectra from TDPC and LMPG is excellent for a 40 kDa homotrimeric multispan membrane protein as part of a much larger micellar complex. The average line-width of the peaks seen in each of these detergents is 25 Hz. The TDPC and LMPG spectra are similar but are also sufficiently *different* from each other and from the assigned DPC spectrum such that the assignments that are available for the DPC peaks (30) cannot in many cases be reliably extrapolated to the TDPC and LMPG cases. The differences in specific peak positions may be explained by the fact that the detergent/DAGK interface is quite extensive, such that variations in the covalent structure of each detergent result in modest but widespread changes in resonance position.

*NOESY NMR Data Shows Differences in DAGK-Detergent Interactions for DPC
Versus LMPC.*

3-D $^1\text{H},^{15}\text{N}$ -NOESY NMR spectra were acquired for U- ^{15}N -DAGK in both DPC and LMPC micelles. We used the “half-filtered” version of this 3-D experiment, which leads to NOE crosspeaks being observed for pairs of proximal protons for which at least one of the two interacting protons is directly attached to ^{15}N . In analyzing this data we focused on NOEs between the tryptophan side chain indole NH proton and detergent protons, observation of which can be taken as an indication of at least transient proximity (<5 angstroms) between the indole proton and protons on the detergent^{Footnote 2}.

Figure 7 shows 2-D strip plots from the ^1H , ^{15}N -NOESY data that shows the NOEs observed between the indole protons and the detergents. In the case of DPC (Figure 7A), it is seen that strong NOEs are observed between all indole NH protons and both choline methyl (3.3 PPM) and alkyl chain protons (1.4 PPM). That NOEs to these spatially distinct parts of DPC are simultaneously observed undoubtedly reflects both the high heterogeneity of detergent/membrane protein interactions (any given snapshot) and also the highly dynamic nature of these interactions. This data indicates that, on the average, the Trp sides chains of DAGK in DPC micelles spend as much time near the

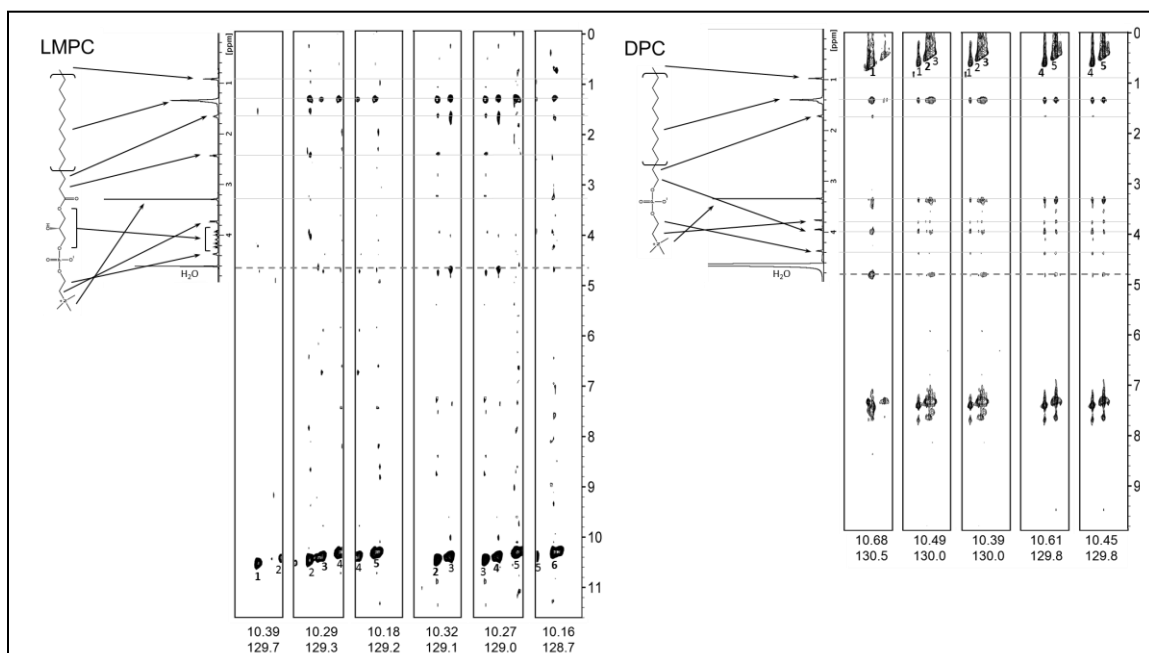


Figure 7: 2-D strips plots showing NOEs to the tryptophan indole NH protons from 3-D ^1H - ^{15}N NOESY-TROSY spectra collected for U- ^{15}N -DAGK in DPC (A) and LMPC (B) micelles at 800 MHz and 45°C. Associated with each set of strips is a 1-D NMR spectrum for pure DPC and LMPC showing resonance assignments. In the strip plots for the DPC case (A) the indole NH proton diagonal peaks appear upfield of the rest of the spectrum rather than in the expected 9.5-10.5 PPM range because the peaks are “aliased” as a result of being outside of the spectral window of the TPPI/States-based NMR experiment used to acquire this data (39). In the LMPC case (B) the peaks appear at the expected chemical shifts because they fell within the observation sweep width. In the LMPC case, side chain-perdeuterated DAGK was used, which is why NOEs between the NH indole protons to their two nearest neighbors on the indole rings are not observed at 7.2-7.8 PPM, unlike the DPC case where DAGK was not deuterated and these NOEs are quite pronounced.

charged choline headgroup as they do with the hydrophobic micelle interior. In the case of LMPC, a significantly different pattern is seen (Figure 7B). While strong NOEs are observed between protons from the acyl chain (1.4 PPM) and the indole peaks, NOEs between the indoles and the choline headgroup are weak or absent. Some NOEs are observed between protons from the glycerol backbone (3.8-4.25 PPM) and the indoles, although these are not as strong as the NOEs to the acyl chain. These results indicate that average position of the indole side chain in LMPC micelles is significantly deeper (towards the apolar micellar interior) than in the case of DPC, such that direct indole/choline interactions are largely avoided in the former case. The Trp side chains are not so deeply buried, however, that NOEs are observed between the indole NH and the terminal methyl group of the aliphatic chain in LMPC, consistent with the Trp side chains being restrained so as to avoid the center of the micelles.

Discussion

Early work on *E. coli* DAGK focused on the catalytic properties of this enzyme under conditions in which it was solubilized using Triton X-100 or alkylglycoside detergents (21-25). Under these conditions DAGK was found to require the presence of added phospholipid in order to exhibit significant catalytic activity. This led to the notion that lipids play a requisite “cofactor” role in promoting DAGK’s catalytic activity. The present work shows that the C₁₄-based detergents LMPC, LMPG, and TDPC were able to sustain specific DAGK activities of at least 10 units/mg under standard assay conditions. Of these, LMPC yielded the highest activity (66 U/mg), with the C₁₂- and C₁₆- analogs of this detergent also sustaining >10 U/mg. Indeed, the V_{max} for DAGK in LMPC, LMPG, and TDPC micelles was seen to be roughly 100, 50, and 20 U/mg respectively, with the 100 U/mg observed for DAGK in LMPC being roughly the same as

the highest activities previously observed for the enzyme in mixed micellar, bicellar, or vesicular conditions (3;31;42;43;48).

There appear to be three key factors that promote DAGK activity. First, the C₁₄ chain was found to be superior to either shorter or longer chains. Most likely, the diameter of micelles containing C₁₄ chains is an optimal match for the hydrophobic span of the transmembrane domain of DAGK. While the diameters of LMPC and LMPG micelles have not been directly measured, from studies (49-52) of C₁₆-lyso-PG and C₈₋₁₂-lyso-PC, it is possible to estimate that the span of the hydrophobic domain of LMPG and LMPC micelles is in the range of 30-35 angstroms, with the thickness of the glycerol backbone/headgroup domain being roughly 10-12 angstroms on each side. Lee and co-workers have examined DAGK's activity in a series of lipid vesicles composed of phosphatidylcholine with two mono-unsaturated chains and found that the enzyme is most active in di(C_{18:1})-PC vesicles (43), which have a hydrophobic span of 30 angstroms, roughly the same as estimated for LMPC and LMPG micelles. This appears to be a good match the observed hydrophobic span of the experimental DAGK structure (26), which is 30-33 angstroms. This highlights the importance of appropriate matching of the transmembrane span of a membrane protein with the thickness of the membrane or membrane-mimetic in which it sits to optimize protein structure, stability and function, as others have previously described (53-57).

A second key factor in promoting activity appears to be the presence of the glycerol spacer between the acyl chain and the charged head group, as reflected by the higher activities observed in the lyso-phospholipids relative to the alkylphosphocholines. The glycerol spacer/backbone is, of course, present in the glycerophospholipids that dominate the composition of the plasma membrane of *E. coli*. It is also interesting to note that the lyso-phospholipids are the only commercially available class of single-chain ionic detergents that has a polar-but-uncharged spacer between the apolar tail and the

charged head group. Our results suggest that DAGK has evolved so as to prefer the presence of a glycerol spacer over an abrupt chain-to-head group transition. This may be a property shared by many other membrane proteins. Other recent studies have highlighted some of the advantages of working with lyso-phospholipids as detergents (58-60), which seem to be particularly effective at sustaining high membrane protein solubility without disrupting structure and function (61-64).

Finally, DAGK exhibited a preference for the zwitterionic phosphocholine head group of LMPC over the anionic head group of LMPG. This result is strikingly similar to results for DAGK activity in lipid vesicles, where Lee et al. observed DAGK to be most active in phosphatidylcholine vesicles compared to vesicles composed of phosphatidylglycerol or phosphatidylethanolamine. This is despite the fact that *E. coli* membranes are bereft of phosphatidylcholine but are rich in phosphatidylglycerol and phosphatidylethanolamine(65). Given that DAGK actually shows a preference for anionic lipids as activators when lipids are added to neutral detergent micelles (24;25), it seems likely the anionic charge density present in LMPG micelles represents “too much of a good thing” from DAGK’s standpoint. We did not test DAGK’s activity in LMPC/LMPG mixtures, but this may be interesting to examine in future work. In any case, charge alone is not the only important property of the detergent headgroup, as illustrated by the fact that the zwitterionic Z3-14 and ASB-14 did not support DAGK activity. This is possibly because the orientation of the positive and negative charge with respect to the main chain are reversed relative to the phosphocholine detergents and virtually all known natural zwitterionic phospholipids.

The most important insight regarding how LMPC, LMPG, and TDPC may promote DAGK’s catalytic activity relative to DPC is provided by NOE measurements (Figure 7). These results indicate that the Trp side chains of DAGK in both DPC and LMPC micelles interact strongly with the detergent aliphatic chains, but avoid contacts

with the chain termini, which are expected to be found primarily in the center of the micelles. This is as expected for Trp side chains based both on the structure of DAGK(26) and the observation that Trp side chains are usually found in the membrane bilayer, but fairly near the surface. However, the NOE data also show that while the Trp side chains of DAGK in LMPC interact almost exclusively with the aliphatic groups and, to a lesser extent, the glycerol spacer, the situation is very different in DPC. Namely, strong interactions of the indole rings with the choline methyl protons on the end of the head group are observed. Apparently, in the absence of the glycerol spacer that is present in both LMPC and in most lipids of native membranes the indole side chains are forced to come in frequent contact with the most polar parts of the micelle. We suggest that it is this inappropriate contact that results in the reduction in the aromatic side chain order evident in the near-UV CD spectrum of DAGK in DPC relative to LMPC conditions, as well as DAGK's lower activity and stability in DPC micelles.

Our conclusions that micelles comprised of certain C₁₄ detergents can sustain DAGK in a stable and nearly fully active form and that these detergents also lead to high quality NMR spectra may be very important for future structural studies of this enzyme. While the structure of the substrate-free form of DAGK was recently determined in DPC micelles, the K_m of the enzyme for MgATP and diacylglycerol are significantly elevated—to the point where structural studies of saturated DAGK-substrate complexes may be very difficult, particularly for complexes that include diacylglycerol. The results of this paper establish that in LMPC and LMPG the K_m for DAGK's substrates are close to their values under ideal conditions and are low enough such that structural studies of saturated binary and ternary complexes using NMR methods are now feasible. This is an important development. While DAGK has previously been shown to be fully active in lipid vesicles (3), bicelles (3), lipid-detergent mixed micelles (3;31), and even amphipols

(39), these alternative membrane-mimetic media have not yet yielded either high quality NMR spectra or well-diffracting crystals of this enzyme.

Conclusions

Great care must be exercised when working with membrane proteins in micelles to insure that potential detergent-induced perturbations of native-like structural or functional properties are taken into account. However, the success of LMPC in sustaining DAGK's high thermal stability and catalytic activity highlights the fact that for at least some integral membrane proteins the best-available detergents appear to exert only very modest perturbations. This is fortunate because some biophysical and biochemical methods are easier to carry out in detergent solutions than in more complex membrane-mimetic media such as bicelles, nanodiscs, lipidic cubic phases, or unilamellar vesicles. The fact that lyso-phospholipids appear to be particularly well-suited for DAGK appears to be closely related to the fact that they are the only class of single-chain detergents that resembles the majority of phospholipids found in nature in that they have a polar-but-uncharged (glycerol) spacer that links the apolar tail to the charged headgroup. The lyso-phospholipids may be worthy of much more widespread use in membrane protein research.

References

- [1] Kim, H. J., Howell, S. C., Van Horn, W. D., Jeon, Y. H., and Sanders, C. R. (2009) Recent Advances in the Application of Solution NMR Spectroscopy to Multi-Span Integral Membrane Proteins, *Prog. Nucl. Magn Reson. Spectrosc.* **55**, 335-360.
- [2] Sanders, C. R. and Sonnichsen, F. (2006) Solution NMR of membrane proteins: practice and challenges, *Magn Reson. Chem.* **44 Spec No**, S24-S40.
- [3] Czerski, L. and Sanders, C. R. (2000) Functionality of a membrane protein in bicelles, *Anal. Biochem.* **284**, 327-333.

- [4] Hanson, M. A., Cherezov, V., Griffith, M. T., Roth, C. B., Jaakola, V. P., Chien, E. Y., Velasquez, J., Kuhn, P., and Stevens, R. C. (2008) A specific cholesterol binding site is established by the 2.8 Å structure of the human beta2-adrenergic receptor, *Structure* 16, 897-905.
- [5] Hunte, C. and Richers, S. (2008) Lipids and membrane protein structures, *Curr. Opin. Struct. Biol.* 18, 406-411.
- [6] Qin, L., Sharpe, M. A., Garavito, R. M., and Ferguson-Miller, S. (2007) Conserved lipid-binding sites in membrane proteins: a focus on cytochrome c oxidase, *Curr. Opin. Struct. Biol.* 17, 444-450.
- [7] Reichow, S. L. and Gonen, T. (2009) Lipid-protein interactions probed by electron crystallography, *Curr. Opin. Struct. Biol.* 19, 560-565.
- [8] Choowongkamon, K., Carlin, C. R., and Sonnichsen, F. D. (2005) A structural model for the membrane-bound form of the juxtamembrane domain of the epidermal growth factor receptor, *J. Biol. Chem.* 280, 24043-24052.
- [9] Chou, J. J., Kaufman, J. D., Stahl, S. J., Wingfield, P. T., and Bax, A. (2002) Micelle-induced curvature in a water-insoluble HIV-1 Env peptide revealed by NMR dipolar coupling measurement in stretched polyacrylamide gel, *J. Am. Chem. Soc.* 124, 2450-2451.
- [10] Kang, C., Tian, C., Sonnichsen, F. D., Smith, J. A., Meiler, J., George, A. L., Jr., Vanoye, C. G., Kim, H. J., and Sanders, C. R. (2008) Structure of KCNE1 and implications for how it modulates the KCNQ1 potassium channel, *Biochemistry* 47, 7999-8006.
- [11] Lee, S. Y., Lee, A., Chen, J., and MacKinnon, R. (2005) Structure of the KvAP voltage-dependent K⁺ channel and its dependence on the lipid membrane, *Proc. Natl. Acad. Sci. U. S. A* 102, 15441-15446.
- [12] MacKenzie, K. R. and Fleming, K. G. (2008) Association energetics of membrane spanning alpha-helices, *Curr. Opin. Struct. Biol.* 18, 412-419.
- [13] Matthews, E. E., Zoonens, M., and Engelman, D. M. (2006) Dynamic helix interactions in transmembrane signaling, *Cell* 127, 447-450.
- [14] Mi, L. Z., Grey, M. J., Nishida, N., Walz, T., Lu, C., and Springer, T. A. (2008) Functional and structural stability of the epidermal growth factor receptor in detergent micelles and phospholipid nanodiscs, *Biochemistry* 47, 10314-10323.
- [15] Faham, S. and Bowie, J. U. (2002) Bicelle crystallization: a new method for crystallizing membrane proteins yields a monomeric bacteriorhodopsin structure, *J. Mol. Biol.* 316, 1-6.

- [16] Nath, A., Atkins, W. M., and Sligar, S. G. (2007) Applications of phospholipid bilayer nanodiscs in the study of membranes and membrane proteins, *Biochemistry* 46, 2059-2069.
- [17] Popot, J. L. (2010) Amphipols, Nanodiscs, and Fluorinated Surfactants: Three Nonconventional Approaches to Studying Membrane Proteins in Aqueous Solutions, *Annu. Rev. Biochem.*
- [18] Prive, G. G. (2009) Lipopeptide detergents for membrane protein studies, *Curr. Opin. Struct. Biol.* 19, 379-385.
- [19] Prosser, R. S., Evanics, F., Kitevski, J. L., and Al-Abdul-Wahid, M. S. (2006) Current applications of bicelles in NMR studies of membrane-associated amphiphiles and proteins, *Biochemistry* 45, 8453-8465.
- [20] Sanders, C. R., Kuhn, H. A., Gray, D. N., Keyes, M. H., and Ellis, C. D. (2004) French swimwear for membrane proteins, *ChemBiochem.* 5, 423-426.
- [21] Bohnenberger, E. and Sandermann, H., Jr. (1983) Lipid dependence of diacylglycerol kinase from *Escherichia coli*, *Eur. J. Biochem.* 132, 645-650.
- [22] Russ, E., Kaiser, U., and Sandermann, H., Jr. (1988) Lipid-dependent membrane enzymes. Purification to homogeneity and further characterization of diacylglycerol kinase from *Escherichia coli*, *Eur. J. Biochem.* 171, 335-342.
- [23] Schneider, E. G. and Kennedy, E. P. (1976) Partial purification and properties of diglyceride kinase from *Escherichia coli*, *Biochim. Biophys. Acta* 441, 201-212.
- [24] Walsh, J. P. and Bell, R. M. (1986) sn-1,2-Diacylglycerol kinase of *Escherichia coli*. Structural and kinetic analysis of the lipid cofactor dependence, *J. Biol. Chem.* 261, 15062-15069.
- [25] Walsh, J. P. and Bell, R. M. (1986) sn-1,2-Diacylglycerol kinase of *Escherichia coli*. Mixed micellar analysis of the phospholipid cofactor requirement and divalent cation dependence, *J. Biol. Chem.* 261, 6239-6247.
- [26] Van Horn, W. D., Kim, H. J., Ellis, C. D., Hadziselimovic, A., Sulistijo, E. S., Karra, M. D., Tian, C., Sonnichsen, F. D., and Sanders, C. R. (2009) Solution nuclear magnetic resonance structure of membrane-integral diacylglycerol kinase, *Science* 324, 1726-1729.
- [27] Zhou, Y. and Bowie, J. U. (2000) Building a thermostable membrane protein, *J. Biol. Chem.* 275, 6975-6979.

- [28] Miller, K. J., McKinstry, M. W., Hunt, W. P., and Nixon, B. T. (1992) Identification of the diacylglycerol kinase structural gene of *Rhizobium meliloti* 1021, *Mol. Plant Microbe Interact.* 5, 363-371.
- [29] Oxenoid, K., Sonnichsen, F. D., and Sanders, C. R. (2002) Topology and secondary structure of the N-terminal domain of diacylglycerol kinase, *Biochemistry* 41, 12876-12882.
- [30] Oxenoid, K., Kim, H. J., Jacob, J., Sonnichsen, F. D., and Sanders, C. R. (2004) NMR assignments for a helical 40 kDa membrane protein, *J. Am. Chem. Soc.* 126, 5048-5049.
- [31] Badola, P. and Sanders, C. R. (1997) *Escherichia coli* diacylglycerol kinase is an evolutionarily optimized membrane enzyme and catalyzes direct phosphoryl transfer, *J. Biol. Chem.* 272, 24176-24182.
- [32] Walsh, J. P., Fahrner, L., and Bell, R. M. (1990) sn-1,2-diacylglycerol kinase of *Escherichia coli*. Diacylglycerol analogues define specificity and mechanism, *J. Biol. Chem.* 265, 4374-4381.
- [33] Riek, R., Pervushin, K., and Wuthrich, K. (2000) TROSY and CRINEPT: NMR with large molecular and supramolecular structures in solution, *Trends Biochem. Sci.* 25, 462-468.
- [34] Weigelt, J. (1998) Single scan, sensitivity- and gradient-enhanced TROSY for multidimensional NMR experiments, *Journal of the American Chemical Society* 120, 10778-10779.
- [35] Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes, *J. Biomol. NMR* 6, 277-293.
- [36] Czisch, M. and Boelens, R. (1998) Sensitivity enhancement in the TROSY experiment, *J. Magn Reson.* 134, 158-160.
- [37] Salzmann, M., Pervushin, K., Wider, G., Senn, H., and Wuthrich, K. (1998) TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins, *Proc. Natl. Acad. Sci. U. S. A* 95, 13585-13590.
- [38] Zhu, G., Kong, X. M., and Sze, K. H. (1999) Gradient and sensitivity enhancement of 2D TROSY with water flip-back, 3D NOESY-TROSY and TOCSY-TROSY experiments, *Journal of Biomolecular Nmr* 13, 77-81.
- [39] Gorzelle, B. M., Hoffman, A. K., Keyes, M. H., Gray, D. N., Ray, D. G., and Sanders, C. R. (2002) Amphipols can support the activity of a membrane enzyme, *J. Am. Chem. Soc.* 124, 11594-11595.

- [40] Vinogradova, O., Sonnichsen, F., and Sanders, C. R. (1998) On choosing a detergent for solution NMR studies of membrane proteins, *J. Biomol. NMR* 11, 381-386.
- [41] Lau, F. W., Nauli, S., Zhou, Y., and Bowie, J. U. (1999) Changing single side-chains can greatly enhance the resistance of a membrane protein to irreversible inactivation, *J. Mol. Biol.* 290, 559-564.
- [42] Pilot, J. D., East, J. M., and Lee, A. G. (2001) Effects of phospholipid headgroup and phase on the activity of diacylglycerol kinase of *Escherichia coli*, *Biochemistry* 40, 14891-14897.
- [43] Pilot, J. D., East, J. M., and Lee, A. G. (2001) Effects of bilayer thickness on the activity of diacylglycerol kinase of *Escherichia coli*, *Biochemistry* 40, 8188-8195.
- [44] Rogers, D. M. and Hirst, J. D. (2004) First-principles calculations of protein circular dichroism in the near ultraviolet, *Biochemistry* 43, 11092-11102.
- [45] Nagy, J. K., Lonzer, W. L., and Sanders, C. R. (2001) Kinetic study of folding and misfolding of diacylglycerol kinase in model membranes, *Biochemistry* 40, 8971-8980.
- [46] Zhou, Y., Lau, F. W., Nauli, S., Yang, D., and Bowie, J. U. (2001) Inactivation mechanism of the membrane protein diacylglycerol kinase in detergent solution, *Protein Sci.* 10, 378-383.
- [47] Li, Q., Mittal, R., Huang, L., Travis, B., and Sanders, C. R. (2009) Bolaamphiphile-class surfactants can stabilize and support the function of solubilized integral membrane proteins, *Biochemistry* 48, 11606-11608.
- [48] Lau, F. W., Chen, X., and Bowie, J. U. (1999) Active sites of diacylglycerol kinase from *Escherichia coli* are shared between subunits, *Biochemistry* 38, 5521-5527.
- [49] Chou, J. J., Baber, J. L., and Bax, A. (2004) Characterization of phospholipid mixed micelles by translational diffusion, *J. Biomol. NMR* 29, 299-308.
- [50] Lipfert, J., Columbus, L., Chu, V. B., Lesley, S. A., and Doniach, S. (2007) Size and shape of detergent micelles determined by small-angle X-ray scattering, *J. Phys. Chem. B* 111, 12427-12438.
- [51] Mendz, G. L., Jamie, I. M., and White, J. W. (1992) Effects of acyl chain length on the conformation of myelin basic protein bound to lysolipid micelles, *Biophys. Chem.* 45, 61-77.
- [52] Vitiello, G., Ciccarelli, D., Ortona, O., and D'Errico, G. (2009) Microstructural characterization of lysophosphatidylcholine micellar aggregates: the structural

- basis for their use as biomembrane mimics, *J. Colloid Interface Sci.* 336, 827-833.
- [53] Botelho, A. V., Huber, T., Sakmar, T. P., and Brown, M. F. (2006) Curvature and hydrophobic forces drive oligomerization and modulate activity of rhodopsin in membranes, *Biophys. J.* 91, 4464-4477.
- [54] Columbus, L., Lipfert, J., Jambunathan, K., Fox, D. A., Sim, A. Y., Doniach, S., and Lesley, S. A. (2009) Mixing and matching detergents for membrane protein NMR structure determination, *J. Am. Chem. Soc.* 131, 7320-7326.
- [55] Holt, A. and Killian, J. A. (2010) Orientation and dynamics of transmembrane peptides: the power of simple models, *Eur. Biophys. J.* 39, 609-621.
- [56] Lee, A. G. (2003) Lipid-protein interactions in biological membranes: a structural perspective, *Biochim. Biophys. Acta* 1612, 1-40.
- [57] Soubias, O., Niu, S. L., Mitchell, D. C., and Gawrisch, K. (2008) Lipid-rhodopsin hydrophobic mismatch alters rhodopsin helical content, *J. Am. Chem. Soc.* 130, 12465-12471.
- [58] Beel, A. J., Mobley, C. K., Kim, H. J., Tian, F., Hadziselimovic, A., Jap, B., Prestegard, J. H., and Sanders, C. R. (2008) Structural studies of the transmembrane C-terminal domain of the amyloid precursor protein (APP): does APP function as a cholesterol sensor?, *Biochemistry* 47, 9428-9446.
- [59] Krueger-Koplin, R. D., Sorgen, P. L., Krueger-Koplin, S. T., Rivera-Torres, I. O., Cahill, S. M., Hicks, D. B., Grinius, L., Krulwich, T. A., and Girvin, M. E. (2004) An evaluation of detergents for NMR structural studies of membrane proteins, *J. Biomol. NMR* 28, 43-57.
- [60] Tian, C., Vanoye, C. G., Kang, C., Welch, R. C., Kim, H. J., George, A. L., Jr., and Sanders, C. R. (2007) Preparation, functional characterization, and NMR studies of human KCNE1, a voltage-gated potassium channel accessory subunit associated with deafness and long QT syndrome, *Biochemistry* 46, 11459-11472.
- [61] Aiyar, N., Nambi, P., Stassen, F., and Crooke, S. T. (1987) Solubilization and reconstitution of vasopressin V1 receptors of rat liver, *Mol. Pharmacol.* 32, 34-36.
- [62] Aiyar, N., Bennett, C. F., Nambi, P., Valinski, W., Angioli, M., Minnich, M., and Crooke, S. T. (1989) Solubilization of rat liver vasopressin receptors as a complex with a guanine-nucleotide-binding protein and phosphoinositide-specific phospholipase C, *Biochem. J.* 261, 63-70.

- [63] Aiyar, N., Valinski, W., Nambi, P., Minnich, M., Stassen, F. L., and Crooke, S. T. (1989) Solubilization of a guanine nucleotide-sensitive form of vasopressin V2 receptors from porcine kidney, *Arch. Biochem. Biophys.* 268, 698-706.
- [64] Huang, P., Liu, Q., and Scarborough, G. A. (1998) Lysophosphatidylglycerol: a novel effective detergent for solubilizing and purifying the cystic fibrosis transmembrane conductance regulator, *Anal. Biochem.* 259, 89-97.
- [65] Shibuya, I. (1992) Metabolic regulations and biological functions of phospholipids in *Escherichia coli*, *Prog. Lipid Res.* 31, 245-299.