Investigation of Protein Structure and Dynamics

PhD Thesis
in partial fulfillment of the requirements
for the degree “Doctor of Philosphy” (PhD)
in the Molecular Biology Program
at the Georg-August-University Göttingen,
Faculty of Biology

submitted by

Benedikt T.C. Frank

born in

Ulm, Germany

Göttingen 2009

Thesis Committee:

Prof. Dr. Christian Griesinger

Prof. Dr. George M. Sheldrick

Prof. Dr. Erez Raz

Affidavit

I hereby declare that this PhD thesis 'Investigation of Protein Structure and Dynamics' has been written independently with no other aids or sources than quoted. This thesis (wholly or in part) has not been submitted elsewhere for any academic award or qualification.

Benedikt T.C. Frank

April, 2009

Göttingen, Germany

Related Publications

Parts of the work presented in this thesis are based on the following publications or manuscripts. I am very grateful to all co-authors of these papers, as well as the people acknowledged in them, for the stimulating and fruitful cooperations.

- Benedikt T.C. Frank, Edward d'Auvergne, Fernando Rodriguez-Castaneda, Stefan Becker and Christian Griesinger. Dynamical insight into CaM/IQ domain motions. *In preparation.*

- Benedikt T.C. Frank, He-Hsuan Hsiao, Erik Meulmeester and Henning Urlaub. ChopTools: an *in silico* tool for the identification and sequence analysis of substantially modified peptides and proteins by highly accurate mass spectrometry. *2009*, submitted.

- He-Hsuan Hsiao*, Erik Meulmeester*, Benedikt T.C. Frank*, Frauke Melchior, and Henning Urlaub. Chop'N'spice, a mass spectrometric approach that allows identification of endogenous SUMO conjugated peptides. *2009*, submitted. (*equally contributing authors)

In addition, results of the work described in this thesis have been presented at international meetings or symposia. I wish to express my kindest regards also to all people involved in these studies.

- He-Hsuan Hsiao, Erik Meulmeester, Benedikt T.C. Frank, Frauke Melchior and Henning Urlaub. Mapping endogenous SUMO sites: A novel approach using ESI-MS and modified database search with common search engines. 57th ASMS Conference on Mass Spectrometry 2009, Abstract No. 1582.

- Benedikt T.C. Frank, Edward d'Auvergne, Fernando Rodriguez-Castaneda, Nicolas Coudevylle, Andrei Leonov, Stefan Becker and Christian Griesinger. Paramagnetic Alignment of a Calmodulin/IQ-Complex: Dynamic Insight into Protein Domain Motions. XXIIIrd ICMRBS 2008, Poster No. 259.

- Fernando Rodriguez-Castaneda, Nicolas Coudevylle, Christophe Fares, Benedikt T.C. Frank, Andrei Leonov, Stefan Becker and Christian Griesinger. Tag-mediated paramagnetic alignment of proteins: Dynamic insights into domain motions in Calmodulin. 47th ENC 2007, Poster No. 016.

Furthermore, I was involved in cooperations leading to the following publication. I wish to express my kindest regards also to all people involved in these studies for the productive collaboration.

- Christophe Farès, Nils-Alexander Lakomek, Korwin F.A. Walter, Benedikt T.C. Frank, Jens Meiler, Stefan Becker, Christian Griesinger. Accessing ns-μs side chain dynamics in ubiquitin with methyl RDCs. *2009* Journal of Biomolecular NMR, advance online publication.

Contents

List of Figures

List of Tables

Acknowledgements

I am very grateful and indebted to:

Prof. Dr. Christian Griesinger of the Max-Planck-Institute for Biophysical Chemistry (MPI-BPC), for supervision, providing an outstanding research environment at his Department for NMR-based Structural Biology, his approachability, open mind, and support of my work.

Prof. Dr. George M. Sheldrick of the University of Göttingen for serving on my thesis committee, and supporting of my work.

Prof. Dr. Erez Raz, now at the Center for Molecular Biology of Inflammation in the University of Münster, for his guidance of my scientific work and occasional lab space for a non-developmental biologist.

Dr. Stefan Becker for numerous helpful discussions and advice concerning the Dead end project, and Karin Giller for her expert advice in molecular biology techniques as well as performing the baculovirus expression tests.

Dr. Henning Urlaub, He-Hsuan Hsiao, Erik Meulmeester and Frauke Melchior for the fruitful collaboration on the ChopTools project.

Dr. Edward d'Auvergne for numerous helpful discussions concerning the calmodulin project, and his excellent analysis framework relax.

Dr. Christophe Fares and Dr. Nils-Alexander Lakomek for helpful discussions on many aspects of NMR spectroscopy, Dr. Volker Klaukien and Gerhard Wolf for peptide synthesis and purification and Dr. Andrei Leonov for synthesis of paramagnetic tags. Nicolas Coudevylle and Fernando Rodriguez-Castaneda for helpful discussions on the calmodulin project.

The many members of the NMR II department, past and present, who have provided friendly and pleasant atmosphere to work in.

Dr. Steffen Burkhardt, and his assistants past and present, Nina McGuiness, Ivana Bacakova and Kerstin Grüniger for their continuous support and excellent organisation of the IMPRS Göttingen.

Burkard Heisen who is not only a helpful colleague at the institut, but an inexpendible partner at BitConf and simply a very good friend at any time not covered by the previous two occupations.

My friends and colleagues Florian, Marc, Christian who have made my time in Göttingen very enjoyable.

Larisa, a very special person to me who I am very glad to have met here.

My parents, Dieter Frank and Renate Frank, for their continuous support throughout my life.

Abstract

This thesis addresses three aspects of protein function:

Part I describes the structural investigation of a calmodulin-peptide complex by NMR spectroscopy. The intracellular calcium sensor calmodulin, model system for a two domain protein, is known to interact with a large variety of different peptides, among them the IQ-motif in the cytoplasmic C-terminal tail of the $Ca_V\alpha_1$ subunit of voltage gated ion channels. A crystal structure of the CaM/IQ-motif complex showed that the CaM N-terminal lobe adopts two conformations suggesting internal conformational plasticity. In order to explore this plasticity and to work towards motional models of the underlying processes, paramagnetic alignment was employed to probe domain dynamics using residual dipolar couplings (RDC), a versatile NMR parameter. The results are compared to existing X-ray structures of the complex. It is concluded that in solution a larger conformational space is sampled by calmodulin than predicted by the available structures, but that these structures are likely to be a subset of this conformational space.

Part II describes the development of a web based tool that can aid with the identification of protein-protein cross-links and peptide-based post-translational modifications by mass spectrometry. Although several well-established and powerful software suites are available (e.g. Mascot [98], Sequest [99]), these are not suitable for addressing a number of scenarios opened up by recent developments. This is especially true in the case of large peptide modifications. Such modifications include post-translational modifications (PTM), most notably Ubiquitin (Ubq) and the Ubiquitin-like modifier (Ubl) families, but also general protein-protein cross-linking. The latter may be endogenous - i.e. disulphide bridges - or engineered through the use of cross-linking reagents, to study the three-dimensional structure of, or the interactions between proteins. To this end, a large number of specialised programmes addressing specific questions have been proposed, each with individual advantages and disadvantages. In contrast, the ChopTools suite is not limited to a specific modifier or affected by the drawbacks of predictive approaches. Instead,

it provides a convenient and ubiquitous way of reformulating the question through the generation of fragment library files. This in turn allows well-established programmes to address the questions at hand thereby bridging the gap between existing analysis schemes and new experimental setups.

Part III describes progress towards determining a high resolution structure of the Dead end protein. This protein has been shown to play a crucial role in fate maintenance, survival and migration of primordial germ cells in the model organism zebrafish (*Danio rerio*). In addition, a particular mutation in the gene encoding the mouse Dead end orthologue results in germ cell loss and testicular tumours. Close homologues of the protein are also found in african clawed frog, chick and human, suggesting a central and general role for this protein in germ cell development. Primary sequence analysis and functional characterisation suggest that the essential core region of the protein consists of a putative RNA or protein interaction domain with an RNP fold and another domain of similar size with unknown structure. Recent studies have shown that Dead end is capable of protecting mRNAs from microRNA (miRNA) mediated repression. A wide range of protein expression and purification strategies are discussed in the context of the Dead end protein. Although no sample amenable to structural investigation could be obtained, the work has laid the foundation of future biophysical and biochemical investigations of this disease relevant protein, and has established a protocol for the preparation of dilute recombinant protein solutions that may prove valuable in the generation of antibodies to the Dead end protein.

Part I

Introduction

Chapter 1

Introduction

Proteins are one of the most important building blocks of life. Essential parts of any organism, they participate in virtually every process within the cell. Consequently, many disciplines have evolved to study proteins and their functions. One of the most prominent fields in this respect is structural biology. Since its advent, marked by the discovery of DNA structure by Watson and Crick in 1953 [1], and thereafter the solution structures of myoglobin [2] and hemoglobin [3] by Kendrew and Perutz respectively, this field has since grown rapidly.

Already at its birth, the field demonstrated that the foundation for understanding function at a molecular level lies in its structure, as the structure of DNA directly suggested a possible mechanism for its replication. Since then, structural biology has continued to contribute to our understanding of molecular function, particularly through the two predominant methods offering information at atomic resolution, X-ray crystallography and NMR spectroscopy.

Whilst X-ray crystallography has provided the majority of structural models available to date, NMR, in addition to structural models has also provided valuable information about macromolecular dynamics. This data is able to expand and complement the static description that high resolution structures provide allowing a detailed understanding of the molecular basis for protein function.

This is the case with the model system introduced in Chapter 2. Calmodulin, a two domain protein and intra-cellular calcium sensor, is known to interact with a large variety of different targets. In complex with a particular target, the IQ peptide from a voltage gated ion channel, and using X-ray crystallography calmodulin is

seen to adapt two different conformations which are proposed to be of functional significance [4]. To investigate how this conformational plasticity manifests itself in solution, and to compare this to the crystal structure, an NMR technique known as paramagnetic alignment was employed. This enables the study of inter-domain orientations and dynamics in this complex, hence complementing the existing two-site model.

In living systems there are numerous protein-protein interactions and complexes, both transient and permanent, that are essential to all cellular functions. Chapters 6 - 8 discuss approaches through which such interactions can be identified. A novel software tool is described which can easily be integrated in existing analysis schemes and which aids the identification of non-covalent interactions, like the calmodulin-IQ interaction described above, as well as more permanent, covalent protein-protein modifications such as ubiquitination.

Before a structural study, or any scientific characterisation outside the context of a living cell, can be undertaken, it is imperative that suitable samples of the protein of interest can be prepared. Chapter 9 introduces the disease implicated and developmentally important Dead end protein, for which no sample preparation protocol has yet been established. First identified in the model organism zebrafish, this protein has been shown to play a crucial role in fate maintenance and has since been linked to testicular cancer [5, 6]. However, the function of this protein has remained largely elusive, in part due to the lack of structural data. Chapters 11 and 12 describe the pursuit towards biochemical characterisation and a high resolution structure of Dead end, by seeking to establish protocols to obtain pure protein samples to facilitate such studies.

Part II

Calmodulin Domain Dynamics

Chapter 2

Background

'... in the winter of our first experiments ... looking on snow with new eyes. There the snow lay around my doorstep - great heaps of protons quietly precessing in the Earth's magnetic field. To see the world for a moment as something rich and strange is the private reward of many a discovery.'

—Eward M. Purcell, Nobel lecture, 1952

2.1 Summary

The intracellular calcium sensor Calmodulin (CaM)[1], model system for a two domain protein, is known to interact with a large variety of different peptides, among them the IQ-motif in the cytoplasmic C-terminal tail of the $Ca_V\alpha_1$ subunit of voltage gated ion channels. A crystal structure of the CaM/IQ-motif complex showed that the CaM N-terminal lobe adopts two conformations suggesting internal conformational plasticity.

In order to explore this plasticity and to work towards motional models of the underlying processes, paramagnetic alignment was employed to probe domain dynamics using residual dipolar couplings (RDC), a versatile NMR parameter.

[1]N.B. Calmodulin itself is an abbreviation for CALcium MODULated proteIN.

2.2 The Calmodulin/IQ system

2.2.1 Calmodulin

CaM is a ubiquitous, calcium-binding protein that can bind to and regulate a plethora of different protein targets, thereby modulating many different cellular functions. In response to cellular calcium signalling, CaM regulates processes such as inflammation, metabolism, apoptosis, muscle contraction, intracellular movement, short-term and long-term memory, nerve growth and the immune response [7, 8]. CaM is expressed in many cell types and can have different subcellular locations in the cytoplasm, or associated with plasma or organelle membranes [7]. Many of the proteins that CaM binds are unable to bind calcium themselves, and therefore use CaM as a calcium sensor and signal transducer.

EF-hands and conformational changes

Calcium binding in CaM is mediated through four 'EF-hand' motifs, occurring in tandem in both the N-terminal domain and the C-terminal domain, connected by a flexible linker in the centre of the protein [9, 10]. Each EF-hand consists of a helix-loop-helix motif in which the 12 or 14 amino acid long loop can coordinate a Ca^{2+} ion. Whether the binding is independent or cooperative has been a matter of debate [11, 12, 13, 14] with recent findings leaning towards models linking calcium binding cooperativity in CaM to conformational changes [15, 16].

The conformational change that CaM undergoes upon binding to calcium exposes methionine-rich, hydrophobic pockets in each of its domains [17] thus enabling CaM to recognize, bind to, and activate a wide range of proteins and enzymes [8, 18, 19]. This intriguing mechanism is facilitated by complex structural rearrangements of the interfaces between the four helices in each domain. In the absence of calcium, termed the apo state, the four helices of each domain adopt an approximately anti-parallel configuration (closed conformation), whereas upon calcium binding, the motif helices are transformed into a nearly orthogonal configuration (open conformation) [8].

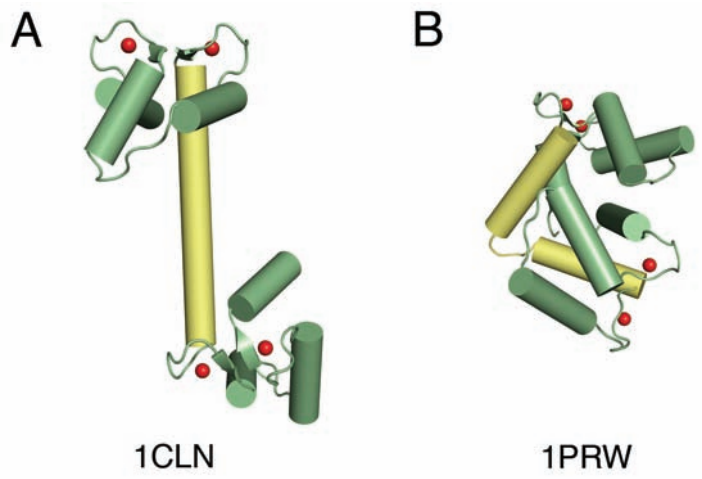

Figure 2.1: Extended (A) and compact (B) conformations of Ca^{2+}CaM in cartoon representations. Ca^{2+} ions are shown in red, the helices of the N- and C-terminal domains that form the extended helix in the PDB structure 1CLN are shown in yellow. The compact structure 1PRW reveals the flexibility of the inter-domain linker.

The inter-domain linker

In addition to structural rearrangements within its EF-hands, CaM also shows considerable flexibility in its inter-domain orientations. Early X-ray data (1985) suggested a dumbbell like shape for calcium saturated CaM (Ca^{2+}/CaM) in which the last and first helices, of the N- and C-terminal domains respectively, form a long continuous helical structure together with the inter-domain linker – a finding to be resembled by all structures of unbound Ca^{2+}/CaM determined over the course of the next 18 years (Fig. 2.1-A) [9]. In contrast, small angle X-ray scattering suggested that CaM in solution should have a significantly less extended shape than observed in the crystal [20], and fluorescence energy transfer measurements in solution indicated a shorter distance between the two domains [21, 22]. Furthermore, NMR data demonstrated considerable flexibility in the inter-domain linker of (calcium free) apo-CaM [17, 23, 24]. It was not until 2003 when the first native Ca^{2+}/CaM X-ray structure was solved that revealed unbound CaM in a compact, rather than extended, conformation (Fig. 2.1-B) [25]. The initially observed α-helical conformation of the central linker has been attributed to crystal packing, whilst compact

conformations have been shown to be of biological relevance, yet the large difference between both structures underlines the flexible nature of the region linking both CaM domains.

The high degree of flexibility inherent to the inter-domain linker plays a key role in enabling CaM to interact with its wide range of target peptides, as demonstrated by numerous structures solved of Ca^{2+}/CaM bound to target peptides or small molecules [8]. In the majority of structures, CaM wraps around amphiphilic helices governed by selectivity imposed through a unique binding site comprising both CaM lobes as well as the linker region [26].

In addition to its biological importance as a calcium sensor and its role in signal transduction, Calmodulin is a paradigm of multi-domain systems, which has been at the centre of many studies investigating conformational changes and inter-domain dynamics in proteins (see Section 2.4).

2.2.2 Voltage gated channels

A prominent example of an α-helical peptide that CaM binds to is the IQ domain of the cardiatic Ca_V 1.2 calcium channel. This voltage gated channel, also referred to as L-type, controls cellular calcium entry in response to changes in membrane potential and thereby governs excitation-contraction coupling, cardiatic action potentials, hormone and neurotransmitter release, and other important cellular processes [27, 28]. Ca_V channels are composed of three major subunits: $Ca_V\alpha_1$, $Ca_V\beta$ and $Ca_V\alpha_2\delta$ [29], of which $Ca_V\alpha_1$ has been shown to serve as a docking platform for CaM. It is also this very subunit that together with the calcium sensor CaM has been identified as a critical component in channel regulation [29].

Two different regulatory modes are differentiated: calcium-dependant facilitation (CDF) and calcium-dependant inactivation (CDI). Both modes are calcium dependant, the former causes a positive feedback response and is triggered by membrane depolarization or a rise in basal Ca^{2+} levels [30, 31], whilst the latter acts upon high Ca^{2+} influx through the channel and results in feed-back inhibition [32, 33]. Yet both contrasting responses have been shown to be regulated by Ca^{2+}CaM bound to the IQ motif of the α_1 subunit [34, 35].

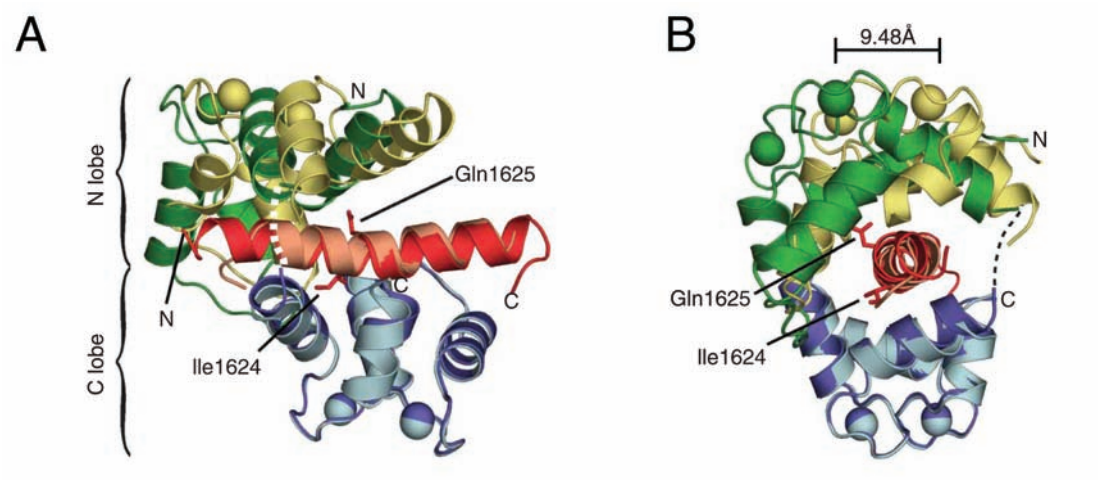

Figure 2.2: X-ray structure of the Ca^{2+}CaM/IQ complex. **A.** N-terminal domains are shown in green and yellow, C-terminal domains in blue and the IQ peptide in red. Key anchor points are indicated. **B.** Rotation of A by 90°. Adapted from [4].

Several independent studies indicate that CaM to be constitutively tethered to the channel it regulates allowing a rapid response to calcium influx [36, 37, 38]. Both CDF and CDI have been shown to be regulated by calcium binding to specific CaM sites, implying that conformational changes within CaM and resulting specific interactions between CaM and the IQ domain are the key to understanding how the two opposing mechanisms may be regulated [37, 39, 40]. The publication of two crystal structures [4, 41] have contributed much to our understanding of the underlying molecular basis of these mechanisms. The former of the two structures, [4], is particularly intriguing as it displays two similar yet distinct conformations in the presence of calcium (see Fig. 2.2), suggesting an inherent conformational plasticity directly at the CaM/IQ interface.

This system provides an ideal model to investigate the extend of the proposed conformational plasticity in solution to further our understanding of how a single sensor, CaM, is able to specifically regulate such a wide range of targets and responses.

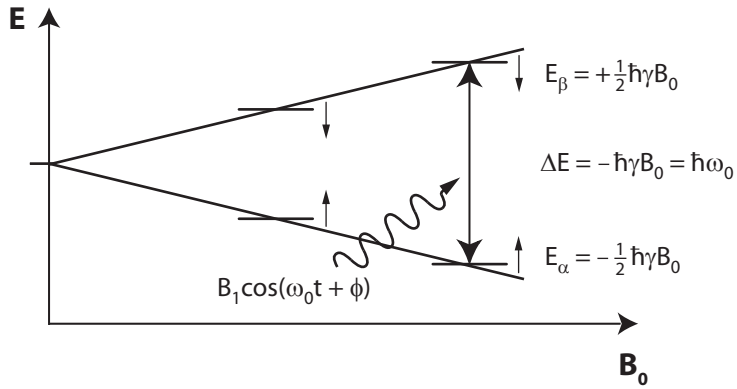

Figure 2.3: Illustration of the Zeeman splitting for a spin-$\frac{1}{2}$ nucleus. Energies (E) are shown in dependence of magnetic field strength ($\boldsymbol{B}_0$). The resonant radio-frequency field is depicted by a wave [45].

2.3 Theory

2.3.1 The NMR phenomenon

Nuclear magnetism is the manifestation of nuclear spin angular momentum, an intrinsically quantum mechanical property that does not have a classical analogue. The physical origins of nuclear spin angular momentum and much of the quantum mechanical theory describing NMR is complex. A thorough treatise of the matter will therefore be sacrificed for the sake of brevity, as a large number of excellent resources covering the material in detail are available in print [42, 43, 44]. The remainder of this section will introduce the fundamental basis of NMR spectroscopy and discuss parameters relevant to the scope of this work.

In the presence of an applied magnetic field $\boldsymbol{B}_0$, the eigenstates of spin angular momentum operators of atomic nuclei with non-zero spin angular momentum present different energy levels. The resulting energy difference is known as Zeeman splitting. NMR spectroscopy saw its birth when Bloch and Purcell (Nobel Prize in physics 1952) proposed to probe transitions between these states by applying an oscillating magnetic field $\boldsymbol{B}_1$, orthogonal to the existing static field, with a frequency ω_0 that is equivalent to the energy difference between the eigenstates (see Fig. 2.3).

Quantitatively, the energy difference ΔE is given by

$$\Delta E = -\hbar\gamma \boldsymbol{B}_0 \ , \tag{2.1}$$

where γ is the gyromagnetic ratio of the nucleus of interest and $\hbar$ is Planck's constant over 2π. Substituting Planck's law yields the corresponding frequency ω_0 of electromagnetic radiation, also referred to as Larmor frequency:

$$\omega_0 = \frac{\Delta E}{\hbar} = -\gamma \boldsymbol{B}_0 \ , \tag{2.2}$$

also known as the fundamental resonance condition. Whilst early experimental setups saw the sweeping of a magnetic field strength range to observe resonance, the field was revolutionized by the onset of Fourier-transform NMR spectroscopy introduced by Ernst and Anderson (Nobel prize in chemistry 1991). Employing short radio frequency pulses covering a frequency range around ω_0 allows the simultaneous excitation of multiple resonance transitions. This concept has been applied to many experimental setups optimized for different NMR parameters, and through the introduction of one or more further orthogonal time dependent variables it is now common practise to record these in the form of correlation maps in multidimensional NMR [44].

2.3.2 Chemical shift

Returning to the Larmor frequency (Equation 2.2 above), magnetic field strengths of several tesla, achieved through superconducting magnets, are not uncommon anymore, yielding resonance frequencies in the megahertz (MHz) range, *i.e.* radio waves. The precise resonance frequency of a particular nucleus however, does not only depend on its gyromagnetic ratio (γ) and the applied field ($\boldsymbol{B}_0$) alone. Rather it is the effective magnetic field at its precise position in space, which is the sum of $\boldsymbol{B}_0$ and any local magnetic fields in its environment, such as those generated by other nuclei or induced by the motions of electrons. This effective resonance frequency carries valuable information about the local magnetic and thus chemical environment of a nucleus and is referred to as chemical shift. Chemical shifts form the basis of structural investigations by NMR, both by the intrinsic structural information they can provide, as well as through the use of this signature frequency in the correlation

of interactions [44]. Chemical shifts are often on the order of 10^6 times smaller than the Larmor frequency, and are thus often quoted in ppm (parts per million) as offsets from the resonance frequency of a reference compound, providing a measure independent of the field strength at which it was recorded.

2.3.3 Scalar coupling

The indirect spin-spin interaction mediated through electrons[2] are known as J-couplings. Their anisotropic component is usually neglected, simplifying the scalar coupling Hamiltonian $\hat{H}_J$ of an IS spin system in the weak coupling limit to

$$\hat{H}_J = 2\pi J \hat{\boldsymbol{I}}_z \hat{\boldsymbol{S}}_z \ , \tag{2.3}$$

where J is the scalar coupling constant, $\hat{\boldsymbol{I}}$ and $\hat{\boldsymbol{S}}$ are nuclear spin operators of spin I and S respectively. This key interaction is often exploited to transfer magnetisation between spin systems in solution state NMR pulse sequences [44].

In addition, ${}^3J$ scalar couplings are able to provide structural information about their environment through their dependence on the dihedral angle formed by the three covalent bonds involved. Despite many efforts to refine existing models to derive precise structural information from J-couplings, theoretical approaches have proven to be generally not applicable due to the complex relationships involved. Instead the originally proposed relationship between ${}^3J$ scalar couplings and dihedral angles, known as the *Karplus equation* is used

$${}^3J = A\cos^2\theta + B\cos\theta + C \ , \tag{2.4}$$

where θ is the dihedral angle, and A, B and C are semiempirical constants that depend upon the nuclei involved in the interaction.

2.3.4 Dipolar coupling

Through-space dipole-dipole interactions between heteronuclei I and S have the following Hamiltonian $\hat{H}_D$ in the weak coupling limit:

[2] Usually covalent bonds, but also H-bonds under certain conditions [44].

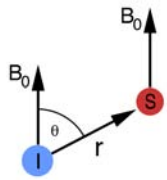

Figure 2.4: The dipolar coupling interaction as described by Equation 2.6.

$$\hat{H}_D = 2\pi D \hat{\boldsymbol{I}}_z \hat{\boldsymbol{S}}_z \ , \tag{2.5}$$

which is of the same form as the weak J coupling Hamiltonian given above, and represents the dominating interaction in solids. D is the dipolar coupling constant, in the laboratory frame, given by

$$D_{IS} = -\frac{\mu_0 \gamma_I \gamma_S h}{(2\pi)^3} \left\langle \frac{3\cos^2\theta - 1}{\boldsymbol{r}^3} \right\rangle \ , \tag{2.6}$$

where μ_0 is the permittivity of free space, γ_I and γ_S the gyromagnetic ratios of I and S, $\boldsymbol{r}$ the distance between the nuclei and θ is the angle that this internuclear vector makes with magnetic field ($\boldsymbol{B}_0$ in the laboratory frame) – c.f. Fig. 2.4. Angular brackets denote motional averaging. In the case of two covalently bonded nuclei, such as ^{1}H and ^{15}N, in an isotropically tumbling molecule in solution, the effect of motional averaging on $\boldsymbol{r}$ is negligible, however the second term averages to zero. This is due to its angular dependence, as fast tumbling in a liquid causes all directions to be uniformly sampled in space.

2.3.5 Residual dipolar couplings

It becomes immediately apparent that, whilst dipolar interactions dominate in solid state NMR, they are normally not observed in solution state NMR due to motional averaging. In order to extract valuable orientational information from dipolar interactions, anisotropy must be introduced into the sample. When the molecule under investigation longer uniformly samples all directions in space, the dipolar interaction no longer averages to zero and thus becomes visible as a residual dipolar coupling.

Many approaches have been introduced to accomplish this partial alignment, most widely used are bicelles, purple membrane fragments, Pf1 phages, alcohol/surfactant phases, stretched, compressed and/or charged gels (for a review see [46]). Through a combination of steric and/or electrostatic interactions, the free tumbling of target proteins is affected thus causing anisotropy in the sample. This process of introducing anisotropy is referred to as partial alignment, where the ordering is much smaller than in crystals[3]. RDCs measured in such aligned samples are a valuable tool for refining structures [47], answering conformational questions and, due to their motional averaging, giving insights into dynamics [48, 49, 50, 51, 52].

Paramagnetic alignment

The alignment media described so far cause external alignment, inducing the same amount of order in all parts of a globular protein. When investigating a multi-domain protein and its inter-domain motions, as is the case with CaM, use of classical alignment media severely complicates the analysis, as relative domain orientations are affected by both inter-domain motions as well as the alignment process of each domain. Paramagnetic alignment poses a very elegant solution to this problem. Through the introduction of a paramagnetic species into one domain, the anisotropic magnetic susceptibility of the paramagnetic ion also confers alignment to this domain. The degree of order in the second domain then only depends on the flexibility of the inter-domain linker and resulting inter-domain motions. This key concept of aligning one domain and investigating the degree of alignment in the second is central to much of the work described in Chapter 4.

Paramagnetic alignment can be achieved by introducing a lanthanide into an endogenous or engineered metal binding site within the protein of interest [53, 54]. Alternatively, a lanthanide can be coordinated by a peptide or ligand based tag which is then in turn attached to the protein [55, 56, 57, 58]. Further details of both methods are discussed in Section 3.2.4.

[3]N.B. Dipolar couplings observed in solids are usually larger by a factor of 10^3 or more.

The alignment tensor

The nature of weak alignment introduces a further unknown into Equation 2.6 that may not be immediately apparent. Analysing RDCs in terms of angular constraints requires knowledge of the amount and direction of order. As this order need not be axially symmetric, it becomes necessary to specify the orientation of a complete ordering frame. This is most conveniently achieved by introducing a probability tensor $\boldsymbol{P}$, which is the orientational probability distribution of the external magnetic field in the molecular fixed reference frame [59]. The principal components of this tensor in a Cartesian coordinate system, P_x, P_y, P_z define a special frame of reference, the alignment frame. In this axis system, $\boldsymbol{P}$ corresponds to a 3×3 symmetric matrix with trace 1, allowing the dynamical average of the term $\cos^2\theta$ in Equation 2.6 to be expressed as

$$< \cos^2\theta > = P_x r_x^2 + P_y r_y^2 + P_z r_z^2 \ , \tag{2.7}$$

where r_x, r_y, r_z are the Cartesian components of the internuclear vector $\boldsymbol{r}$. Instead of the probability tensor $\boldsymbol{P}$, a slightly different definition of molecular order is often used in the literature, the alignment tensor, $\boldsymbol{A}$ [60]. The alignment tensor is the traceless part of $\boldsymbol{P}$ and given by

$$\boldsymbol{A} = \boldsymbol{P} - \frac{1}{3}\,\mathbf{1} \ . \tag{2.8}$$

This definition is analogous to the one common in liquid crystal literature where it is referred to as Saupe matrix [61].

Rewriting Equation 2.6 in terms of the principal components A_{zz}, A_{yy}, A_{xx} of the alignment tensor, where $|A_{zz}| > |A_{yy}| > |A_{xx}|$, yields the residual dipolar coupling

$$D_{rdc}(\theta,\phi) = -\frac{\mu_0 \gamma_I \gamma_S h}{(2\pi r)^3}\left[A_{zz}(3\cos^2\theta - 1) + (A_{xx} - A_{yy})\sin^2\theta\cos(2\phi)\right] \ . \tag{2.9}$$

This equation is also commonly written in terms of the axial D_a and rhombic R components of $\boldsymbol{A}$, which are defined as follows:

$$D_a = \frac{1}{3}\left(A_{zz} - \frac{A_{xx} + A_{yy}}{2}\right) = \frac{A_{zz}}{2} \ , \tag{2.10}$$

$$R = \frac{A_{xx} - A_{yy}}{3} \ , \tag{2.11}$$

giving

$$D_{rdc}(\theta, \phi) = -\frac{\mu_0 \gamma_I \gamma_S \hbar}{2\pi^2 r^3} \left[D_a (3\cos^2\theta - 1) + \frac{3}{2} R \sin^2\theta \cos(2\phi) \right] \ . \tag{2.12}$$

Practical implications

Considerations regarding the preparation of paramagnetically aligned samples are discussed in Section 3.2.4. When measuring such a sample, the RDC is observed as an addition to scalar couplings, as the internal Hamiltonian is the sum of the scalar $\hat{H}_J$ and dipolar $\hat{H}_D$ Hamiltonians. This can be rationalised by considering the Hamiltonian $\hat{H}$ of two spins in a magnetic field, which can be split into two parts

$$\hat{H} = \hat{H}_{ext} + \hat{H}_{int} \ , \tag{2.13}$$

describing the interactions with external magnetic fields[4] ($\hat{H}_{ext}$) and between themselves ($\hat{H}_{int}$). $\hat{H}_{int}$ in turn is the sum of the scalar and dipolar Hamiltonians (c.f. Equation 2.3 and Equation 2.5)

$$\hat{H}_{int} = \hat{H}_J + \hat{H}_D \ . \tag{2.14}$$

Consequently, two samples are required to measure RDCs, one isotropic sample in which the dipolar interaction averages to zero and thus provides scalar couplings, and a second anisotropic sample from which the sum of scalar and dipolar interactions can be extracted. The RDC is then given by the difference between the two measurements. Two predominant principles are employed to extract scalar and dipolar interactions: line positions as e.g. used in popular in-phase anti-phase (IPAP) HSQC experiments, or J-modulation as applicable to constant time HSQC and triple resonance experiments[5].

[4] External magnetic fields refers to both $\boldsymbol{B}_0$ and $\boldsymbol{B}_1$.

[5] For a comprehensive review on available pulse sequences and methodologies, please refer to [62].

2.3.6 The Pseudo-contact shift

The presence of paramagnetic species in the vicinity of a protein causes two prominent effects, pseudo-contact shifts (PCS), and paramagnetic relaxation enhancement (PRE). PCS arise from through-space dipolar interactions with rapidly relaxing unpaired electrons of the lanthanide. This effect is dependant on the distance between interacting spins and the magnetic susceptibility tensor, χ, which governs the interaction of the paramagnetic dipole moment with the external magnetic field ($\boldsymbol{B}_0$). In NMR spectroscopy PCS manifest as large changes δ^{PCS} in chemical shifts of affected nuclear spins, which, in analogy to Equation 2.12, can be expressed in terms of the axial ($\Delta\chi_{ax}$) and rhombic ($\Delta\chi_{rh}$) components of the magnetic susceptibility tensor χ

$$\Delta\delta^{PCS} = \frac{1}{12\pi \boldsymbol{r}^3}\left[\Delta\chi_{ax}(3\cos^2\theta - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\theta\cos 2\phi\right] , \tag{2.15}$$

where θ and ϕ are the polar coordinates describing the paramagnetic centre vector nuclear spin with respect to the principal axes of the $\Delta\chi$ tensor, and $\boldsymbol{r}$ is the distance between the nuclear spin and the metal ion.

Finally, magnetic anisotropy induced alignment, it should be noted that D_a and R in Equation 2.12 are related to the analogous axial χ_{ax} and rhombic χ_{rh} components of the magnetic susceptibility tensor χ as follows:

$$D_a = \frac{1}{2}\frac{B_0^2}{15\mu_0 kT}\,\Delta\chi_{ax} \tag{2.16}$$

and

$$R = \frac{1}{2}\frac{B_0^2}{15\mu_0 kT}\,\Delta\chi_{rh}\ . \tag{2.17}$$

2.3.7 Paramagnetic relaxation enhancement

The second prominent effect caused by paramagnetic species is a relaxation mechanism similar to the Nuclear Overhauser effect (NOE), known as paramagnetic relaxation enhancement (PRE). In the case of short-lived electronic lifetimes as found in paramagnetic lanthanides the predominant term of this transverse relaxation enhancement R_2^{PRE} in the slow tumbling limit is due to Curie-spin relaxation [63]

$$R_2^{PRE} \propto \frac{\gamma_I^2 \boldsymbol{B}_0^2 \left[g_e S_e (S_e+1)\right]^2 \tau_r}{\boldsymbol{r}^6 T^2} \ , \tag{2.18}$$

where g_e are the electronic g-factor and S_e the spin of the electron, τ_r the rotational correlation time of molecular tumbling, $\boldsymbol{r}$ the distance, T the temperature and $\boldsymbol{B}_0$ the external magnetic field strength. Although PREs have been employed to derive distance restraints [64, 65], the paramagnetic broadening they introduce is often undesirable and has a dramatic effect in internal paramagnetic alignment.

2.4 Dynamics

NMR has emerged as a powerful tool to investigate protein dynamics, partly because of the availability of a multitude of dynamic parameters. Commonly used experiments and the time-scales they are sensitive to are summarised in Fig. 2.5. Early studies have concentrated on intra-molecular dynamics, such as backbone and side-chain dynamics. With the advent of paramagnetic alignment techniques it has become possible to extend dynamics studies to inter-domain motions, as in such a system the alignment process does not interfere with the analysis of the motion.

Given alignment of one domain, the strength of alignment of the second, and consequently also the size of its alignment tensor relative to that of the first, are dependant on inter-domain motions. In an entirely rigid system, the alignment of both domains is identical, thus the alignment tensors of each domain match. In contrast, inter-domain mobility leads to a reduction of alignment in the second domain which may be described in terms of the dynamically averaged alignment tensor $\bar{x}_i$ of the mobile domain over a time t_{max} as

$$\bar{\chi}_i = \left\langle \int_0^{t_{max}} R^{-1}\left(\Omega_t\right) \cdot \chi_i \cdot R(\Omega_t)\, dt \right\rangle \ , \tag{2.19}$$

where $R(\Omega_t)$ is a time-dependant rotation matrix. This trajectory and ensemble averaged alignment tensor can be modelled by approximating the conformational space by N discrete states, each with an associated probability p_c

$$\bar{\chi}_i(\theta) = \sum_{c=1}^{N} p_c\, R^{-1}\left(\Omega_c\right) \cdot \chi_i \cdot R(\Omega_c) \ , \tag{2.20}$$

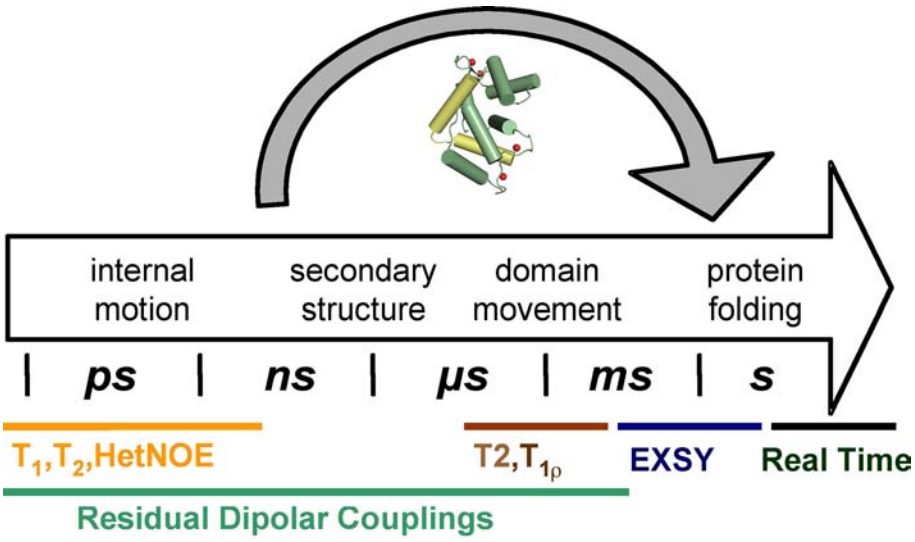

Figure 2.5: NMR parameters and the time-scales these parameters are sensitive to, relative to the correlation time of the molecule under investigation.

thus removing the time-dependence from the equation [54]. This 'N-state model' has been extended to determine maximum allowable probabilities (p_{max}) for fixed orientations [66, 67]. Recently, these to models have been absorbed into a single model, the MAP(R) analysis [68].

Chapter 3

Materials & Methods

Quidquid latine dictum sit, altum videtur.
—Latin proverb.

3.1 Materials

3.1.1 *E. coli* strains and plasmids

Plasmids

Plasmid	Resistance	Features	Reference
pET28a	Kan	XCaM wild type (Genbank 1DMO)	F. Rodriguez MPIbpc, Göttingen
pET28a	Kan	XCaM T146C mutant	as above
pET28a	Kan	XCaM N60D mutant	N. Coudevylle MPIbpc, Göttingen
pET28HMT	Kan	N-termial His_6-tag, MBP, TEV cleavage site $Ca_V\alpha 1_c$ IQ domain	F. van Petegem UCSF

Table 3.1: Names, features and references of plasmids used

Protein sequences

For the full amino acid sequences of all calmodulin constructs used, please refer to Section A.1 in the appendix.

E. coli strains

Strain	Genotype	Reference
Bl21 (DE3)	F^-, *ompT*, *hsd*S_B, (r_{B^-}, m_{B^-}), *dcm*, *gal* λ(DE3)	[69]
XL2-Blue	recA1 endA1 garA96 thi-1	Stratagene
	hsdR17 supE44 relA1 lac	
	$[FproABlac^qZ\triangle M15Tn10(Tet^r)AmyCam^r]$	

Table 3.2: Names, genotypes and references of bacterial strains used in CaM expression

3.1.2 Growth media and antibiotics

The composition of growth media for *E. coli* cultures is listed in table Table 3.3. All media were either autoclaved or filtered (0.22 μm) before use. Concentrations for plasmid selection were 100 μg/L and 70 μg/L for ampicillin and kanamycin respectively. Agar plates were made by autoclaving a solution containing 15 g agar in 1 L LB medium.

Table 3.3: Growth media and antibiotics

Name	Amount	Substance	Note
2 x YT-Medium	16 g	Trypton	
	10 g	Yeast extract	
	5 g	NaCl	
	add to 1 L	ddH_2O	
LB - Medium	10 g	Trypton	
	5 g	Yeast extract	
	10 g	NaCl	
	add to 1 L	ddH_2O	
M9-Minimal medium	6.8 g	Na_2HPO_4	
	3 g	KH_2PO_4	
	0.5 g	NaCl	
	1 g	NH_4Cl / $[^{15}N]H_4Cl$	
	4 g	Glucose / $[^{13}C_6]$-Glucose	
	2 mL, 1 M	$MgSO_4$	
	50 μL, 2 M	$CaCl_2$	
	0.03 g	Thiaminhydrochloride	
	10 mL	Trace elements	
	add to 1 L	ddH_2O	
Trace elements	80 mL	ddH_2O	add in precisely this order
	0.6 g	$FeSO_4$ x 7 H_2O	stir 10 min
	0.094 g	$MnCl_2$ x 2 H_2O	stir 10 min
	0.08 g	$CoCl_2$ x 6 H_2O	stir 10 min
	0.07 g	$ZnSO_4$ x 7 H_2O	stir 10 min
	0.03 g	$CuCl_2$ x 2 H_2O	stir 10 min
	0.002 g	H_3BO_3	stir 10 min
	0.025 g	$(NH_4)_6Mo_7O_24$ x 4 H_2O	stir 10 min
	add to 100 mL	ddH_2O	stir over night
Ampicillin stock	25 mg/mL	Ampicillin (sodium salt)	sterile filtration, store at -20 °C
Gentamycin stock	20 mg/mL	Gentamycin	sterile filtration, store at 4 °C
Kanamycin stock	70 mg/mL	Kanamycin	sterile filtration, store at -20 °C
IPTG stock	1 M	IPTG	sterile filtration, store at -20 °C

3.1.3 Chemicals & enzymes

Chemicals and enzymes used in this work are listed in Table 3.4. Standard chemicals not listed here were supplied by Merck KGaA (Darmstadt, Germany) and Fluka (Neu-Ulm, Germany). This table also contains reagents used in the investigation of the Dead end protein in Chapter 11.

Table 3.4: Chemicals and enzymes

Chemical/enzyme	**Supplier**
Ampicillin, APS, Disodium-hydrogen phosphate,	Roth, Karlsruhe, Germany
Ethidium bromide, EDTA, Hepes, Imidazole,	
Magnesium-chloride hexahydrate, Potassium dihydrogen phosphate,	
Rotiphorese Gel 30, TEMED, Tryptone,	
Yeast extract, β-Mercaptoethanol	
α-Lactose monohydrate, Adenin, Guanosin, Thymin,	Sigma-Aldrich, Seelze, Germany
Uracil, Cytosin, NAD	
Agar, DNase I, IPTG	AppliChem GmbH, Darmstadt, Germany
Agarose, Kanamycin	GibcoBRL, Karlsruhe, Germany
Ammonium chloride (>98% ^{15}N)	Cambridge Isotope Lab., Andover, USA
Bench Mark protein ladder	Invitrogen, Karlsruhe, Germany
Boric acid	J.T. Baker BV, Deveter, Netherlands
Complete^{TM}-Protease-Inhibitor-Tabletten, EDTA-free	Roche Diagnostics, Mannheim, Germany
D_2O	Deutero GmbH, Kastellaun, Germany
$^{13}C_6$-D-Glucose (> 98% ^{13}C), D_8-Glycerin (98% ^{13}C)	Spectra Stable Isotopes, Columbia, USA
Dithiothreitol (DTT)	Gerbu, Gaiberg, Germany
*Bam*HI, *Eco*RI, *Hind*III, *Nde*I, CIAP	Fermentas, St. Leon-Rot, Germany
Fastruler DNA ladder	Fermentas, St. Leon-Rot, Germany
Guadinium-HCl	Roth, Karlsruhe, Germany
HiTrap Sepharose Q HP	Amersham Biosciences AB,
	Uppsala Schweden
Hot Star Taq, Ni-NTA Agarose	Qiagen, Hilden, Germany
Nucleobond AX PC 100, NucleoSpin Plasmid, Extract II	Macherey-Nagel, Düren, Germany
Phusion^{TM} polymerase	Finnzymes, Espoo, Finland
SDS	Serva, Heidelberg, Germany
Thrombin	Amersham Biosciences, Uppsala, Sweden
Tris	VWR, Hassrode, Belgium
Turbo *Pfu* DNA Polymerase, dNTP	Stratagene, La Jolla, USA

3.1.4 Laboratory equipment

Laboratory equipment and consumables used in this work are listed in Table 3.5. This table also contains equipment used in the investigation of the Dead end protein in Chapter 11.

Table 3.5: Laboratory equipment

Application	Name	Manufacturer
Balances	Sartorius B 3100 S & AC 210 S	Sartorius, Göttingen, Germany
Centrifugation	Beckmann-Coulter Avanti J-20 & J-301,	Beckmann, Krefeld, Germany
	Rotors: JLA 9.100, JA 25.50	
	Eppendorf centrifuge 5415D & 5804	Eppendorf Wesseling-Brenzdorf, Germany
	Heraeus Biofuge primo	Kendro, Hanau, Germany
	Heraeus Multifuge 4 KR	
Chromatography	Disposable columns, 1-5 mL & 2-10 mL	Pierce, Rockford, IL, USA
	NiNTA Agarose	Qiagen, Hilden, Germany
	Glutathione Sepharose™ HP	Amersham Pharmacia Biotech AB
		Uppsala, Sweden
Concentrators	Vivaspin 5 kDa & 10 kDa	Sartorius, Göttingen, Germany
Dialysis	Spectra Por Membra MWCO 5 kDa & 10 kDa	Roth, Karlsruhe, Germany
	Slide-A-Lyzer® dialysis casettes	Pierce, Rockford, IL, USA
	3.500 MWCO & 10.000 MWCO	
Electrophoresis	Kodak Electrophoresis Documentation-	Eastman Kodak Co., New York, USA
	and Analysis System 120	
	Power Pac 300, PAGE: Mini-PROTEAN 3 Cell	BioRad, München, Germany
	Agarose gel electrophoresis: Mini-Sub Cell GT	
Filtration	Rotilabo 0.22 μm PES	Roth, Karlsruhe, Germany
	Steritop Express™ Plus 0.22 μm	Millipore Co., Billerica, USA
FPLC	Äkta Purifier, Äkta Basic,	Amersham Pharmacia Biotech,
	Frac-100 fractionator, HiLoad	Freiburg, Germany
	HiTrap™ Q HP 1 mL	
	HiTrap™ SP XL 1 mL	
	XK 16 column	
	Phenyl Sepharose™ 4B	
Freezer	-80 °C, MDF-U71V Ultra low temperature	SANYO Electric Co., Ltd,
		Osaka, Japan
	-20 °C freezer	Liebherr, Germany
French Press	French Pressure Cell Press	Thermo Electron Corporation, Germany
Incubation	Infors Multitron HT	Infors Multitron, Einsbach, Germany
	Certomat R	B. Braun Biotech International,
		Melsungen, Germany
Lyophylisation	Christ Alpha 2-4	B. Braun Biotech International,
		Melsungen, Germany
NMR	AVANCE 400, AVANCE 600, AVANCE 700,	Bruker, Karlsruhe, Germany
	AVANCE 800, AVANCE 900,	

Table 3.5: Laboratory equipment

Application	Name	Manufacturer
	Shigemi NMR sample tube, 5 mm	Tokyo, Japan
pH-measurement	PB11 PY-P10	Sartorius, Göttingen, Germany
Pump	Pharmacia LKB-Pump P1	Pharmacia, Uppsala, Sweden
Sonication	Sonoplus GM 2200	Bandelin electronic, Berlin, Germany
	Sonotrode UW 2200	
	Spike TT13	
	Sonoplus GM 3100	
	Sonotrode UW 3100	
	Spike MS 72	
	Ultrasound waterbath Sonorex RK 103 J	
Spectroscopy	UV/VIS-Spectrophotometer	Hewlett-Packard 8453, Böblingen, Germany
	Microplate reader Benchmark	BioRad, Hercules, USA
	96 well plate 'UV star'	Greiner bio-one, Frickenhausen, Germany
	96 well plate 'μClear'	Greiner bio-one, Frickenhausen, Germany
	CD spectrometer Chirascan	Applied Photophysics, Leatherhead, UK
PCR cycler	HYBAID PCR sprint	Perkin Elmer, Wellesley, USA
Waterbath	Polystat	Fisher Bioblock Scientific, Illkirch, France
Water treatment	Ultra Clear	SG Wasseraufbereitung & Regenerierstation GmbH, Barsbüttel, Germany

3.2 Methods

3.2.1 Peptide synthesis

The IQ peptide used in this study is derived from the sequence of the consensus isoleucine-glutamine (IQ) motif of human L-type voltage-dependent calcium channel (GenBank accession number: AAA02501, residues: 1582-1615). Synthetic IQ peptides with sequences as depicted in table Table 3.6 were synthesised in FMOC solid-phase synthesis on a 433A Peptide Synthesizer (Applied Biosystems, Forster City, CA, USA).

Table 3.6: IQ peptide sequences

Peptide	Amino acid sequence
IQ	DEVTVGKFYATFLIQEYFRKFKKRKEQGLVGKPS
IQ-2	GHMDEVTVGKFYATFLIQEYFRKFKKRKEQGLVGKPS

3.2.2 Protein expression & purification

Calmodulin expression

Expression vectors (c.f. Table 3.1.1) were transformed into *E. coli* (c.f. Table 3.1.1) as described on page 107 in Section 10.2. Subsequent production of isotope labelled protein loosely followed the protocol described in [70], with changes as outlined below.

The plate was incubated overnight at 37 °C to obtain single colonies. A colony was picked and was inoculated into 3 mL of LB medium (containing 70 μg/mL kanamycin); the culture was incubated at 37 °C for at least 6 h. 50 mL of M9 minimal medium (see Table 3.3 in this chapter) including trace elements and antibiotic for plasmid selection (70 μg/mL kanamycin) were prepared and inoculated with 50 μL of the 6 h culture and incubated overnight at 37 °C. Further 1000 mL of M9 medium (containing 70 μg/mL kanamycin) were prepared, inoculated with 50 mL of the overnight culture and incubated at 37 °C. Cell growth was monitored regularly by measuring the optical density at 600 nm. Induction proceeded by addition of IPTG (to a final concentration of 1 mM) to the cells growing in the logarithmic phase (OD_{600} 0.8). 6 h after induction cells were harvested by centrifugation (5000g, 30 min, 4 °C). If necessary, the cell pellet was frozen in N_2 (l) and stored at -80 °C.

Calmodulin purification

The purification protocol for CaM used in this thesis, based on early biochemical studies of CaM, exploits the hydrophobic properties of CaM in its calcium bound state. This strategy is preferred, as it abolishes the need to introduce an affinity tag, which may be detrimental to the overall yield.

The bacterial pellet was thawed on ice and resuspended in 40 mL of lysis buffer (50 mM Tris-HCl, pH 7.4, 4 mM $MgCl_2$, 8 mg Lysozyme, 1 tablet of Complete inhibitor mix (*Roche*)). Following a short incubation (10 min, on ice), the cells were lysed by sonication (8 pulses of 20 s, TT13, 10% - 90% intensity). The cell debris was removed by ultracentrifugation (50000g, 30 min).

After centrifugation, EDTA was added to the supernatant (to a total concentration of 2.5 mM) and two sequential Trichloroacetic acid (TCA) precipitations were performed. TCA was added up to a final concentration of 2.84% from a 50% (w/v) stock solution, and the pH of the solution was adjusted to 6.0 with NaOH. Following incubation (1 h) and centrifugation (20000g, 20 min), the supernatant was precipitated again by addition of TCA to a final concentration of 3.4%. After further incubation (1 h), the pellet was collected by centrifugation (20000g, 20 min) and dissolved in 10 mL resuspension buffer (1 M Tris-HCl, pH 7.4, 0.5 mM PMSF). The solution was then dialyzed overnight against 2 L of dialysis buffer (50 mM Tris-HCl, pH 7.4, 0.5 mM PMSF).

The second purification step consisted of two successive phenylsepharose affinity chromatographies. The first one occurred by gravity flow through a phenylsepharose matrix. The sample was incubated in the matrix for 3 h, in a buffer containing 50 mM Tris-HCl, pH 7.4, 5 mM EDTA and 0.5 mM PMSF. The eluate was collected and the $CaCl_2$ concentration adjusted to 15 mM. The second affinity chromatography was done in a Äkta Prime FPLC. The column consisted of 10 mL of phenylsepharose. The sample was loaded and the column washed (50 mM Tris-HCl, pH 7.4, 1 mM $CaCl_2$ and 0.5 mM PMSF). A second wash followed with a buffer identical as the one before, with the exception of the additional presence of 0.5 M NaCl. Finally, CaM bound to the column was eluted (50 mM Tris-HCl, pH 7.4, 5 mM EDTA) and collected in fractions of 0.5 mL. Fractions containing CaM (as assessed by SDS-PAGE) were pooled, dialyzed overnight against distilled water and lyophilized.

For purification of the CaM T146C mutant, all buffers described in this section were saturated with N_2 (g) and supplemented with 1 mM DTT to prevent cysteine-oxidation.

Expression of recombinant IQ peptide

As in the case of synthetic IQ peptide (see page 28), the sequence is derived from consensus isoleucine-glutamine (IQ) motif of human L-type voltage-dependent calcium channel (GenBank accession number: AAA02501, residues: 1582-1615). An expression vector (c.f. Table 3.1.1) was transformed into *E. coli* (c.f. Table 3.1.1) as described on page 107 in Section 10.2. Cell growth, induction and harvesting procedures were identical to the ones used for CaM (see page 29).

3.2.3 Complex reconstitution

Synthetic IQ peptide

Lyophilized CaM (see page 29) was dissolved in NMR buffer (Bis-Tris 20 mM pH 6.8, KCl 150 mM, $CaCl_2$), and lyophilized IQ peptide was dissolved in the same buffer by sonication. A range of dilutions (0.005 mM - 1.0 mM for both CaM and IQ) was prepared for both solutions. Subsequently, the IQ solution was titrated into CaM solution until a molar ratio of 1:1.05 (CaM:IQ) was achieved. Titration was performed at 4 °C, 20 °C and 37 °C, with and without shaking. Further titrations were performed with IQ solutions supplemented with DMSO (1% - 20%) or varying salt concentrations (0 mM, 10 mM, 50 mM, 150 mM, 500 mM).

Recombinant IQ peptide

CaM was expressed and purified as described in Section 3.2.2 on page 29. The tripartite fusion protein containing a hexa-histidine tag, maltose binding protein and IQ peptide, with a TEV protease cleavage site between the last two components, was expressed as described above. The reconsitution protocol below is in part derived from the procedure described in [4].

The bacterial pellet of His-MBP-IQ expressing cells was thawed on ice and resuspended in 40 mL of lysis buffer (50 mM K-HEPES (pH 7.4), 250 mM KCl, 1 mM $CaCl_2$, 0.5 mM PMSF, lysozyme 0.2 mg/mL). Following a short incubation (10 min, on ice), the cells were lysed by sonication (8 pulses of 20 s, TT13, 10% - 90% intensity). The cell debris was removed by ultracentrifugation (50000g, 30 min). The

cell lysate was filtered (0.22 μm) and bound to equilibrated Ni-NTA agarose resin (Qiagen) by incubation (1 h, 4 °C, slow shaker) in binding buffer (50 mM K-HEPES (pH 7.4), 250 mM KCl, 1 mM $CaCl_2$, 0.5 mM PMSF, 10 mM imidazole). The resin was washed with binding buffer containing 20 mM imidazole, and then equilibrated with CaM binding buffer (50 mM K-HEPES (pH 7.4), 100 mM KCl, 1 mM $CaCl_2$, 0.5 mM PMSF, 10 mM imidazole). Lyophilized CaM was dissolved in CaM binding buffer and loaded onto the resin. Following incubation (1 h, 4 °C, slow shaker), the resin was washed once more with CaM binding buffer, before eluting the CaM/IQ complex with CaM binding buffer containing 400 mM imidazole. The eluate was dialyzed against TEV cleavage buffer (50 mM Tris-HCl pH 8.0, 0.5 mM PMSF, 1 mM $CaCl_2$) and then digested with TEV protease (25 μg/mL, room temperature, over night).

Subsequently, the sample was dialysed against loading buffer (20 mM Tris-HCl (pH 8.8), 100 mM KCl, 1 mM $CaCl_2$) and loaded on a HiTrap Sepharose Q HP ion exchange column (GE Healthcare). Following a washing step with the same buffer, the complex was eluted with a linear gradient (20 mL, 1 mL/min) of 0 - 30% elution buffer (20 mM Tris-HCl (pH 8.8), 1 M KCl, 1 mM $CaCl_2$).

Residual affinity tag fragments were removed by incubation with Ni-NTA agarose (Qiagen) yielding a pure CaM/IQ complex in the flow-through. For complex formation involving the CaM T146C mutant, all buffers described in this section were saturated with N_2 (g) and supplemented with 1 mM DTT to prevent oxidation of the free cysteine.

3.2.4 Alignment

In order to measure RDCs and PCS, lanthanides need to be introduced. This section describes methods to introduce a lanthanide into the CaM/IQ complex at a molar ratio close to 1:1 (complex:lanthanide). Lanthanides used in this work are listed in Table 3.7.

Table 3.7: Lanthanides

Lanthanide	Compound	J	Supplier
Dy^{3+}	$DyCl_3 \cdot H_2O$	15/2	Chempur, Karlsruhe, Germany
Er^{3+}	$ErCl_3$ anhydrous	15/2	Aldrich Chem Co, Milwaukee, USA
Tb^{3+}	$TbCl_3$ anhydrous	6	Aldrich Chem Co, Milwaukee, USA
Tm^{3+}	$TmCl_3$ anhydrous	6	Aldrich Chem Co, Milwaukee, USA
Yb^{3+}	$YbCl_3 \cdot H_2O$	7/2	Chempur, Karlsruhe, Germany

Tag based alignment

The CaM T146C mutant, which contains a single, surface accessible cysteine was used in all tagging reactions. EDTA-based lanthanide binding tags and tagging procedures have been described in the literature [55, 56, 71, 72]. CaM/IQ samples however, require lanthanide loading prior to performing the tagging reaction, as the sample buffer contains free calcium that otherwise competes with lanthanides for the EDTA ligand. A number of strategies were evaluated, and the one described in this section was found to yield the most reproducible results.

5x fold molar excess of an EDTA-based tag (compared to the target protein) was solubilized in water. 15x fold molar excess of lanthanide was dissolved in water and added to the tag solution. Following incubation (1 h, room temperature), the pH was adjusted to 6.5 with KOH. After further incubation (15 min, room temperature), an equimolar amount of EDTA (compared to the lanthanide) was added to this solution to bind excess lanthanide. This process can be observed by the disappearance of white $Ln^{3+}(OH^-)_n$ precipitates. Thereafter, excess EDTA was saturated with calcium by adding an equimolar amount of $CaCl_2$ (compared to the EDTA concentration) to the solution. This solution was incubated with the CaM-T146C/IQ complex (2 h, room temperature).

The tagged protein solution was washed twice with NMR buffer in a concentrator and then concentrated to a final protein complex concentration of 1 mM (Vivaspin 6 membrane, MWCO 5000 Da, *VivaScience*).

Internal alignment - the N60D mutant

Bertini and coworkers have been able to engineer a number of calmodulin mutants with useful properties [53, 54]. Most notable of all mutants, is the N60D mutant, which was chosen for this work because the engineered lanthanide binding site has both a high affinity for lanthanides as well as a reduced affinity for calcium. Available protocols for selective lanthanide loading of this site in the literature assume calcium-free CaM as starting material, which is incompatible with complex reconstitution protocol for CaM/IQ established in this work. Strategy considerations are discussed in Chapter 4 on page 41.

Pure CaM/IQ complex in a buffer containing excess calcium (as yielded by the protocol discussed on page 31) was dialyzed against 2 L lanthanide loading buffer (20 mM bis-TrisHCl, pH 6.8, 150 mM KCl, 125 μM $CaCl_2$). This solution was subjected to the NMR sample preparation routine as described below, before recording a ^{1}H-^{15}N-HSQC. Lanthanide was then titrated from a concentrated stock solution to a molar ratio of 1:0.95 (CaM/IQ:lanthanide). The titration was monitored by further ^{1}H-^{15}N-HSQC spectra.

3.2.5 NMR sample preparation

The protein solution was concentrated to a volume of 270 μL (Vivaspin 6 membrane, MWCO 5000 Da, *VivaScience*). Buffered D_2O (150 mM KCl for N60D samples, 150 mM KCl and 20 mM $CaCl_2$ for T146C and wild-type samples) was added to a final concentration of 10% (v/v) before transferring the solution to a Shigemi tube (5 mm, *Shigemi Inc, Allison Park, PA*). The tube was then centrifuged (1 min) in a hand centrifuge (Hettich Zentrifugen, Tuttlingen, Germany) and sealed with its plunger.

3.2.6 NMR measurements

All spectra were recorded at a sample temperature of 303 K using Bruker Avance spectrometers, and standard Bruker pulse sequences, with the exception of the IPAP-HSQC, which was adapted from [73].

3.2.7 Data analysis

Computer programmes

The following table lists programmes that were used in this work.

Table 3.8: NMR Software

Name	Application	Reference
PyMOL	Molecular graphics	[74]
CARA	Assignment	[75]
Echidna	Paramagnetic assignment, $\Delta\chi$ tensor determination	[76]
NMRPipe	Processing	[77]
TopSpin2	Processing, spectrometer control	Bruker, Karlsruhe, Germany
XWin-NMR	Processing, spectrometer control	Bruker, Karlsruhe, Germany
Sparky	Spectal analysis	[78]
NMRdev	RDC & PCS analysis	
Numbat	RDC & PCS analysis	[79]
Pales	RDC analysis	[80, 81]
relax	RDC & PCS analysis, Model-free, N-state model and centre of mass analyses, versatile framework	[82]

Chapter 4

Results

'Commonplace as such experiments have become in our laboratories, I have not yet lost that sense of wonder, and of delight, that this delicate motion should reside in all ordinary things around us, revealing itself only to him who looks for it.'

—Edward M. Purcell, Nobel lecture, 1952

Whilst Purcell was referring to the precession of nuclear spins, intra-molecular and domain motions are an equally enigmatic phenomenon, albeit on a slightly larger scale. This chapter seeks to probe such motions using the tools introduced in Chapter 2 and Chapter 3.

4.1 Feasibility and RDC precision

The two structural models proposed for the CaM/IQ complex differ by an approximately 15° rotation in the second domain (c.f. Fig. 2.2). To assess whether RDCs in conjunction with paramagnetic alignment are an appropriate structural probe to investigate and discriminate these conformations, the maximum observable difference in RDC data was determined.

Several alignment tensors from paramagnetic alignment in another system were considered to quantify the expected degree of alignment. For each of these tensors, theoretical RDCs were calculated for a uniform distribution of unit vectors on a sphere, yielding all possible RDCs that could be obtained for such a tensor. Subsequently, the tensor was rotated through 15° to reflect the difference in the

CaM/IQ structures, and RDCs for the unit vector distribution were recalculated. Subsequently, the change in RDC value between both tensor orientations were calculated.

The maximum observable change was found to be approximately 10 Hz, depending on the input tensor, and occurs for unit vectors rotating through the magical angle leading to a sign change in the corresponding RDCs. The average change in RDC value for a 15° rotation was found to be approximately 1.5 Hz and thus well above the error in RDC measurements.

4.2 Complex reconstitution

4.2.1 Expression & purification

Established protocols in the literature for purification of CaM in complex with the IQ peptide studied in this work are based on co-expression of both interaction partners [4]. This approach however, is undesirable for studies by NMR spectroscopy as isotopic labelling of both proteins increases spectral crowding. Therefore alternative strategies allowing only CaM or the IQ-peptide to be labelled were explored.

Both CaM and a His_6-MBP-IQ fusion construct were expressed independently in *E. coli* and subsequently reconstituted into a 1:1 complex on column. The resulting complex could then be successfully purified using a protocol based on [4], as described in Section 3.2.3 on page 31. This approach proved highly reproducible, although yields were limited by the numerber of purification steps involved in the procedure.

Complex reconstitution using synthetic peptides presents a very efficient strategy that is especially attractive with respect to selective lanthanide loading to achieve alignment. Several synthetic peptides were titrated with purified CaM. This, however, triggered aggregation. A large range of conditions was tested for two different peptides, but the aggregation issue could not be overcome, leaving the recombinant approach as the method of choice.

Further details on all conditions and procedures investigated are given in Section 3.2.3 on page 31.

4.3 Paramagnetic alignment

Residual dipolar couplings (RDCs) provide an elegant tool to address both conformational and dynamical questions. Unlike external alignment, paramagnetic alignment is especially suited to induce RDCs in two-domain proteins such as CaM [54]. Fitting separate alignment tensors to each domain allows the study of relative domain orientations, and the study of domain motions as the reduction of the alignment tensor in the non-aligned domain is an inevitable consequence of such motions.

Paramagnetically aligned samples are most often prepared through the use of lanthanide (Ln^3+) ions. Although a number of lanthanides are available, not all elements are equally suitable to this kind of study. Ideally, the lanthanide of choice should yield a large magnetic susceptibility tensor to reduce measurement errors on RDCs whilst keeping the PRE to a minimum to avoid losing large numbers of resonances due to extensive broadening (c.f. Section 2.3.7). An overview of Ln^{3+} ions and their paramagnetic properties is given in Fig. 4.1. It is apparent that lanthanides with f8 - f13 electronic configurations yield good compromises between the strength of alignment and PRE, whilst Gd^{3+} and Sm^{3+} are not suitable, due to the absent and minute alignment tensors of both lanthanides respectively.

Two different strategies for incorporating Ln^{3+} ions into CaM/IQ samples to achieve paramagnetic alignment were evaluated – tag-based alignment and internal alignment.

4.3.1 EDTA-based tags

Initial trials focused on small organic compounds that are based on the ligand EDTA and can be covalently attached to a cysteine residue through a disulphide bond [55, 56, 71, 72]. As wild-type *Xenopus* CaM is devoid of cysteine residues, serine and threonine residues with approximately 40% surface accessibility were identified as targets for mutagenesis [85]. The accessibility threshold is based on two considerations: surface accessibility of the residue to facilitate attachment of the tag and stronger alignment by minimisation of the mobility of the tag.

When introducing a paramagnetic tag, it is imperative that the tagging reaction goes to completion and all tag species are successfully loaded with the lanthanide

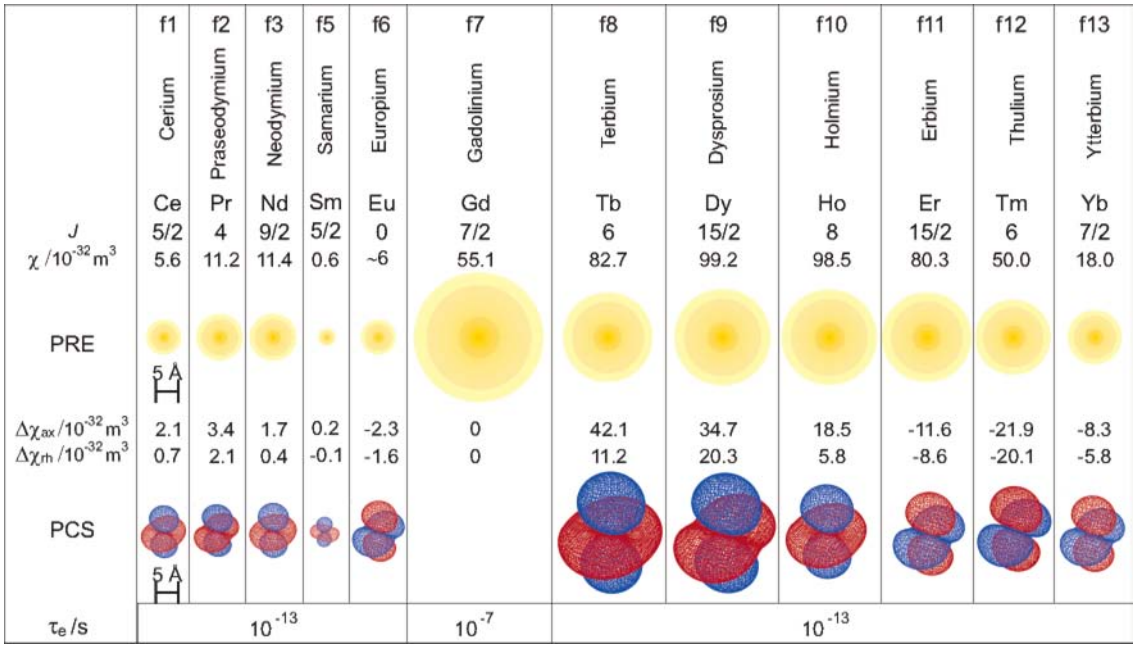

Figure 4.1: Paramagnetic properties of selected Ln^{3+} ions.
The PRE is represented as by a distance plot, where the yellow sphere indicates the distance from the metal ion at which ^{1}H NMR signals of macromolecules with τ_c = 15 ns are broadened by 80 Hz at a field of 18.8 T. Corresponding electronic relaxation rates are given at the bottom. Representative isosurfaces are plotted for PCS by ±5 ppm using tensors reported in [83]. Figure adapted from [84].

of choice. Incomplete tagging, in addition to, or incomplete loading results in a doubling of resonances – one set originating from the isotropic, the other set from the anisotropic species. Whilst, it may be possible to discriminate and correctly assign a proportion of anisotropic signals belonging to residues displaying a large PCS, residues with small PCS yield two overlapping signals that prevent subsequent extraction of accurate RDCs.

Despite extensive testing of tagging and loading conditions, no reproducible protocol for tagging and lanthanide loading of the CaM/IQ complex could be established[1]. Maximum RDCs that could be extracted were typically between 3 Hz and 4 Hz.

[1]For a description of the protocol with the highest success rate, please refer to Section 3.2.4.

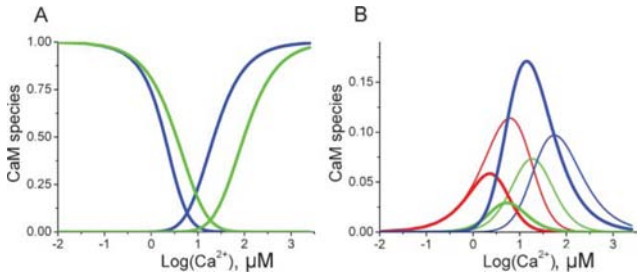

Figure 4.2: Comparison of independent and cooperative Ca^{2+} binding to CaM.
A. Concentration dependant apo and fully saturated CaM species in the absence (green) and presence (blue) of cooperativity.
B. Intermediate CaM species as a function of Ca^{2+} concentration. Ca_1Cam – red, Ca_2CaM – green, Ca_3CaM – blue. Thin curves indicate the absence of cooperativity, thick lines show cooperative Ca^{2+} binding.
Figure adapted from [86].

4.3.2 Internal alignment

In order to achieve stronger alignments than that possible with EDTA based tags, and to overcome problems in complete loading, the possibility of introducing the lanthanide directly into CaM/IQ without using a tag was explored. EF hands in CaM can bind lanthanides, however with wild-type CaM it is difficult to achieve selective loading of a single site. To avoid a mixture of differentially loaded CaM species, Bertini and coworkers have investigated a number of point mutations in CaM's EF hands and their effect on both Ca^{2+} and Ln^{3+} binding. Striking effects can be observed for a particular mutation, N60 → D (N60D) in site II, which showes a highly increased affinity for lanthanides and at the same time a reduced affinity for calcium in this site [53]. Using a step-wise titration protocol adding 3 equivalents of Ca^{2+} followed by one equivalent of Ln^{3+}, obtaining a well defined Ca_3Ln-CaM sample was feasible, with selective lanthanide loading in site II. This has been successfully employed to investigate domain dynamics in free CaM [54].

This procedure, however, is incompatible with the recombinant complex reconstitution discussed above. Tight binding of the IQ peptide to CaM is only given in CaCaM rather than apoCaM, and thus it is necessary to work with an excess of

CaM during the course of the reconstitution. Although the affinity of site II in CaM is higher for lanthanides than for calcium, displacing calcium from site II through addition of Ln^{3+} would require adding an excess of lanthanide due to the unfavourable equilibrium of free calcium in solution. This is undesirable as excess lanthanide, whether in solution or bound to non-canonical sites (e.g. the linker region), causes significant broadening due to PRE. Removing Ca^{2+} from CaM is also not an option as this would lead to dissociation of the CaM/IQ complex by reducing the affinity of CaM for the peptide.

Recent studies of CaM's multi-site calcium binding properties have lead to the construction of mathematical models accounting for equilibria and cooperativity effects involved in this mechanism [86, 87]. These models predict that it is possible to maintain a stable, site specific Ca_3CaM species in the open conformation in the presence of a small reservoir of free calcium (see Fig. 4.2). Therefore it was hypothesised that the calcium saturated species Ca_4CaM/IQ can be selectively transformed into a Ca_3CaM/IQ species by reducing the excess of calcium present after complex reconstitution to precisely the reservoir needed to prevent the equilibrium to shift towards apoCaM. In Ca_3CaM the EF hands still retain their open conformation, thus keeping the complex intact. Given that the calcium reservoir required to maintain this equilibrium is small enough, it is possible to then selectively introduce Ln^{3+} into the high affinity site II without requiring an excess of lanthanide.

This hypothesis was tested by bringing Ca_4CaM/IQ in excess Ca^{2+}, as available through recombinant complex reconstitution, into a range of low [Ca^{2+}] buffers to determine the optimal calcium concentration in the buffer that maintains the complex and allows selective substitution of calcium for lanthanide in site II. Taking the concentrations proposed by Valeyev et. al. [86] as starting points, it was confirmed that it is indeed possible to reduce the complex to Ca_3CaM/IQ using a lower reservoir Ca^{2+} concentration, and subsequently selectively add lanthanide to yield the desired Ca_3LnCaM/IQ species without dissociating the complex or having to cope with excess lanthanide in solution. The addition of lanthanide was monitored by ^{1}H-^{15}N-HSQC as described in Chapter 3.

4.4 Assignment

At 303 K, a ^{1}H-^{15}N HSQC of CaM/IQ shows a single set of sharp signals without doubling of resonances or excessive broadening, suggesting that either a single conformation is present in solution, or that possible motions are taking place in the fast exchange regime (c.f. Fig. 4.3). To address structural and dynamical properties of the complex further, first each resonance must be unequivocally assigned to an individual nucleus within the protein. The ^{1}H-^{15}N HSQC is reminiscent of other other CaM:peptide complexes, but shows distinct differences to these spectra, as well as to free Ca_4CaM. An overlay of two ^{1}H-^{15}N-HSQCs of free and IQ-bound CaM is shown in Fig. 4.3. Unlike for the Ca_4CaM/IQ complex, assignments have been published for a number of CaM species and complexes. Consequently assignments can be conveniently transferred to CaM/IQ in identical or less crowded regions of the spectrum where there is no ambiguity regarding the correct assignment. This is not possible with the majority of the resonances in the central, more crowded region, hence requiring *de novo* assignment of these resonances.

4.4.1 Isotropic CaM/IQ

Triple resonance experiments have been the state of the art with respect to sequential resonance assignment in proteins ever since their first application to CaM in 1990 [88]. These experiments based on one- and two-bond scalar coupling interactions (c.f. Section 2.3.3) present a very attractive alternative to classical strategies based on short-range NOEs. For a comprehensive review on triple resonance experiments, please refer to [89].

A new ^{15}N-^{13}C-labelled CaM/IQ sample was produced to facilitate triple resonance experiments. The 3D experiments subsequently recorded to assign the CaM/IQ complex, and the sequential information that they can provide are summarised in Table 4.1.

Fig. 4.4 illustrates the connectivity information, in this case $^{13}C^{\alpha}_{i} \rightarrow {}^{13}C^{\alpha}_{i-1}$ in a HNCA triple resonance experiment. Complementary experiments are grouped and may help to discern weak resonances. Considering the second stip (R126) in Fig. 4.4, it is not possible from this spectrum alone to determine which of the

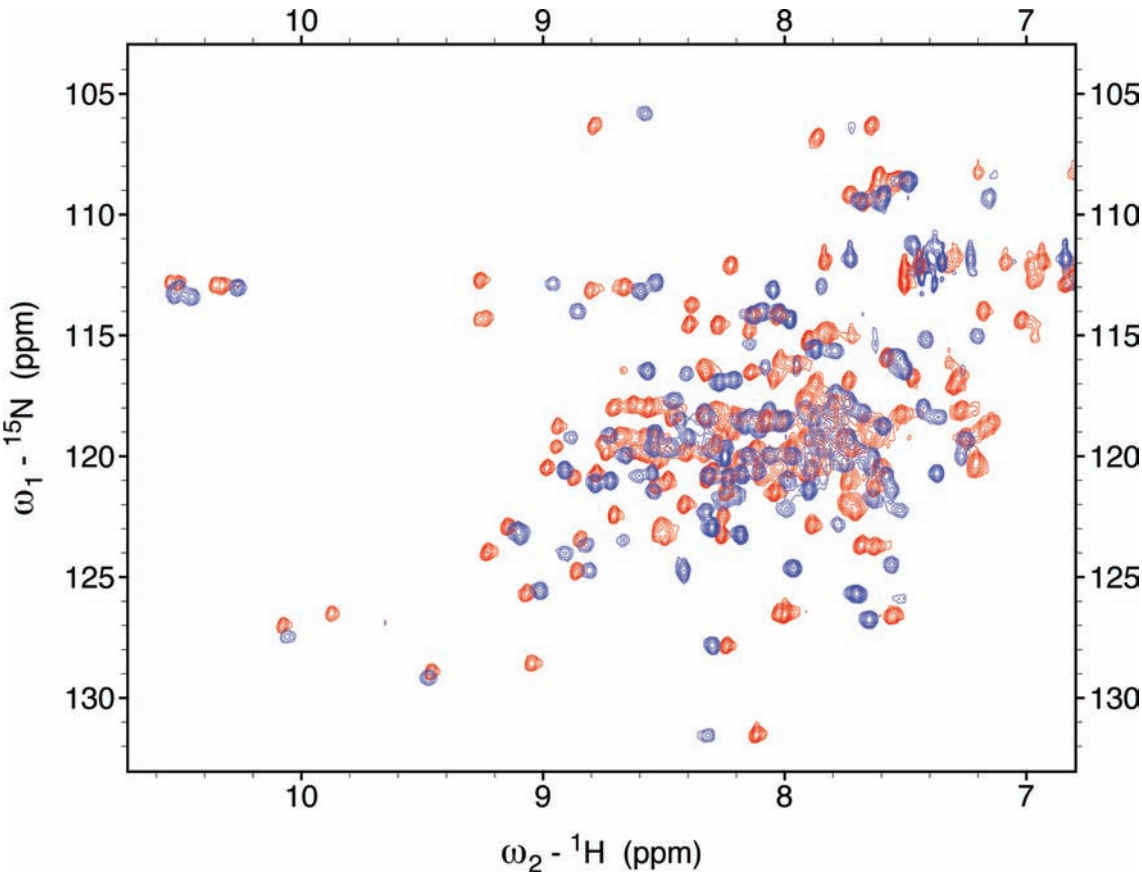

Figure 4.3: ^{1}H-^{15}NHSQC overlay of free Ca_4CaM (blue) and Ca_4CaM/IQ (red).

resonances corresponds to C_α, and which to $C_{\alpha-1}$. This ambiguity is easily resolved by comparison with an HN(CO)CA spectrum. Despite the increased resolution of three dimensional spectra recorded at high field, there may still be spectral overlap, especially in proteins containing repetitive motfis as is the case with CaM's EF hands. Therefore, multiple sets of complementary experiments were recorded, as listed in Table 4.1.

Triple resonance spectra were processed with NMRPipe [77] and assigned using the CARA software package [75]. ^{1}H, ^{15}N and ^{13}C resonance frequencies are listed in in the Appendix.

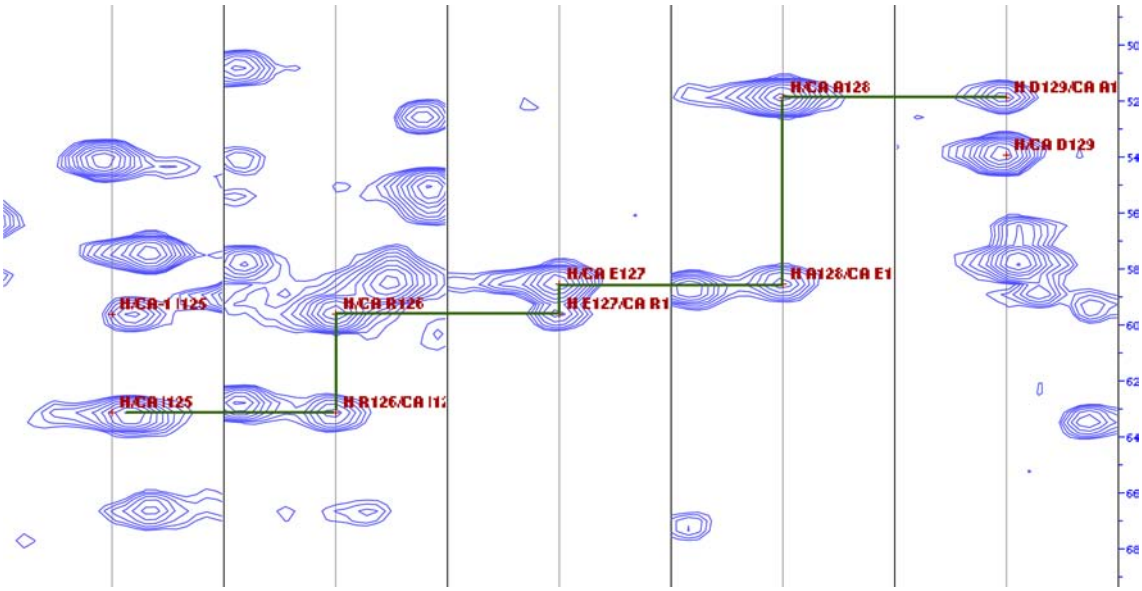

Figure 4.4: Strips extracted from a HNCA triple resonance spectrum of ^{15}N-^{13}C CaM/IQ. Sequential connectivities are illustrated in green, the $^{13}C_{\alpha}$ dimension is indicated on the right and the ^{1}H dimension is along the x-axis.

Table 4.1: Triple resonance experiments conducted on ^{13}C-^{15}N-CaM/IQ

Experiment	Correlation observed
HNCA	$^{1}H_i^N$–$^{15}N_i$–$^{13}C_i^{\alpha}$
	$^{1}H_i^N$–$^{15}N_i$–$^{13}C_{i-1}^{\alpha}$
HN(CO)CA	$^{1}H_i^N$–$^{15}N_i$–$^{13}C_{i-1}^{\alpha}$
HN(CA)CO	$^{1}H_i^N$–$^{15}N_i$–C'$_i$
	$^{1}H_i^N$–$^{15}N_i$–C'$_{i-1}$
HNCO	$^{1}H_i^N$–$^{15}N_i$–C'$_{i-1}$
HNCACB	$^{13}C_i^{\alpha}$–$^{15}N_i$–$^{1}H_i^N$
	$^{13}C_{i-1}^{\alpha}$–$^{15}N_i$–$^{1}H_i^N$
	$^{13}C_i^{\beta}$–$^{15}N_i$–$^{1}H_i^N$
	$^{13}C_{i-1}^{\beta}$–$^{15}N_i$–$^{1}H_i^N$
CBCA(CO)NH	$^{13}C_i^{\alpha}$–$^{15}N_{i+1}$–$^{1}H_{i+1}^N$
	$^{13}C_i^{\beta}$–$^{15}N_{i+1}$–$^{1}H_{i+1}^N$

4.4.2 Paramagnetic samples

HSQC spectra of paramagnetic CaM/IQ samples were notably different from the isotropic wild-type sample (Fig. 4.5). This is to be expected as the the point muta-

tion, N60D or T146C, introduces changes in the local magnetic environemnt in the vicinity of the altered residue, thus affecting chemical shifts of nearby spins (c.f. Section 2.3.2). Much more drastic however, than this localised effect, are the changes brought about by the addition of lanthanides. Spins in close proximity (approx. 10-15 Å) experience a very strong PRE, broadening their singals beyond detection (c.f. Section 2.3.7). This effect is less pronounced with EDTA-based tags, but comes into full effect in the case of the N60D mutant which harbours a lanthanide directly within the N-terminal domain and thus effectively 'bleaching' a large proportion of its resonances.

Further to PRE, PCS cause significant changes in paramagnetic spectra (c.f. Section 2.3.6). Inversely proportional to the distance between the affected nucleus and the lanthanide, this interaction is most drastically observed for nuclei close to the paramagnetic centre, yet beyond the 'blind-zone' caused by PRE – a situation applicable to most observable resonances in the N-terminal domain of Ca_3LnCaM/IQ samples resulting in considerable shifts up to several ppm. Consequently, transferring an existing isotropic assignment to a a paramagnetic spectrum becomes a non-trivial task, as illustrated by Fig. 4.5. This is especially true for crowded regions of the spectrum.

4.4.3 Assignment strategies

PCS are observed as 'diagonal' shifts in HSQC spectra, which can be rationalised by inspecting Equation 2.15 in Chapter 2. Both distance and direction vector from the paramagnetic centre to each of the correlated spins in this experiment, ^{1}H and ^{15}N, are very similar, thus yielding a similar change in chemical shift in each dimension. This crude yet popular method of relating diamagnetic and paramagnetic peaks works well in non-crowded spectra provided the maximum observable PCS is small – smaller than observed inter-peak distances in the diamagnetic spectrum. As is evident from Fig. 4.5, PCS of up to 1 ppm are observed in the case of N60D-CaM, thus limiting the usefulness of this approach due to increasing ambiguities, demanding a more thorough method.

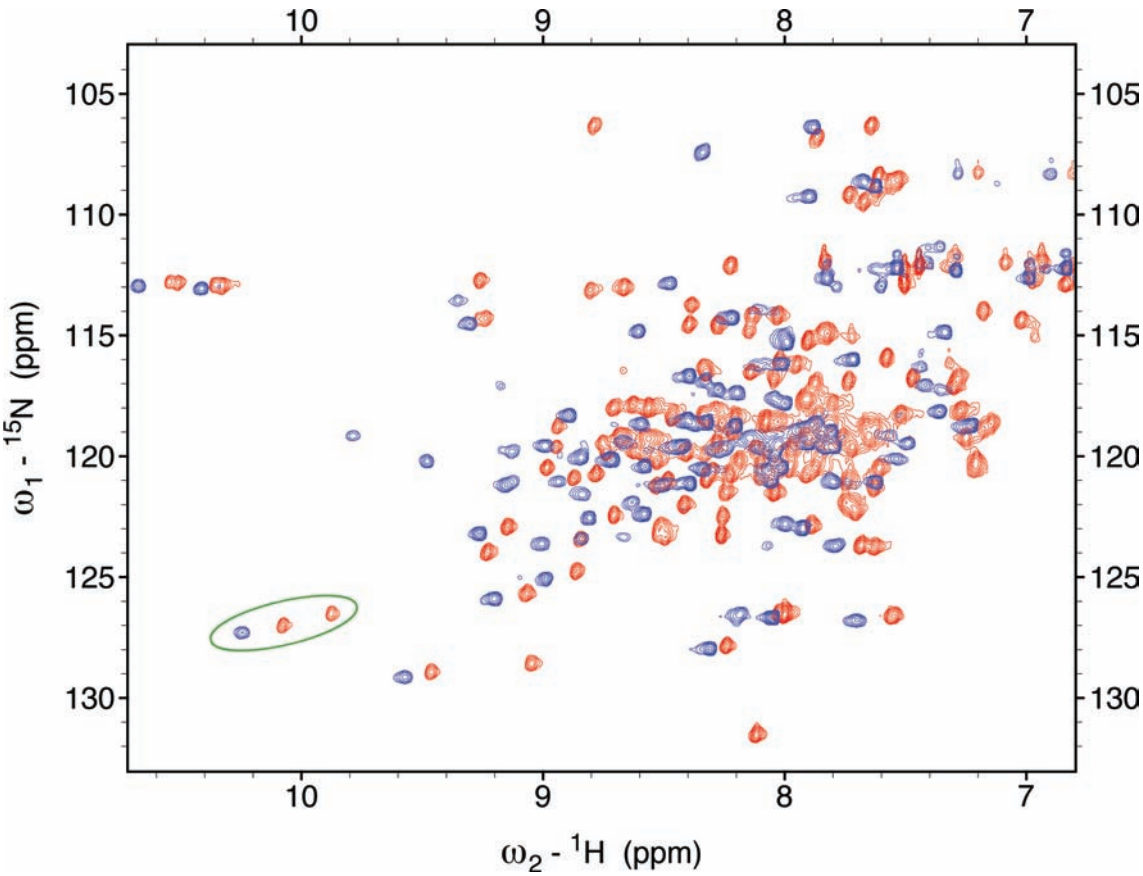

Figure 4.5: ^{1}H-^{15}N HSQC overlay of isotropic (red) and Tb^{3+} paramagnetically aligned (blue) CaM-N60D/IQ. The anisotropic spectrum clearly contains fewer resonances due to PRE bleaching, and remaining resonances are subject to PCS. Highlighted in green is a case in which it is not straight forward to transfer the assignment from the isotropic spectrum to the anisotropic one.

Given knowledge of the $\Delta\chi$ tensor, lanthanide position and three dimensional structure of the protein, it is possible to back-calculate expected PCS using Equation 2.15. If an assignment for the corresponding diamagnetic species is available, back-calculated PCS may then be compared with peaks observed in the paramagnetic spectrum to facilitate assignment. An iterative combination of this approach and the 'diagonal search' method is implemented in the programme *Echidna* [76]. Diamagnetic and paramagnetic spectra are compared to identify unambiguous peak pairs with nearly equal shifts in both dimensions. Although in theory only five in-

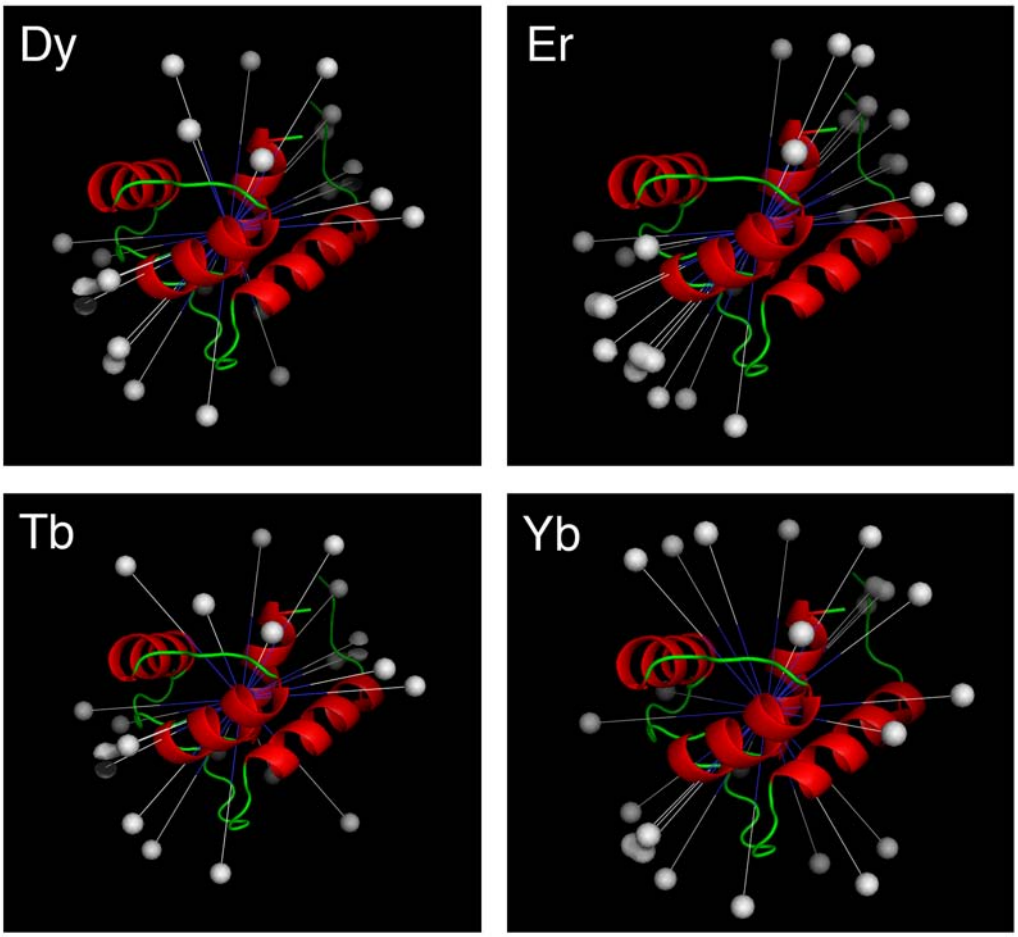

Figure 4.6: Vector distributions of orientations sampled by RDCs in the N-terminal domain of CaM N60D for selected lanthanides. Dy, Tb and Yb show a satisfactory sampling, whilst Er only shows a limited sampling of orientations.

dependent values are required to determine an alignment tensor, in practise eight is a more realistic number. Given at least eight unambiguous assignments, PCS are back-calculated from the $\Delta\chi$ tensor to discriminate ambiguous peak pairs previously identified in the diagonal search. A new tensor is then calculated using the new PCS data, and the cycle is iterated several times.

Echidna was also applied to facilitate assignment of paramagnetic CaM/IQ samples. For the majority of samples, this reproduced and extended the manual assignments made on the basis of 'diagonal search', yet many peaks remained unas-

signed. In addition, *Echidna* is reported to have a approx. 5% mis-assignment rate, which could lead to drastic effects in the case of the CaM N-terminal domain, as PRE leaves only few assignable residues.

Assignment strategies based on triple resonance experiments, as described for isotropic CaM above, are of limited use with paramagnetic samples due to PRE bleaching which abolishes a significant number of connectivities. However, this is only an issue for *de novo* assignments. By comparison to the existing assignment, the additional information gained from triple resonance experiments was shown to resolve the majority of ambiguities and thus enabling successful assignment of paramagnetic samples.

It should be noted that despite the use of triple resonance spectra, in contrast to the C-terminal domain, on average only 15 resonances could be effectively assigned in the N-terminal domain, due to excessive PRE bleaching and the fact that data from the linker region is detrimental to analysis of inter-domain conformations and motions. At first sight this appears to be a very small number, yet there are two important factors to consider. Whilst 15 RDCs from a 75 amino acids domain are insufficient to refine a structure or investigate local dynamics, this number is indeed sufficient to investigate domain conformations. Fitting an alignment tensor requires knowledge of only five independent parameters in theory (c.f. Section 2.3.5), and about eight in practise given that they provide an adequate sampling of orientations in space, i.e. are not parallel or in the same plane. The spatial sampling by assigned residues in the CaM N-terminal domain is visualised in Fig. 4.6. In contrast, almost all resonances can be assigned in the C-terminal domain due to its distance from the parmagnetic centre and thus adequate sampling is easily achieved.

4.5 Data acquisition & signal processing

4.5.1 Acquisition

As the strength of alignment and the maximally observable RDC increases with magnetic field strength (c.f. Equation 2.16 and Equation 2.17 on page 19), all experiments were recorded at the highest field available – 900 MHz. IPAP HSQC

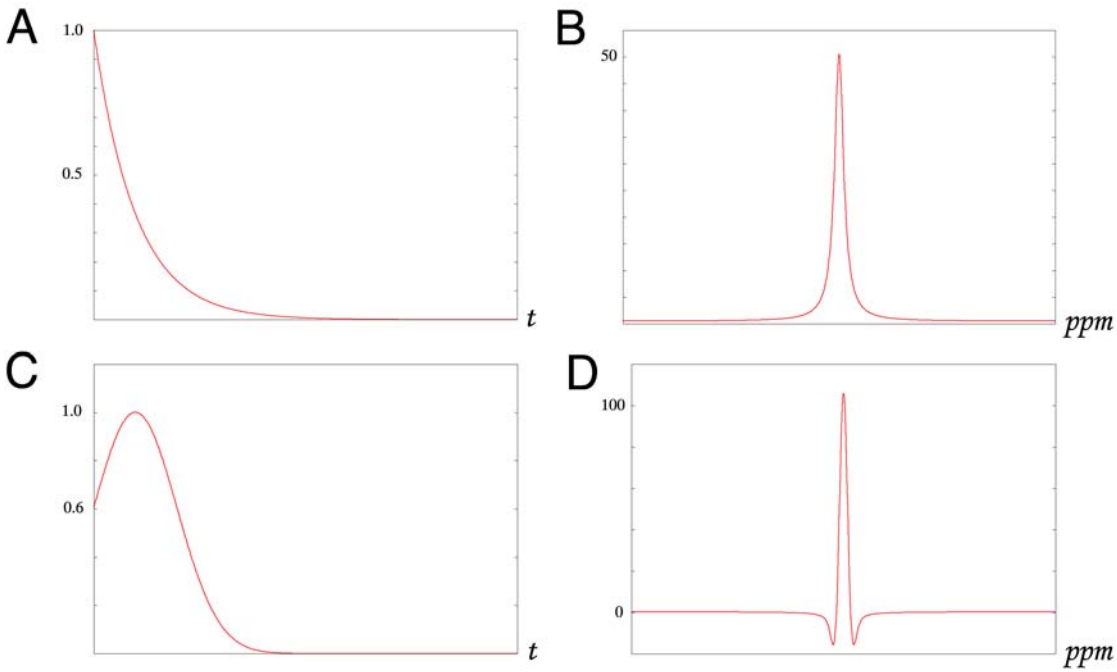

Figure 4.7: Comparison of exponential and Lorentz-to-Gauss window functions. (A) shows an inverse exponential, (B) its Fourier transformation (real part). (C) shows a Lorentz-Gauss function, (D) the real part of its Fourier transformation. Y-axes are labelled with relative units for comparison. The Fourier transformation of the Lorentz-Gauss function shows a higher intensity and smaller line-width at half height, but is accompanied by truncation like artefacts at its base.

experiments were employed to extract RDC – for experimental details please refer to Chapter 3.

4.5.2 Processing

Processing raw spectral data before Fourier transformation has long been standard procedure in NMR. This includes phase correction to yield absorption line shapes and zero filling to increase signal to noise[2]. Largely, standard methods were followed – with one important exception. For sensitivity and resolution enhancement so called window functions are often applied to the FID. Window functions are weighting

[2]For a comprehensive review on all types of NMR data processing see [90].

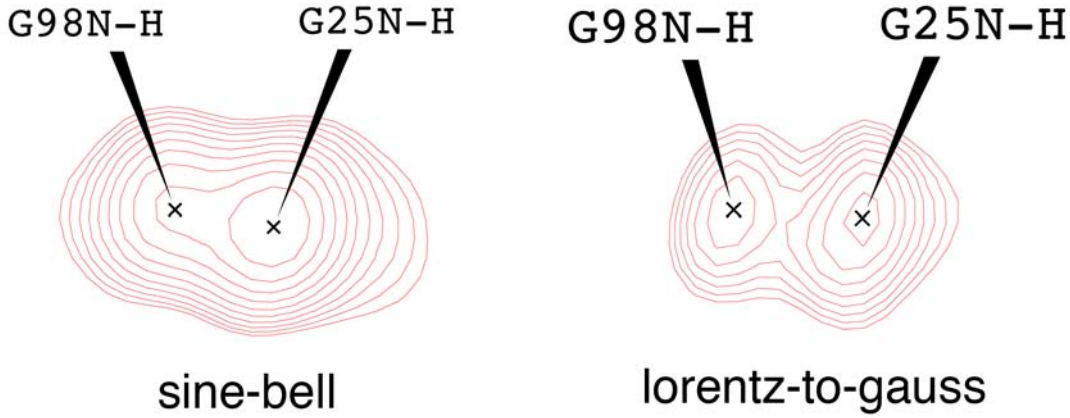

Figure 4.8: Effects of sine-bell and Lorentz-to-Gaussian window functions on the effective resolution in crowded parts of the CaM/IQ spectrum. Using a sine-bell function, the peaks clearly overlap, preventing the extraction of accurate line positions for RDC measurements of G98. The Lorentz-to-Gaussian transformation leads to clear seperation and improvement of the spectrum.

functions multiplied with the FID to optimise the signal after Fourier transformation. Increasing functions (a rising exponential would be the most basic example) are used to increase resolution by compensating for the exponential decay of the NMR signal, thereby reducing line-width in the Fourier transformed spectrum, at the expense of sensitivity (due to the increased contribution of noise towards the end of the FID). The opposite effect, increase in sensitivity at the expense of resolution is achieved by applying an increasing exponential. Consequently, in most cases a compromise is required which has lead to the development of many weighting functions, the most popular representatives are *sine bell* functions.

In the processing scheme of IPAP-HSQC spectra however, a less popular but very beneficial window function was chosen – the Lorentz-to-Gaussian transformation, which alters the line shape from the typical Lorentzian to a Gaussian. Although this amplifies truncation artefacts[3] at the signal base, it reduces line-width and increases sensitivity (Fig. 4.7). Consequently, this method is not suitable when

[3]Often caused by a truncated FID due to experimental constraints, oscillations around the base of the peak in the corresponding spectrum are observed resembling a *sinc* function. Popular counter measures are linear prediction and maximum entropy methods [44, 90].

aiming to calculate peak integrals or determining factors relative to the base plane which is distorted. With regard to extracting line positions however, this method proved to be an invaluable tool that is often underestimated. This applies both to counteracting paramagnetic broadening, as well as seperating individual peaks in crowded regions of a spectrum (Fig. 4.8).

4.6 RDC data

4.6.1 Alignment tensor fits

The first question to be addressed with respect to conformations in the CaM/IQ complex was whether one of the two conformations[4], A or C, is exclusively present in solution (c.f. Fig. 2.2 on page 11). To this end, RDCs were measured in the complex and fitted to the crystal structure using the PALES software package [81]. For both structure A and C, three fits were calculated. (i) fit to the entire structure, (ii) fit to the N-terminal domain only and (iii) fit to the C-terminal domain only. Assuming that the CaM/IQ complex is rigid, all three alignment tensors corresponding to the same structure are expected to be identical. In this case the quality of the fit can clearly discriminate between the two conformations. Conversely, in the presence of inter-domain motions, the alignment tensor of the mobile (C-terminal) domain is expected to be reduced due to motional averaging, however this does not affect the quality of the fits for the N- and C-terminus. The fit to the entire structure however, is expected to deteriorate.

Quality factors and alignment tensor components D_a, R are given in Table 4.2 and Table 4.3 respectively. In addition, the alignment tensors for individual domain fits are visualised in Fig. 4.9 and Fig. 4.10 (pages 54-55).

From the RDC data it is apparent that neither structure A, nor structure C are present exclusively in solution. This is shown both by the lowered quality of the fit to each entire structure, as well as the decrease in tensor size in the un-aligned domain when fitting separate tensors for each domain.

[4]The nomenclature for the these conformations was adapted from the original publication of the structure [4]

Table 4.2: Quality factors (Q) of alignment tensor fits to the N-terminal, C-terminal or both domains of structures A and C.

	Structure A			Structure C		
Ln^{3+}	**N-term**	**C-term**	**both**	**N-term**	**C-term**	**both**
Dy^{3+}	0.260	0.253	0.533	0.387	0.237	0.584
Er^{3+}	0.133	0.245	0.246	0.198	0.268	0.294
Tb^{3+}	0.330	0.280	0.376	0.371	0.202	0.553
Yb^{3+}	0.390	0.390	0.396	0.296	0.346	0.447

Table 4.3: Axial and rhombic components of alignment tensor fits to the N-terminal, C-terminal or both domains of structures A and C.

		Structure A			Structure C		
Ln^{3+}		**N-term**	**C-term**	**both**	**N-term**	**C-term**	**both**
Dy^{3+}	D_a	-8.94e^{-4}	8.02e^{-4}	6.51e^{-4}	-9.48e^{-4}	8.51e^{-4}	6.61e^{-4}
	R	-3.15e^{-4}	4.21e^{-4}	3.42e^{-4}	-2.15e^{-4}	4.79e^{-4}	3.44e^{-4}
Er^{3+}	D_a	4.58e^{-4}	-3.12e^{-4}	-3.10e^{-4}	-3.12e^{-4}	-2.76e^{-4}	-3.16e^{-4}
	R	3.05e^{-4}	-1.46e^{-4}	-1.49e^{-4}	-1.46e^{-4}	-1.35e^{-4}	-1.56e^{-4}
Tb^{3+}	D_a	-7.27e^{-4}	6.69e^{-4}	-6.54e^{-4}	-7.41e^{-4}	-5.91e^{-4}	6.50e^{-4}
	R	-3.97e^{-4}	4.24e^{-4}	-3.60e^{-4}	-4.68e^{-4}	-3.60e^{-4}	4.31e^{-4}
Yb^{3+}	D_a	-2.22e^{-4}	-2.06e^{-4}	-2.10e^{-4}	-2.82e^{-4}	-2.13e^{-4}	-2.04e^{-4}
	R	-3.76e^{-5}	-5.18e^{-5}	-3.77e^{-5}	-8.93e^{-5}	-7.18e^{-5}	-3.95e^{-5}

Consequently, in solution CaM/IQ either exists as a mixture of conformers A and C, or neither structure is an adequate description of the solution state of CaM/IQ. The possibility of structures A and C to be both present in solution in either slow or intermediate exchange can already be excluded at this stage, as the ^{1}H-^{15}N HSQC spectrum clearly shows only one set of signals, and all expected resonances are present in the isotropic spectrum (c.f. Fig. 4.3 on page 44).

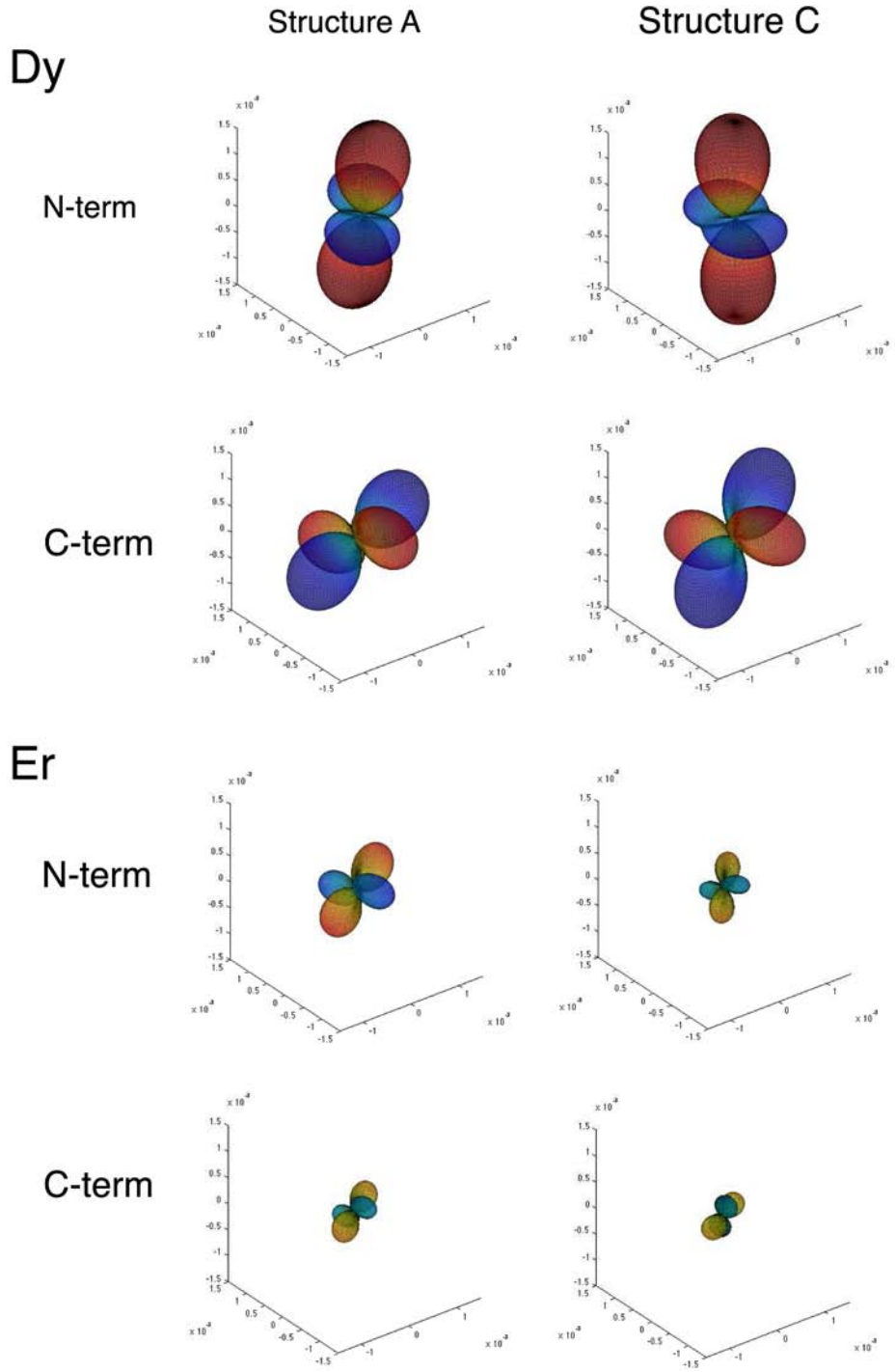

Figure 4.9: Alignment tensors (Dy^{3+} and Er^{3+}). Postive values are plotted in blue-green, negative values in red-yellow.

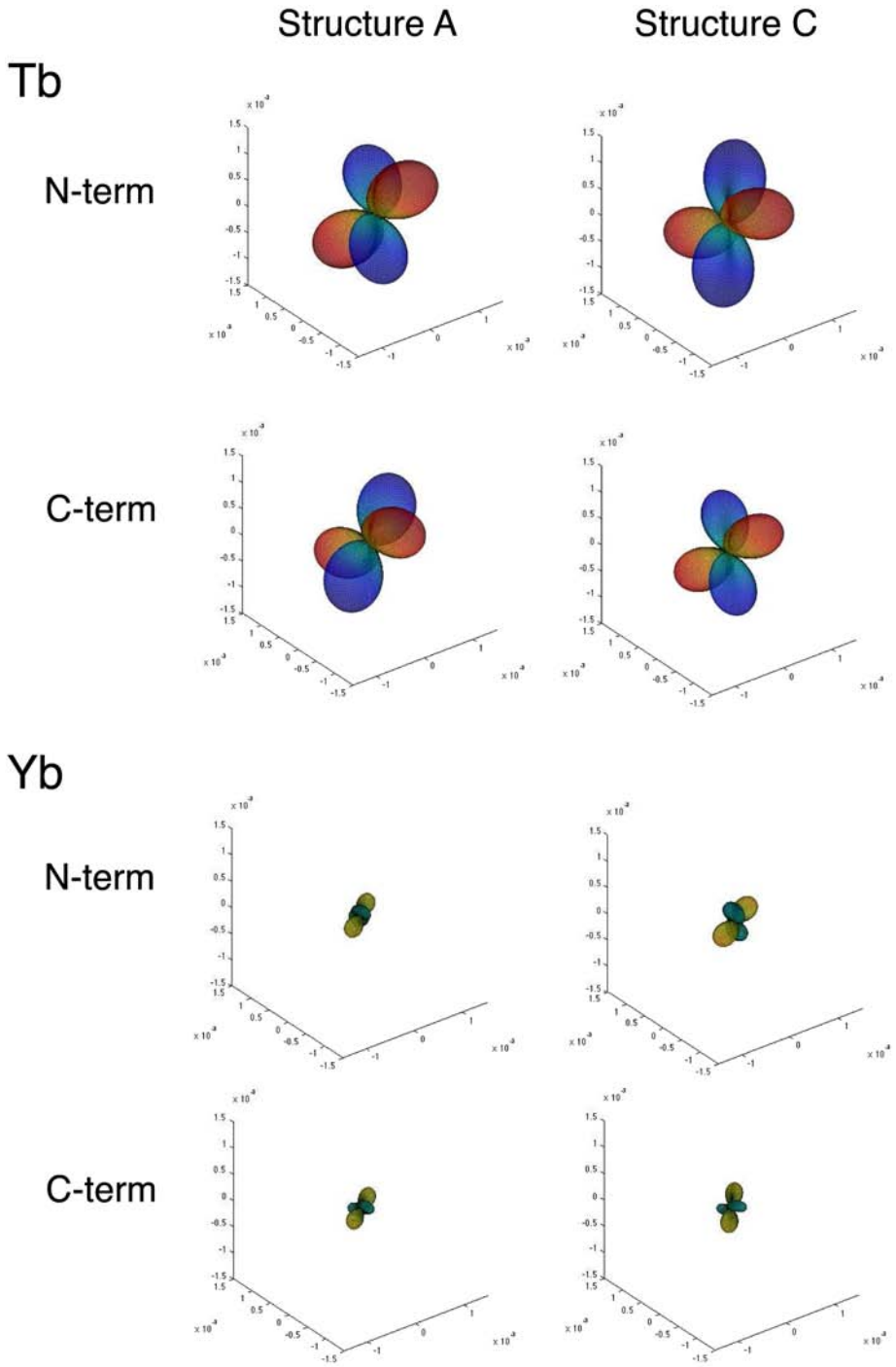

Figure 4.10: Alignment tensors (Tb^{3+} and Yb^{3+}). Positive values are plotted in blue-green, negative values in red-yellow.

4.6.2 Investigating structural populations

As neither structure A nor structure C are exclusively present in solution, the scenario of fast exchange between different populations of structures A and C were investigated. Both structures were assumed to be rigid, such that both domains of a single structure are described by the same alignment tensor.

In order to facilitate comparison of alignment tensor orientations, the aligned N-terminal domains of both structures were superimposed to place the tensors fitted to each structure within the same molecular frame. An experimentally determined alignment tensor for the N-terminal domain of structure A was then used to back-calculate RDCs for the C-terminal domains of both structures. Given two populations in fast exchange in solution, average RDCs for both solutions would be observed. Thus the back-calculated RDCs for both structures were averaged assuming a range of populations for states A and C (10:0, 9:1, ..., 0:10). The population averaged RDCs were then in turn used to fit alignment tensors for both C-terminal domains, which correspond to the alignment tensor that would be observed if the assumption of two rigid structure populations in fast exchange is correct. This process is summarised in Fig. 4.11.

To test the hypothesis, the population averaged tensors were compared to experimentally determined C-terminal alignment tensors. These tensors along with the populations they represent are visualised in Fig. 4.12 - Fig. 4.14. It is apparent, that as the population ratios shift from one conformer to the other, the tensor undergoes a small rotation that corresponds to the relative orientations of structure A and C. However, none of the population averaged tensors resemble the experimentally determined tensors in their orientation. N- and C-terminal tensors derived from experimental data are related through a much larger rotation than is predicted by population averaged tensors for two rigid states. Furthermore, in contrast to the experimentally determined tensors of the C-terminal CaM domain, the back-calculated tensors based on the population model do not undergo a significant reduction in size – tensors corresponding to mixed populations are of the same magnitude as starting and end points which are based on a single structure only.

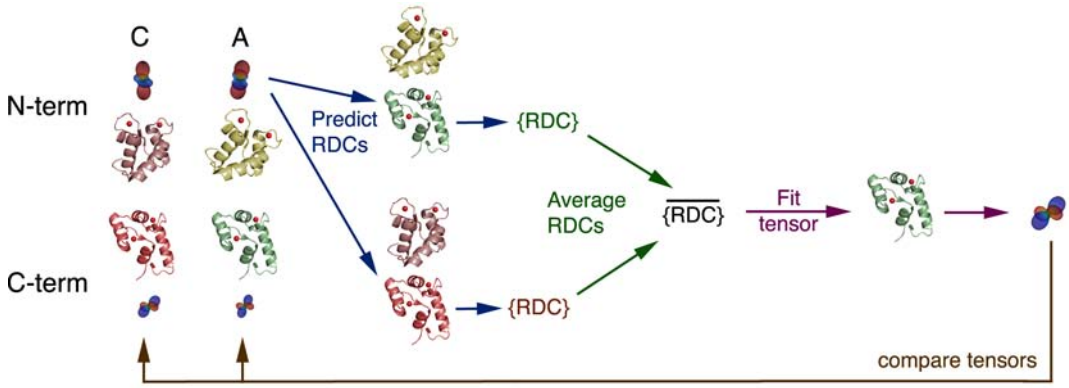

Figure 4.11: Conformational population analysis scheme.

As a two-state population model can not explain the experimental data, it can be concluded that in solution the CaM/IQ complex is not adequately described by the X-ray derived, 2-state structure model.

4.6.3 N-state models

A single discrete conformation of CaM/IQ in solution, different from the conformations proposed in the X-ray structure, can be excluded due to the reduction in alignment tensor magnitude which is indicative of mobility of the unaligned domain (c.f. Table 4.3). A further indication for increased flexibility in this complex stems from the original X-ray data that proposed the two structures A and C [4]. This revealed a third CaM/IQ species, termed structure B, which partially resembled structure A, but was not well defined due to the lack of sufficient density indicating the possibility of mobility even within the crystal.

Obtaining a precise description of molecular mobility is difficult to achieve, owing to the unfavourable data to parameter ratio. Instead, mobility has often been reduced to mere order parameters quantifying the amount of motion [91]. Recently, the development of N-state model theories and its derivatives has opened the possibility for obtaining a much richer description of the motion, including a model of the conformational space that is sampled [54, 66, 68].

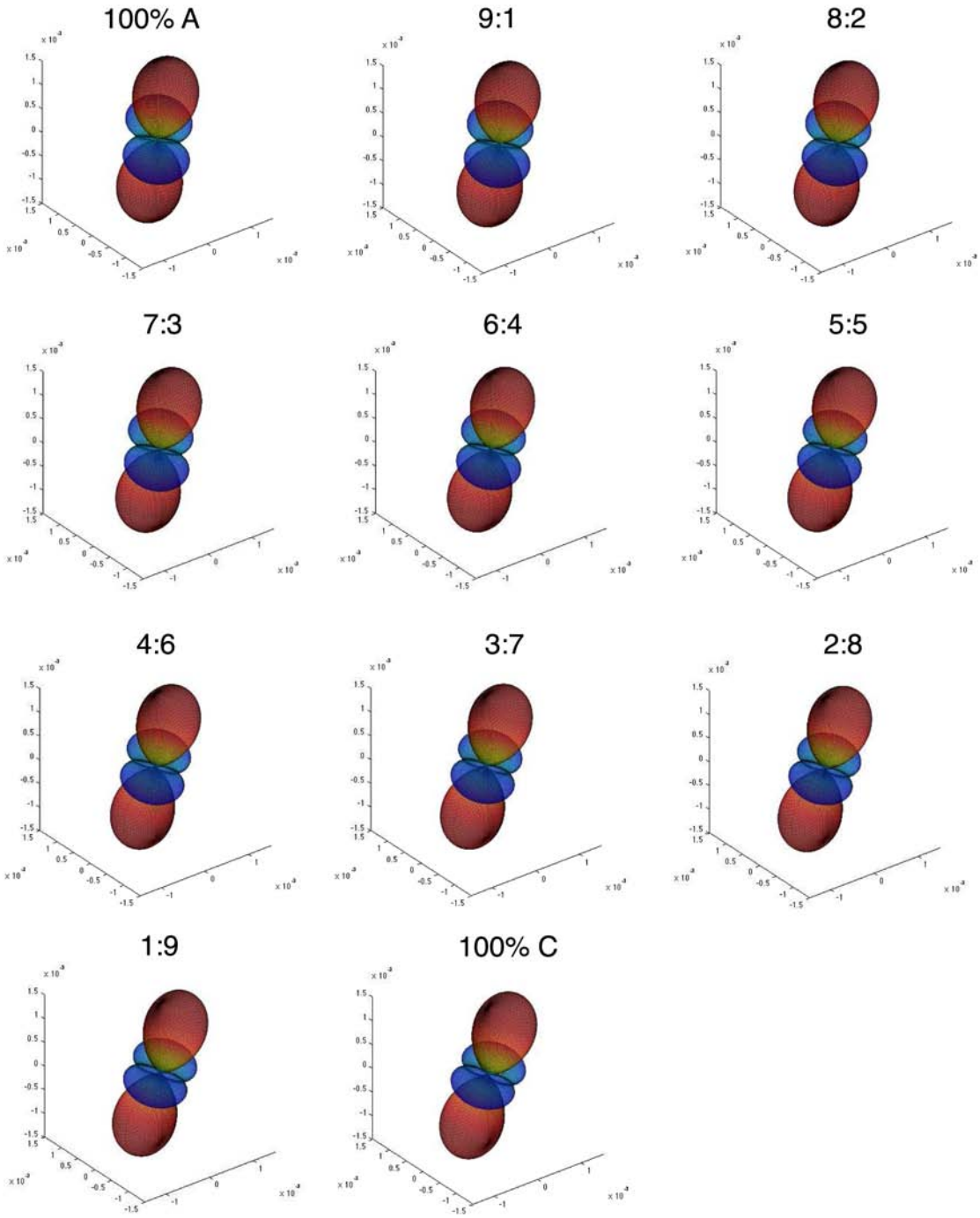

Figure 4.12: Predicted C-terminal alignment tensors for different structure populations (Dy^{3+}).

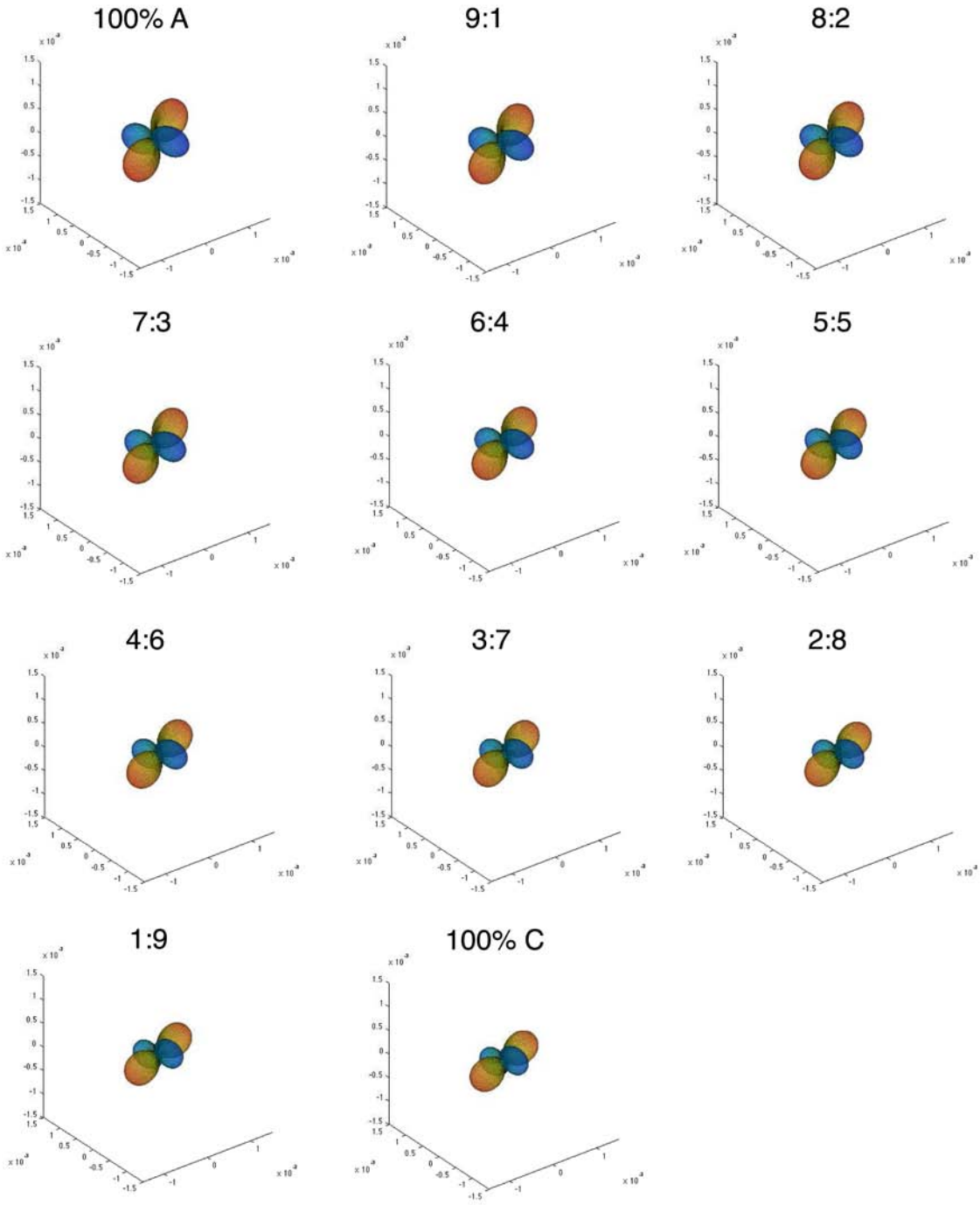

Figure 4.13: Predicted C-terminal alignment tensors for different structure populations (Er^{3+}).

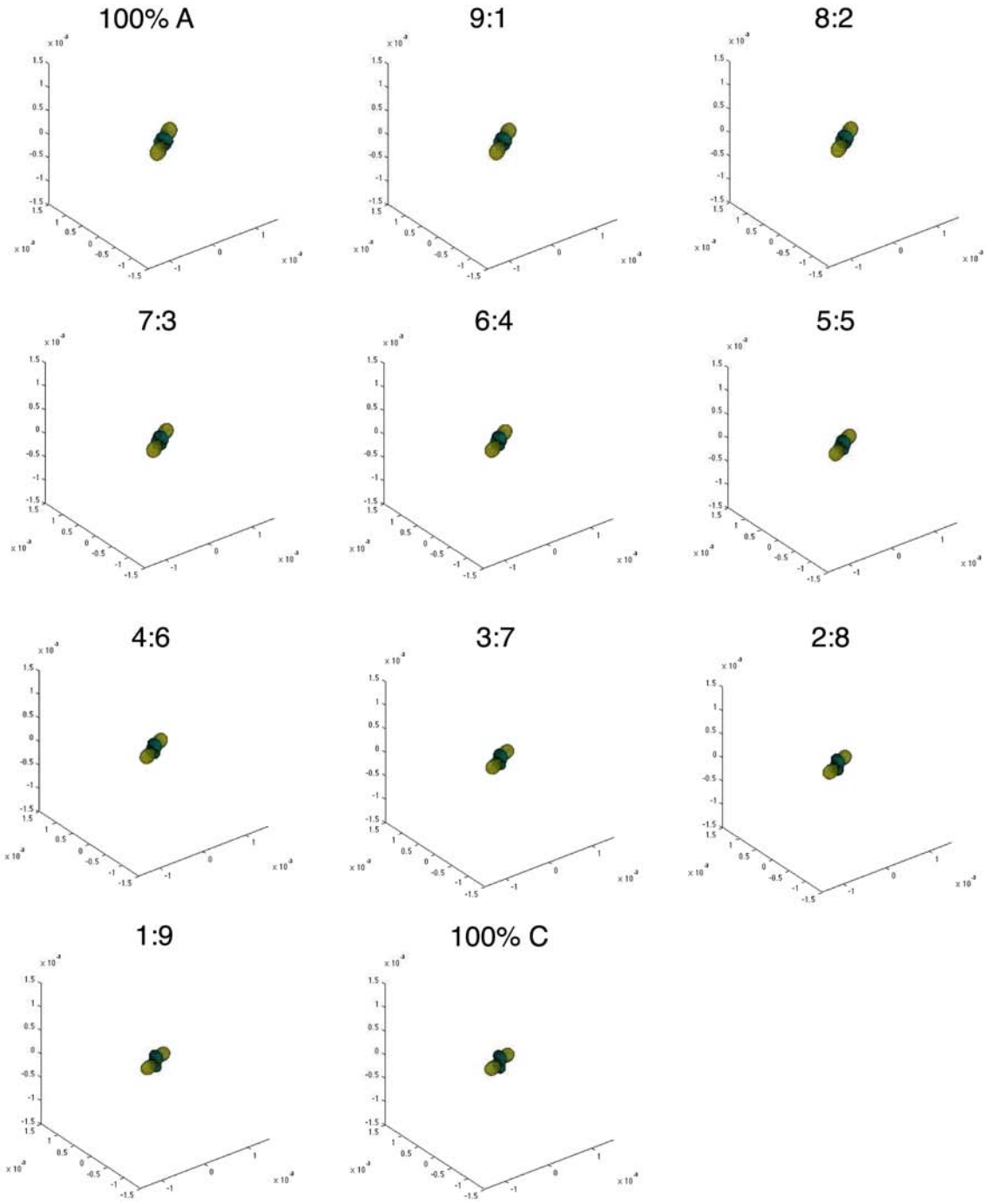

Figure 4.14: Predicted C-terminal alignment tensors for different structure populations (Yb^{3+}).

Using the N-state model framework implemented in the analysis framework 'relax' [92], the available RDC data were utilised in an N-state analysis of the CaM/IQ complex. However, trials to fit a 3-, 4- or 5- state model failed at the minimisation stage.

Chapter 5

Discussion

The CaM/IQ complex presents an interesting model system, both with respect to methods development in the field of protein dynamics, as well as the challenges in understanding the complex regulation of cellular ion channels. To make this system amenable to NMR investigations, a protocol for the reconstitution of isotope labelled samples was established, through which either part of the complex may be labelled according to the question at hand. In addition to a suitable sample, resonance assignment is a further requirement for NMR studies which was fulfilled through the use of triple resonance experiments.

Paramagnetic alignment opens up a very elegant approach to studying inter-domain conformations [68]. Two different strategies, tag-based alignment and intra-molecular alignment were evaluated. Both strategies are subject to inherent drawbacks, weaker alignment in the case of flexible paramagnetic tags, and loss of resonances due to PRE with the use of protein internal lanthanide binding sites. A combination of the strengths of both methods, strong alignment through a rigid tag on the outside of the protein, would be highly desirable. Although currently unavailable, progress has recently been made using tags with different chelators (Fabian Peters, personal communication) or two attachment sites to the protein (Fuyuhiko Inagaki, personal communication). In this work, internal alignment making use of the CaM N60D mutation [53] was chosen over tag-based alignment due to increased sample preparation reproducibility in conjunction with the CaM/IQ complex.

Despite initial complications imposed by the nature of the complex – four metal binding sites and calcium dependant interaction – by delicate manipulation of the

complex equilibrium, it was possible to selectively implant a lanthanide ion into one specific site within the protein whilst maintaining the complexation with the IQ peptide in a reliable manner. The resulting samples showed large PCS, requiring a re-assignment of each spectrum. Different strategies as employed in the field [76] were evaluated, and the use of triple resonance experiments was identified to yield the most reliable results due to the additional information provided over other approaches. This is especially true when dealing with crowded spectra, or when PCS exceed separation of peaks in a spectrum.

Owing to the small distance from the paramagnetic centre to other spins in the protein, many resonances are heavily broadened in internally aligned samples. Before proceeding with an analysis of RDC or PCS restraints it is of critical importance to evaluate the remaining signals and ensure that there are both enough signals – more than the theoretically required five signals – and that the underlying vector orientations provide an adequate sampling in space. Orientations were found to be well sampled in Dy^{3+}, Tb^{3+} and Yb^{3+}, whereas in the case of Er^{3+} the sampling may be adequate but is clearly suboptimal. Whilst short of the possibility to provide local structure information for the use in refinement, once these prerequisites are fulfilled, alignment tensors describing domain orientations and mobilities can be derived.

Close inspection of the resulting alignment tensors revealed that neither structure A nor structure C are exclusively present in solution. A two-site exchange between both structures can also be excluded as the single complete set of signals observed abolishes slow and intermediate exchange regimes, and a fast exchange regime can be excluded from the population averaging analysis. In conjunction with the reduction in alignment tensor size (about 10%) of the non-aligned domain and the different orientation of said tensor with respect to the aligned domain, it is concluded that the system exhibits motion on the ps - ms time-scale as sampled by RDCs. Conformers A and C are likely to be within the conformational space of this motion, but these states alone do not fully describe the behaviour of the CaM/IQ complex in solution.

This may be the reason for the third species found in the original X-ray data of the CaM/IQ complex – structure B – which was present in the unit cell but contained too little density to provide high resolution data. Furthermore, high flexibility of the inter-domain linker and inter-domain motions of CaM also in complex with target

peptides is a re-emerging theme in the literature [17, 23, 24, 68, 93, 94]. In addition to studies focusing on CaM motions, the plethora of different peptide targets and CaM/target structures, encompassing one or more peptides, is an indication of high flexibility in the inter-domain linker [41, 95, 96, 97].

Formulating a complete description of the underlying motion is a difficult task due to both the data to parameter ratio and the different time-scales that may be involved. Nevertheless, structural ensembles as provided by N-state model analyses [54] or their weighted counterpart of maximum allowable probabilities for a set of conformations MAP(R) [68] can provide a good approximation to such a description by defining the conformational space of the motion. Current trials to fit an N-state model to the data presented in this work failed due the large number of local minima encountered. This is an inherent draw-back of the simplex minimisation algorithm employed, which is prone to fail when faced with multiple local minima. A global minimisation algorithm such as simulated annealing or a genetic algorithm may present a more feasible strategy in this case and is worth investigating.

Part III

ChopTools: an *in silico* toolbox to investigate protein modifications

Chapter 6

Background

6.1 Summary

Mass spectrometry (MS) has become one of the most powerful protein-analytical techniques in life sciences and has established an entire new research field: proteomics. Recent developments of novel mass spectrometers and new dissociation methods have pushed the boundary of mass spectrometry capabilities allowing an ever-increasing array of biological questions to be addressed. However, these developments have not been fully matched by corresponding analysis software. Although several well-established and powerful software suites are available (e.g. Mascot [98], Sequest [99]), these are not suitable for addressing a number of scenarios opened up by recent developments. This is especially true in the case of large peptide modifications. Such modifications include post-translational modifications (PTM), most notably Ubiquitin (Ubq) and the Ubiquitin-like modifier (Ubl) families, but also general protein-protein cross-linking. The latter may be endogenous - i.e. disulphide bridges - or engineered through the use of cross-linking reagents, to study the three-dimensional structure of, or the interactions between proteins. To this end, a large number of specialised programmes addressing specific questions have been proposed, each with individual advantages and disadvantages. In contrast, the ChopTools suite is not limited to a specific modifier or affected by the drawbacks of predictive approaches. Instead, it provides a convenient and ubiquitous way of reformulating the question through the generation of fragment library files. This in

turn allows well-established programmes to address the questions at hand thereby bridging the gap between existing analysis schemes and new experimental setups.

6.2 Mass spectrometry

6.2.1 Principles and instrumentation

Mass spectrometry can be described as the smallest scale in the world, not because of the instrument's size, but because of the size of what it weighs - molecules. This is achieved by measuring the mass-to-charge (m/z) ratio of the molecule (or molecule fragments) in an ionised state.

To this end, most mass spectrometers consist of three elementary parts: an ion source, a mass analyzer and a detector. During operation, samples introduced into the machine through a sample inlet are first ionised and subsequently separated in space or in time based on their mass-to-charge ratio. Finally the abundance of ions of each mass-to-charge ratio is measured by a detector (for a schematic, see Fig. 6.1) [100].

Although many different ionisation methods have been proposed to date, with respect to solid and liquid biological samples two techniques predominate: electron-spray ionisation (ESI) [101] and matrix-assisted laser desorption/ionization (MALDI) [102][1]. Both techniques are particularly suitable to large biomolecules due to their high mass ranges and good sensitivities at low sample concentrations. Furthermore, both they achieve reliable vapourization and ionisation without destroyed or severely fragmenting the target molecule. As ions are rather reactive and short-lived, ionisation and subsequent manipulation is conducted in a vacuum.

Upon ionisation, lone ions are accelerated in vacuum towards the mass analyzer by electromagnetic fields. Different types of mass analyzers have been proposed, yet they are all operate on the same underlying principle of separating ions based on

[1] Ionisation methods in mass spectrometry have been reviewed in [103].

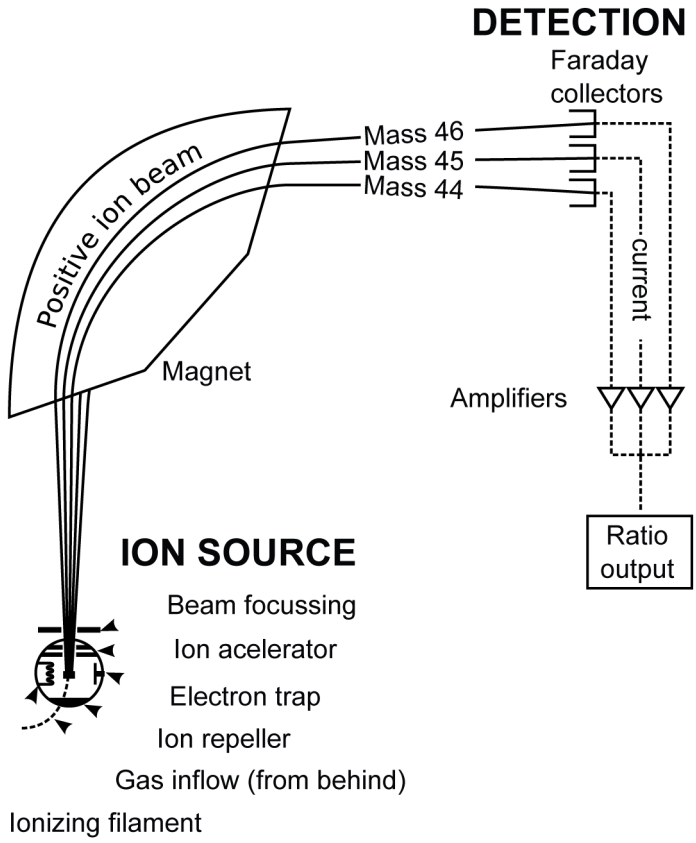

Figure 6.1: Schematic representation of a mass spectrometer
At the ion source, introduced samples are ionised, accelerated and focused. In the second stage, ions in the beam are separated by electromagnetic fields and finally detected as a function of their mass-to-charge ratio in the third stage.

their m/z ratio through electromagnetic fields[2]. The motion of a charged particle in an electromagnetic field in vacuum is governed by two laws:

Newton's second law of motion[3]

$$\boldsymbol{F} = m\boldsymbol{a}$$

[2]for reviews on recent developments in mass analyzers, see [104] and [105] or refer to other comprehensive resources on mass spectrometry in general, such as [100], [106] or [107].

[3]Newton's second law of motion only holds in a non-relativistic case, i.e. at velocities much smaller than the speed of light, an assumption which is valid in the case of mass spectrometers.

where $\boldsymbol{F}$ is the force on the ion, m its mass and $\boldsymbol{a}$ the resulting acceleration, and the Lorentz force law

$$\boldsymbol{F} = Q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$$

where $\boldsymbol{F}$ is the force on the ion, Q the ion charge, $\boldsymbol{E}$ the electric field, $\boldsymbol{v}$ the ion velocity and $\boldsymbol{B}$ the magnetic field. Combining both equations, it immediately follows that

$$(m/Q)\boldsymbol{a} = \boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}$$

where m/Q is the mass-to-charge ratio, which is usually quoted as a dimensionless number m/z, where z is the number of elementary charges (e) on the ion ($z = Q/e$). This differential equation is the classic equation of motion for charged particles, and together with the particle's initial conditions, it completely determines the particle's motion in space.

Modern mass spectrometry setups often utilize more than a single mechanism for separating ions to increase resolution and sensitivity. Tandem mass spectrometry for instance, also known as MS/MS, couples two mass analyzer stages, often separated by an additional fragmentation stage. Other common setups employ non-MS techniques before the ionisation stage, such as liquid chromatography, known as LC-MS when coupled to MS, or even separation of the input mixture by SDS-PAGE before application to a coupled MS setup.

6.2.2 Applications in proteomics

In addition to applications in structure determination, a tool valued by many synthetic chemists, and the field of metabolomics, MS has become one of the most powerful protein-analytical techniques in life sciences and has established an entire new research field: proteomics. It is the large-scale study of proteins, particularly their structures and functions, but also includes modifications made to particular sets of proteins [108, 109]. Besides high-throughput identification of proteins from highly complex mixtures, dynamics, expression profiles, the study of protein modifications and protein-protein interactions has been nurturing the demand for new methodological developments [110, 111].

The most prevalent approach is of a bottom-up nature: a typical work-flow comprises digestion of proteins with specific endoproteinases, separation of the generated by peptides by (nano) liquid chromatography (LC), ionization of the eluting peptides in the mass spectrometer, selection and subsequent fragmentation (sequencing) of a peptide within the mass spectrometer. Finally, searching the fragment spectra against a database by using so-called search engines allows identification of the corresponding protein.

6.3 Large peptide modifications

Further to many other applications, ranging from mere identification to quantitative metabolomics, MS has also become increasingly valuable in the analysis of protein modifications. During its lifetime, a protein may undergo a wide variety of alterations that have critical effects on its function. For example during cell signaling many enzymes and structural proteins can undergo phosphorylation. The addition of a phosphate to particular amino acid, such as serine or threonine, or more rarely tyrosine, causes a protein to become a target for binding or interacting with a distinct set of other proteins recognizing the phosphorylated domain [112].

6.3.1 Ubiquitin and Ubiquitin-like modifiers

In addition to small modifications, proteins can also be subject to alteration through large peptide based modifications. The most well-known example is ubiquitination, a process during which an activated peptide originating from the small protein Ubiquitin is added to a target protein by an E3 Ubiquitin ligase [113]. The conjugation involves an isopeptide bond between the carboxyl group of the modifier and the -amino group of a lysine residue within the targets. Attachment of Ubls to specific targets involves an enzymatic cascade: first the Ubls are processed to expose their C-terminal di-Glycine motif. The matured Ubl is then transferred to its target via a cascade of E1 (activating), E2 (conjugating) and E3 (ligase) enzymes. In many cellular processes, such modified proteins are thereby labelled for degradation. Besides

this function, ubiquitination also controls the stability, function, and intracellular localization of a wide variety of proteins.

Ubiquitin is only one of many examples of peptide based modifications. Ubiquitin-like modifiers (Ubl) have been attracting more and more attention over the last years [114], especially their most prominent representative, SUMO, which has also shown to be involved in a wide range of cellular processes [115]. Using mass spectrometry it is possible to unambiguously both identify modification targets as well pinpoint modification sites of the modifiers involved in these processes.

6.3.2 Cross-linking

In the context of mass spectrometry, a further type of large peptide 'modification' is often investigated - protein-protein cross-linking. Although technically this does not constitute a modification in the true sense of the word, on the molecular level, and therefore, with respect to mass spectrometry, the situation is very similar to Ubls[4]. Protein cross-links may be of two predominant types, either endogenous or engineered. Endogenous cross-links occur as internal disulphide bridges that many proteins contain, and can provide valuable information about the tertiary structure of a protein. More common yet, are artificially introduced cross-links which are put to good use in many a proteomic analysis. Through the use of appropriate reagents, it is possible to link interacting proteins through cysteine or lysine residues under many experimental conditions, including *in vivo* scenarios. In a subsequent mass spectrometric analysis it is then possible to unambiguously identify the direct interaction partners of a given protein [116, 117].

[4]Subjecting cross-linked complexes to proteolytic cleavage yields cross-linked fragments which may be viewed as a fragment of the protein of interest with a large peptide modification conjugated to it.

6.4 Existing software

6.4.1 Search engines

Following the typical work-flow of a MS based proteomics experiment as outlined above, one obtains the molecular weight of the fragment ions originating from the sample (MS/MS experiment). To identify the corresponding proteins, the determined molecular masses of single peptides and the respective masses of the fragment ions are searched for matches in databases with the help of so-called search engines (Mascot [98], Sequest [99], ProteinPilot [118], ProteinProspector [119], PROWL [120]). Search engines compare the experimentally determined mass-spectrometric values with values generated *in silico*, derived from the FASTA sequences of the proteins in the database. Each search engine applies a different specific algorithm to score (i.e. to weight) the results and to obtain a list of proteins present in the sample.

Search engines also introduce the opportunity to search for (post-translational) modifications at various amino-acid positions of the protein sequence by adding the exact mass of the modification to that of any given amino acid. The position of the (post-translational) modification within the protein is determined by highly accurate mass analysis of the modified peptide sequence and, in addition, sequence analysis of the modified peptide (fragmentation pattern) that pinpoints the modified amino acid.

However, whilst search engines are capable of taking experimental conditions (e.g. proteases used, experimental mass deviation, modifications, instrument used) into account, they rely solely on databases containing putative peptide sequences for identification.

Search engines and conjugated modifiers

Commonly used databases for MS-based proteomic research contain protein sequences in FASTA format [121, 122] derived from annotated gene sequences and/or reversed protein sequences in order to determine the rate of false positives in MS-based database search [123]. A user who wishes to identify conjugated proteins of

two proteins or part of it currently has two choices: either the user has to manually apply a fused linear sequence to a database he/she searches against, or if not applying a fused FASTA sequence, identify two protein sequences in (by) normal database search (if protein sequences are available in the database). Manual generation of a fused protein sequence is only straightforward if the site(s) of conjugation (e.g. the N - and/or C terminus) is/are defined. However, when a user wishes to identify a hitherto unknown conjugation site between a protein that might be modified at any amino acid position within the sequence of a second protein, all possible conjugation sites in all possible combinations of the proteins have to be covered in order to generate a modified FASTA file.

Such a scenario is relevant in the MS-based analysis of I) protein-protein crosslinks where two proteins are conjugated naturally or by chemical crosslinking [124] or II) such post-translational modifications where the modification is a protein itself as it is the case for ubiquitin (Ub) and ubiquitin-like (Ubl) polypeptides [125], as described above.

When such modified proteins are being analysed in an MS-based proteomic workflow, enzymatic hydrolysis generates e.g. tryptic fragments of the modified protein still carrying a tryptic fragment derived from the modifier. In cases where the peptide derived from the modifier is relatively small the entire mass of the remaining modification is relatively small and can be added for MS-based database searching to e.g. any lysine residue of the FASTA sequence in the database for identification by using the available search engines. For instance in the case of modification with ubiquitin *per se* and SUMO in yeast [125], only GG or EQIGG after complete endoproteolic cleavage remains attached to a lysine residue of the modified protein. However, in the case of Ubiquitin modification *per se* and Ubl in yeast, incomplete enzymatic cleavage generates much larger endoproteolytic fragments.

Search engines and protein-protein cross-links

The situation is similar in the analysis of protein-protein crosslinks, but is dramatically challenged by the fact that in contrast to Ub and Ubl modifiers where the modifier is always attached with its C-terminus to a Lysine-residue in the modified protein the actual conjugation sites within both the modifier and modified protein

sequences are not known. Consequently, any combination between endoproteolytic fragments containing acceptor sites (e.g. Lysine and Cystein residues) derived from both proteins is possible. Therefore unambiguous identification of larger or unknown modifications generated from two conjugated peptides requires the determination of the exact mass of conjugated product and the complete sequence information derived from both the peptides that are conjugated. Such branched-(conjugated) peptide sequences that would allow an MS-based database search are unavailable in the databases that common search tools make use of. Moreover, incomplete enzymatic cleavage both within the target peptide and within the modifier peptide can lead to an even larger variety of theoretically observable fragments.

6.4.2 Alternative software tools

Various computational strategies have been proposed to overcome this problem of which only relatively few are available freely as web-server versions. These typically try to predict possible modification sites for specific Ubls (e.g. SUMOsp [126], SUMOplot[5], FindSUMO [127], Ublfinder [128]) or to match possible (crosslink) modifications with experimental MS data provided by the user (SUMmOn [129], MS-Bridge [130], CLPM [131], xQuest [132]). Site-prediction tools, which are often restricted to specific modifications, clearly suffer from the fact that proteins could be modified on residues at positions that do not conform to known consensus sites. Matching engines, on the other hand, are usually restricted to certain types of spectra (MS, MS/MS, LC-MS/MS) and may prove difficult for first-time users.

6.5 ChopTools

The following chapters describe a novel software tool called ChopTools that can assist in the identification of large modifications, both protein-protein and Ubl-crosslinks. It does not require any a priori information about consensus sites, and is neither restricted to certain crosslinking reagents, nor to certain Ubls, nor to

[5]Not published through peer review, available online at `http://www.abgent.com/tools/sumoplot_login`

certain types of mass spectra. Instead, ChopTools provides a fast and easy way to generate a new, modified FASTA sequence or an entire database for use with existing, well-established MS-dependent database search tools that users of modern mass spectrometers are already familiar with.

Chapter 7

Implementation

Available at `http://choptools.gwdg.de`

7.1 Strategy

In order to overcome the limitations of prediction methods and pattern search tools, whilst at the same time being ubiquitously applicable to any peptide based modification, the strategy of generating a fragment library was pursued (c.f. Chapter 6). A similar strategy has first been discussed in conjunction with the analysis of protein-protein crosslinking MS data [133], but to date no software is publicly available to facilitate the generation of the required FASTA library files, and a user wishing to pursue this strategy would have to manually generate the required libraries. ChopTools is able to create a library of all possible and any number of modifications and fragmentation scenarios at any amino-acid residues of a modified protein, for any modifier peptide derived from the modifier. The novel FASTA sequence library serves as a basis for existing algorithms that are implemented in commercially available search engines such as MASCOT or Sequest, so that any (user-defined) modifications at any amino-acid residues are identified by database search subsequent to the MS and MS/MS experiments in the output format of the search engines used. In addition, m/z mass lists corresponding to all fragments in the library file may be generated for use as an inclusion list in the spectrometer, or to aid in the annotation of spectra.

7.2 Workflow

In general, ChopTools generates modified FASTA [121, 122] sequences of proteins of interest. The modified FASTA file contains at every position of a putatively modified acceptor site within the modified protein a sequence of the modifier in various forms, i.e with zero, one, two, etc. miscleavages. The entirely modified sequence is added in FASTA format to an available database which can then be searched by common MS search engines in order to identify unambiguously the modified peptide subsequent to MS and MS/MS analyses.

In detail, ChopTools allows the user to generate a library of all theoretically possible peptide fragments. Sequences of target proteins of interest can be submitted either as FASTA files or, alternatively, in the single-letter amino-acid code. Two modules are available: ChopNStitch is tailored to protein-protein crosslinks, whilst ChopNSpice allows ubiquitin-like modifications to be investigated.

7.2.1 Chop'N'Stitch

ChopNStich generates crosslinked fragments from the input sequences it is provided with. Additionally, one can specify the precise mass of a linking reagent when generating inclusion lists. In the case of ChopNSpice, the kind of (post-translational) modification to be investigated can be chosen from a large list of known (Ub and Ubl) modifiers (40 modifiers from 6 species)[1], or submitted as single-letter amino-acid code (e.g. in the case of novel or any user-defined modifiers).

Fig. 7.1 illustrates the 'chop' and 'stitch' steps implemented in the ChopNStitch module to investigate protein-protein crosslinks. The first step ('Chop', see Fig. 7.1) performed is an *in silico* cleavage of all input sequences, using the enzyme or enzyme combination that the user has specified (as previously used in his/her experiment)[2]. This cleavage step generates a library of all possible digestion fragments. Upper and lower boundaries for miscleavages, within both target proteins can be fine-tuned

[1]For a detailed list of all modifiers and their sequences implemented in the current version of ChopNSpice, please refer to Table B.2 in the Appendix.

[2]For a detailed list of all proteases and their corresponding cleavage rules that are implemented in ChopNStich and ChopNSpice, please refer to Table B.1 in the Appendix.

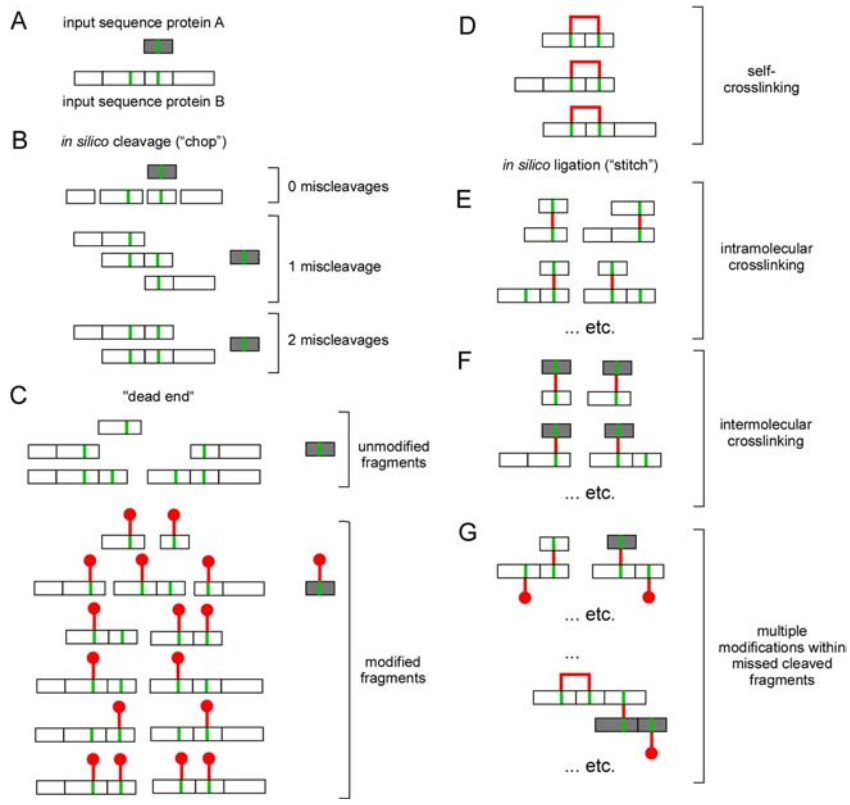

Figure 7.1: ChopNStitch workflow

A. Input data: one target (B) and one modifier sequence (A) are provided. These may contain protease sites (black) and crosslinking sites (green). **B.** 'Chop': *in silico* digest of all input sequences (target, modifier). Upper and lower boundaries for allowed missed cleavages can be set independently for both target and modifier. The user can choose to exclude unmodified fragments from the output. **C.** Addition of the linker reagent: fragments that contain linkage sites (e.g. Cys or Lys), are fused to dead end linker reagents. Fragments containing more than one site, may be modified multiple times. The user can choose to exclude dead end fragments from the output. **D.** 'Self-linking': fragments containing several linkage sites can be linked to themselves. The user may choose to exclude self-linked fragments from the output. **E.** 'Stitch' within a sequence: all combinations of fragments from the target sequence are ligated. **F.** 'Stitch' across sequences: all combinations of fragments from sequence A fused to sequence B are generated. **G.** Combinations of different stitches are also calculated and can include any combination of ligation, self-linkage and dead-end linkages within the same fusion peptide.

independently (e.g. up to four miscleavages in the modified protein and two in the modifier. If the user wants to achieve maximum coverage of all possible peptide modifications, or wants to investigate crosslinking efficiency, he/she can choose to include dead end linker reagents, which are attached to fragments containing the linkage site of interest. Subsequently, *in silico* crosslinking ('stitch') takes place on three levels: fragments containing more than one putative link site may be linked to themselves ('self-link'), all combinations of modifier fragments linked to fragments of the modifier ('intra-link') and all combinations of the modified protein fragments ligated to modifier fragments. Depending on the user settings, combinations of different linkage levels are also generated if the fragments contain a sufficient number of modification sites, e.g. fragments of the same protein linked to each other, but at the same time containing dead-end links and/or self-links attached to further sites.

7.2.2 Chop'N'Spice

Fig. 7.2 illustrates the 'chop' and 'spice' steps of the ChopNSpice module. Analogous to the ChopNStitch module, the input sequences are subjected to *in silico* digestion ('chop'), generating a library of all possible digestion fragments. Upper and lower boundaries for miscleavages, both within the target proteins and within the modifier (Ub and Ubl), can be fine-tuned independently.

The second step ('Spice', Fig. 7.2) identifies those peptides that contain putative acceptor sites (as specified by the user, e.g. any lysine residue) and attaches the C-terminus of the modifier to that particular peptide[3]. If acceptor sites are endoproteinase cleavages sites (as in the case of Lys-C or trypsin) then only those peptides that harbour lysine as a missed cleavage site are identified to which the modifier is attached. If several modifier peptides were generated during the endoproteinase cleavage step, all different modifier peptides will be added independently. Optionally, multiple modifier sequences and different combinations thereof may also occur within a single peptide fragment. Each module allows very fine control over all the above parameters. For example, upper and lower boundaries for miscleavages can

[3]Of note, during the testing phase of the programme it was experimentally observed that branched peptides fragment in a manner like that of linear fused peptides in MS/MS [134].

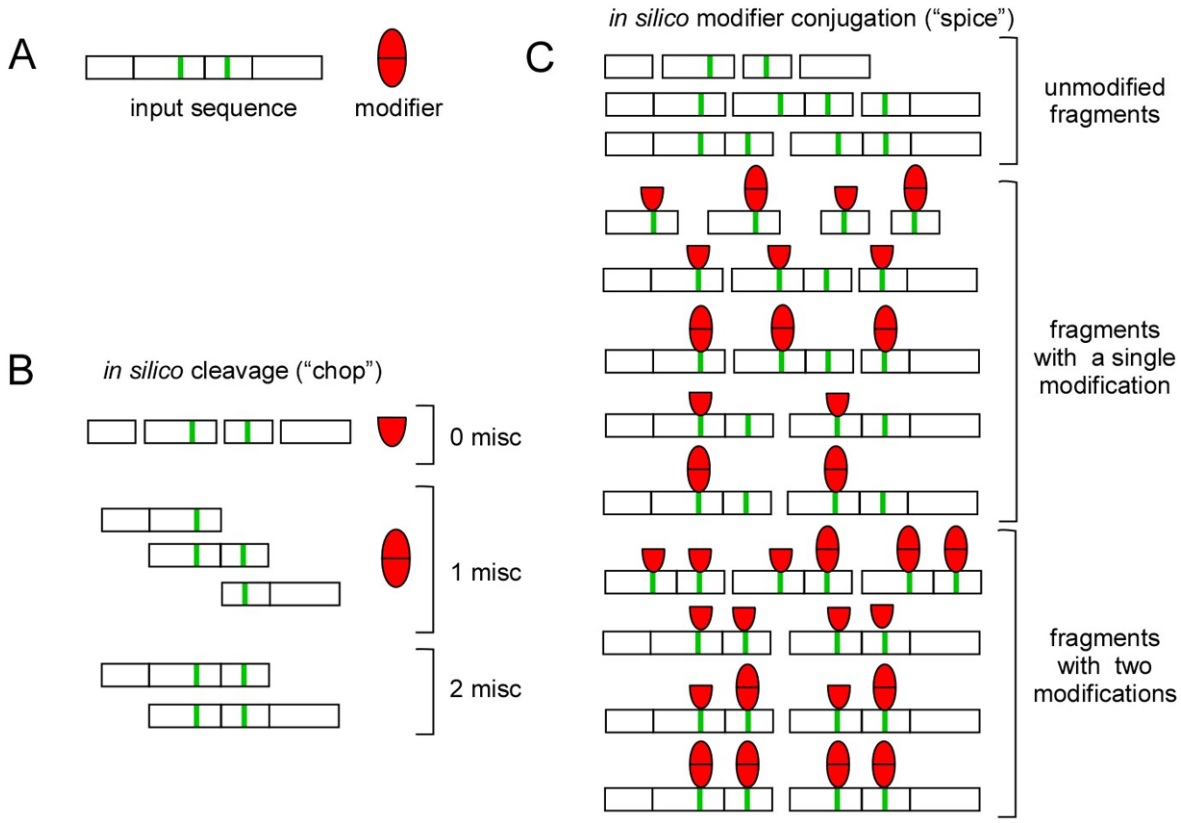

Figure 7.2: ChopNSpice workflow

A. Input data: one (or more) target protein sequences are provided. These may contain protease sites (black) and putative modification sites (green). Also shown is the sequence of a modifier (red), which may also contain a protease site (black).

B. 'Chop': in silico digest of input sequences (target and modifier). Upper and lower boundaries for allowed missed cleavages can be specified independently for both the target and the modifier.

C. 'Spice': the modifier is conjugated to cleaved fragments that contain a modification site (green). The user can additionally choose to include unmodified fragments in the output, as well as allowing more than one modification per fragment to take place.

be set independently for the modified protein and the modifier, or different sets of combinations can be excluded from the calculation.

7.3 Output options

7.3.1 Peak list of modified peptides

Monoisotopic and average m/z values according to the different charge states of the modified peptides (that could be observed in MALDI and ESI-MS, viz, from +1 to +5, see Fig. 7.3-A for example) after endoproteolytic digestion of the modified proteins can be calculated[4] by ChopTools and then used further, e.g. as an inclusion list or to help the user with the interpretation of simple MALDI spectra. m/z peak lists may subsequently be downloaded in two different formats: either in a tabular format as also displayed in the web browser, or in a list format for direct use as an inclusion list for, e.g., data-dependent data acquisition in the mass spectrometer. Importantly, such a peak list includes only putatively modified peptides. Confirmation that observed masses which match the theoretically generated values are indeed modified peptides must be obtained by sequencing, i.e. by MS/MS experiments.

.

7.3.2 Integration into a database for search with MS-based search engines

A much more powerful approach in analysing modified peptides is the sequencing of these peptides and subsequent evaluation of the corresponding MS/MS spectra by the use of available search engines (e.g. the widely employed MASCOT and Sequest search engine packages) in combination with database searching. To make this possible, a software package must be able to deal with branched peptides. Importantly, the masses of a linear and a branched peptide composed of the same amino acids are identical. Therefore, ChopTools reports branched fragments (as illustrated in Fig. 7.1 and Fig. 7.2) in a linear fashion, as fusion peptides consisting of peptides derived from the modified protein and modifier (see above). MS search programmes for the analysis of MS and MS/MS data are able to make use of knowing which peptides originated from the same protein and thus to identify the

[4]Two calculation modes are supported: average and monoisotopic mass. The calculation is based on masses listed in Table B.3 in the Appendix.

A

Result

>P37840|SYUA_HUMAN Alpha-synuclein - Homo sapiens(Human).

Peptide	m/z 2^+	m/z 3^+	m/z 4^+
MDVFMK	385.682410	257.457365	193.344843
GLSK	202.628817	135.421637	101.818046
AK	109.578596	73.388156	55.292936
EGVVAAAEK	437.237449	291.827391	219.122362
...	...	...	...
MDVFMKGLSK	578.298669	385.868205	289.652972
ELGMEEEDVIEVYQEQTGGMDVFMKGLSK	1646.258918	1097.841704	823.633097
IADNHTPKELGMEEEDVIEVYQEQTGGMDVFMKGLSK	2084.481601	1389.990159	1042.744438
GLSKAK	302.194856	201.798996	151.601066
ELGMEEEDVIEVYQEQTGGGLSKAK	1370.155105	913.772495	685.581190
IADNHTPKELGMEEEDVIEVYQEQTGGGLSKAK	1808.377788	1205.920950	904.692532
AKEGVVAAAEK	536.803488	358.204750	268.905382
ELGMEEEDVIEVYQEQTGGAKEGVVAAAEK	1604.763737	1070.178250	802.885506
IADNHTPKELGMEEEDVIEVYQEQTGGAKEGVVAAAEK	2042.986420	1362.326705	1021.996848
EGVVAAAEKTK	551.808770	368.208272	276.408023
ELGMEEEDVIEVYQEQTGGEGVVAAAEKTK	1619.769019	1080.181771	810.388147
IADNHTPKELGMEEEDVIEVYQEQTGGEGVVAAAEKTK	2057.991702	1372.330227	1029.499489
...	...	...	...

B

Result

>P37840|SYUA_HUMAN Alpha-synuclein - Homo sapiens(Human).
MDVFMKJGLSKJAKJEGVVAAAEKJTKJQGVAEAAGKJTKJEGVLYVGSKJTKJEGVVHGVATVAEKJTKJEQVT
NVGGAVVTGVTAVAQKJTVEGAGJMDVFMKGLSKJELGMEEEDVIEVYQEQTGGMDVFMKGLSKJIADNHTPKEL
GMEEEDVIEVYQEQTGGMDVFMKGLSKJGLSKAKJELGMEEEDVIEVYQEQTGGGLSKAKJIADNHTPKELGMEE
EDVIEVYQEQTGGGLSKAKJAKEGVVAAAEKJELGMEEEDVIEVYQEQTGGAKEGVVAAAEKJIADNHTPKELGM
EEEDVIEVYQEQTGGAKEGVVAAAEKJEGVVAAAEKTKJELGMEEEDVIEVYQEQTGGEGVVAAAEKTKJIADNH
TPKELGMEEEDVIEVYQEQTGGEGVVAAAEKTKJTKQGVAEAAGKJELGMEEEDVIEVYQEQTGGTKQGVAEAAG
KJIADNHTPKELGMEEEDVIEVYQEQTGGTKQGVAEAAGKJQGVAEAAGKTKJELGMEEEDVIEVYQEQTGGQGV
AEAAGKTKJIADNHTPKELGMEEEDVIEVYQEQTGGQGVAEAAGKTKJTKEGVLYVGSKJELGMEEEDVIEVYQE
QTGGTKEGVLYVGSKJIADNHTPKELGMEEEDVIEVYQEQTGGTKEGVLYVGSKJEGVLYVGSKTKJELGMEEED
VIEVYQEQTGGEGVLYVGSKTKJIADNHTPKELGMEEEDVIEVYQEQTGGEGVLYVGSKTKJTKEGVVHGVATVA
...

Figure 7.3: ChopTools example results

A. Example output in m/z calculation mode. The user can specify which charge states she wants to compute (1+, 2+, 3+, 4+, 5+).

B. Example output in FASTA library mode. Target protein fragments are colour coded in grey, modifier fragments in red, recognized modification sites in green, and the virtual amino acid 'J' to serve as a cleavage marker in downstream processing is colour coded in blue.

corresponding protein in a database. To facilitate the exploitation of this powerful feature in conjunction with the fragment library generated by ChopTools, the user can choose to fuse all fragments into a large fusion protein that contains all the fragments (modified and non-modified peptides) that originate from the same protein. Fragments within the fusion protein are separated by the virtual amino acid

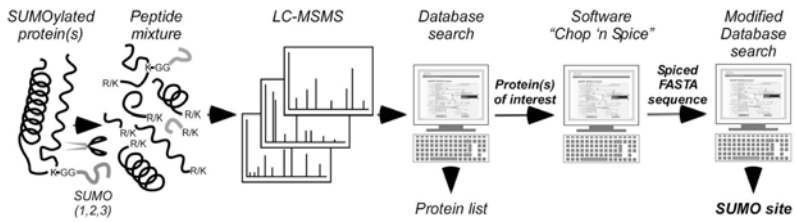

Figure 7.4: Database search scheme building on ChopNSpice
Sumoylated proteins are digested with endoproteinases and analyzed by LC-MS/MS. The corresponding proteins are identified by a database search using search engines (MASCOT and/or Sequest). Putatively sumoylated protein sequences are 'chopped' and 'spiced' (c.f. Fig. 7.2) and the spiced FASTA sequences are added to the database. The search with the search engine is repeated to identify the sumoylated peptide with its corresponding acceptor site (adapted from [134])

'J'. The user defines a virtual 'J-endopeptidase' within a downstream analysis tool such as MASCOT or Sequest. As a result, database search tools can be provided with a library of all possible fragments (modified and non-modified peptides) whilst retaining all their strengths such as fragmentation b/y ion series and information on which fragments originated from the same protein. An example output of this kind is displayed in Fig. 7.3-B. Alternatively, each library fragment can be assigned its own unique FASTA record, obviating the need for the virtual amino acid 'J'. In the case of protein-protein crosslinking, dead-end linker reagents attached to cysteine and lysine residues are reported as the virtual amino acids 'B' and 'Z' respectively, whilst complete linkage reactions are represented by the virtual amino acid 'O'. Of note, crosslinking ligation products are reported twice in the output, once in the order A-B, once B-A, to allow downstream search engines to generate b and y ion series for each of the fused fragments. A download option for a text-based result file for downstream processing is available. Fig. 7.4 illustrates an example of how a FASTA library file generated by ChopNSpice fits into the overall analysis workflow of identifying modification sites from a complex sample.

7.4 Access and usage modes

ChopTools allows great flexibility of use by providing various different means of access:

1. Web interface – The easiest way of using ChopTools is its web interface, requiring only a web browser and an internet connection.

2. Web service – Automated high-throughput analysis schemes. Power users, or users wanting to generate very large libraries, can make use of a web service offering machine-to-machine interoperability. The web service accepts instructions in XML via HTTP-POST and responds directly with the result in the desired format.

3. Local installation – Both the ChopNSpice and ChopNStitch modules are distributed under GPLv3 and may be downloaded and run as a local copy. They can be run either as a web interface / web server on a local server, or on the command line.

Extensive information on all three modes of use, including detailed tutorials and the DTD specifications of the XML input formats for both modules, are available online via the ChopTools website[5].

7.5 Validation - three test cases

The usefulness of the strategy to generate FASTA library files containing branched fragments represented in a linear fashion has already been demonstrated for the analysis of protein-protein crosslinking scenarios [133]. To further validate this approach, *in vitro* sumoylation reactions were performed on RanGAP1, Sp100 and p53 by Erik Meulmeester of the Department of Biochemistry I in the University of Göttingen. He-Hsuan Hsiao of the bioanalytical mass spectrometry group in the Max-Planck Institute of biophysical chemistry then subjected the samples to LC-MS/MS and subsequently tested different approaches to identify SUMO acceptor

[5] `http://choptools.gwdg.de`

sites in the three proteins. Using standard database search tools alone (Mascot, Sequest), it was not possible to identify any sumoylated peptides, an observation shared by other groups [129]. In contrast, by generating FASTA library files for each of the protein and searching these with Mascot and Sequest (c.f. Fig. 7.4), well-known SUMO modifications were confirmed on RanGAP1 on lysine 526 [135], p53 on lysine 386 [136], and Sp100 on lysine 297 [137] (for primary data, please refer to [134]. In addition to identifying all known modifications, it was possible to identify several new acceptor sites [134].

Chapter 8

Discussion

It is unworthy of excellent men to lose hours like slaves in the labor of calculation which could be relegated to anyone else if machines were used.

—Gottfried Wilhelm von Leibnitz

ChopTools successfully bridges the gap between well-established and very powerful software suites and the analysis needs of modern proteomics approaches. Previously users wishing to apply the approach of linearized peptide libraries in their analysis, or simply in the need of mass lists for the analysis of Ubl modification or crosslinking experiments, had to go through the laborious process of generating these resources by hand. Now ChopTools offers a convenient platform to take over this duty.

ChopTools is available as a free webserver offering distinct advantages over distributed software intended for local installations. Users may conveniently run the software without installation from any machine independant of location or configuration, and can rest assured to always be using the most recent version of the program. Furthermore, the intrinsic nature of a web server implies parallelisation. Multiple requests can be issued to the web service in parallel which in turn is able to directly make use of parallel computing facilities available to the operating system on the server or server cluster. In addition to the web server, a local copy is available under a GPL license to users with limited internet access, or users who wish to modify or extend the software to suit their needs.

Unlike many other software packages that address Ubl modification or crosslinking, ChopTools is not limited to one particular type of analysis, or a particular type

of mass spectrum. The tool is applicable to any Ubl modification or crosslinking scenario and can easily be integrated in existing analysis schemes. The availability of a web service in addition to the browser based user interface enables machine to machine interoperability and is particularly attractive to users who wish to intergrate ChopTools in any stage of a high-throughput analysis workflow. The modular nature of the 'Spice' and 'Stitch' components allows each of the components to be used both seperately or in combination with each other. The latter would for example be of interest in a scenario where a Ubl modification is to be investigated in the context of disulphide bridges.

The fact that ChopTools is easily accessible, versatile and easy to integrate into existing MS analyses workflows make it a simple-to-use, yet very powerful tool for any mass spectroscopist investigating large peptide modifications, be they Ubl- or other protein-based post-translational modifications or proteinprotein crosslinking.

Current limitations

The crosslinking module 'ChopNStitch' currently only generates fusion peptides containing at most two digestion fragments. Ligation products of higher order, e.g. more than two fragments linked to each other in a chain like fashion can theoretically occur. However with respect to mass spectrometry, these are expected to be of little significance due to the high masses involved, as these are extremely difficult to detect in practise. Nevertheless, support for higher order linkage products will be implemented in the near future. Furthermore, 'ChopNStitch' is currently geared towards homo-link reagents (e.g. Cysteine-Cysteine or Lysine-Lysine), but hetero-linkers (e.g. Cysteine-Lysine) will be included in the next release.

Downstream analysis tools that take advantage of FASTA library files generated by ChopTools (e.g. Mascot, Sequest) identify peaks in MS spectra by generating expected fragmentation patterns of the peptides in the library file and matching them to observed fragments in the spectrum. Although the observed fragmentation of a branched peptide has been shown to be very similar to the expected fragmentation of the linear peptide representation [134], the two fragmentation patterns (b- and y-ion series) are not 100% identical. The most abundant ions are predicted correctly, which allows reliable identification of the modified species in the spectrum

as demonstrated in [134]. However, care must be taken when comparing quality factors for such matches generated by downstream analysis software packages, as such quality factors take the entire fragmentation pattern into account. Therefore, whilst the expected scores will be sufficient to qualitatively identify modifications of interest, the quality factor of the match has to be carefully investigated before being used in comparison with other quality factors.

Future perspectives

The generation of an array of large-scale modification libraries (e.g. the entire Swiss-Prot human/NCBI-nr treated with common enzymes such as trypsin etc. and modified by a Ubl) is currently underway and will soon be available for download via the ChopTools website. These libraries will contain all possible modifications and endoproteolytic peptides for various given Ubl/protease combinations and thus potentially constitute a very useful resource for proteomic studies.

Part IV

Structural Basis of Dead End Function

Chapter 9

Background

'If you want to understand function, study structure.'
—Francis Crick [138]

9.1 Summary

The Dead end protein has been shown to play a crucial role in fate maintenance, survival and migration of primordial germ cells in the model organism zebrafish (*Danio rerio*). In addition, a particular mutation in the gene encoding the mouse dead end orthologue results in germ cell loss and testicular tumours. Close homologues of the protein are also found in african clawed frog, chick and human, suggesting a central and general role for this protein in germ cell development. Primary sequence analysis and functional characterisation suggest that the essential core region of the protein consists of a putative RNA or protein interaction domain with an RNP fold and another domain of similar size with unknown structure. Recent studies have shown that Dead end is capable of protecting mRNAs from microRNA (miRNA) mediated repression.

9.2 The *dead end* phenotype

The progenitors of the gametes, primordial germ cells (PGCs), are set aside early during development of an organism and differentiate independently of the soma [139, 140, 141, 142]. In most species, PGCs migrate from the site at which they are

specified toward the developing gonad where they will differentiate into sperms or eggs in the adult (see Fig. 9.1 A) [142]. The *dead end* (*dnd*) gene was first identified in the model organism zebrafish as a germ plasm component by means of a large-scale whole-mount in-situ screen for genes expressed in PGCs, *dnd* RNA was shown to be exclusively expressed in PGCs where it localises to perinuclear germ granules as assessed by expression of a *dnd*-GFP fusion protein (Fig. 9.1 B) [5]. This characteristic expression pattern - restricted to germ cells and localised to perinuclear germ granules - has also been observed in other model organisms such as chick, mouse and *Xenopus laevis* (african clawed frog) [143, 144, 145, 146]. Interestingly, when *dnd* translation is inhibited by the injection of *dnd*-specific morpholino-antisense oligonucleotides into zebrafish embryos, PGCs lose their motility and are thus unable to migrate towards the gonad (Fig. 9.2 A-B) [147]. The immotile Dead end knocked down cells eventually lose the expression of several RNAs that are normally expressed in the PGCs and die (Fig. 9.2 C-F). A very similar phenotype is also observed in the model organism *Xenopus laevis* [146]. Dead end function is thus evolutionary conserved and essential for germ cell fate maintenance and survival [5].

In mouse, *dnd* is also required for germ cell survival [148]. Specifically, the Ter mutation in *dnd*, which is a single base substitution introducing a premature stop codon at amino acid position 178, exhibits germ cell loss. Strikingly, the same mutation in a different genetic background is manifested as formation of testicular germ cell tumours [6] and partial embryonic lethality [148]. The implications of studying the Dead end protein thus range from basic Germ Cell Biology to Tumour Biology with particular relevance to testicular germ cell tumours in humans.

9.3 Known functions and mechanisms: RNA binding and miRNA inhibition

Recent findings have shown that Dead end modulates MicroRNA (miRNA) dependant regulation of protein expression by binding to mRNA [149]. miRNAs are inhibitors of gene expression constituting a family of approximately 22 nucleotide

Figure 9.1: Primordial germ cells and *dead end* in zebrafish.
A. Primordial germ cells are specified at random starting points during early development. Over a time period of 24 h they migrate from their site of origin to the developing gonad where they will differentiate into sperms or eggs in the adult.
B. The expression pattern of *dnd* in the developing embryo is clearly restricted to primordial germ cells.

long RNAs widely expressed in metazoans [150], and are implicated in the regulation of a wide variety of cellular processes, as well as the definition of cell fate [151, 152].

Most miRNAs are transcribed by RNA polymerase II as long RNAs that are converted to approximately 70 nucleotide long pre-miRNAs by the RNAse Drosha [153]. Upon transport to the cytoplasm, the RNAse Dicer converts the pre-miRNA into mature, 22 nucleotide fragments, of which one strand is incorporated into the RNA-induced silencing complex (RISC) [154, 155]. In animals, a two to eight nucleotide long seed sequence at the miRNA 5' end leads to association with 3' untranslated regions (UTR) of mRNAs, which in turn suppresses gene expression by inhibiting translation [151, 156]. This process has also been found to be linked to mRNA degradation in a number of cases [157].

In the study referenced above, Reuven Agami and coworkers were able to show, through immunoprecipitation of whole cell extracts, that human Dnd1 (HDnd1) specifically binds to endogenous p27 and cyclophilin mRNA [149]. This interaction,

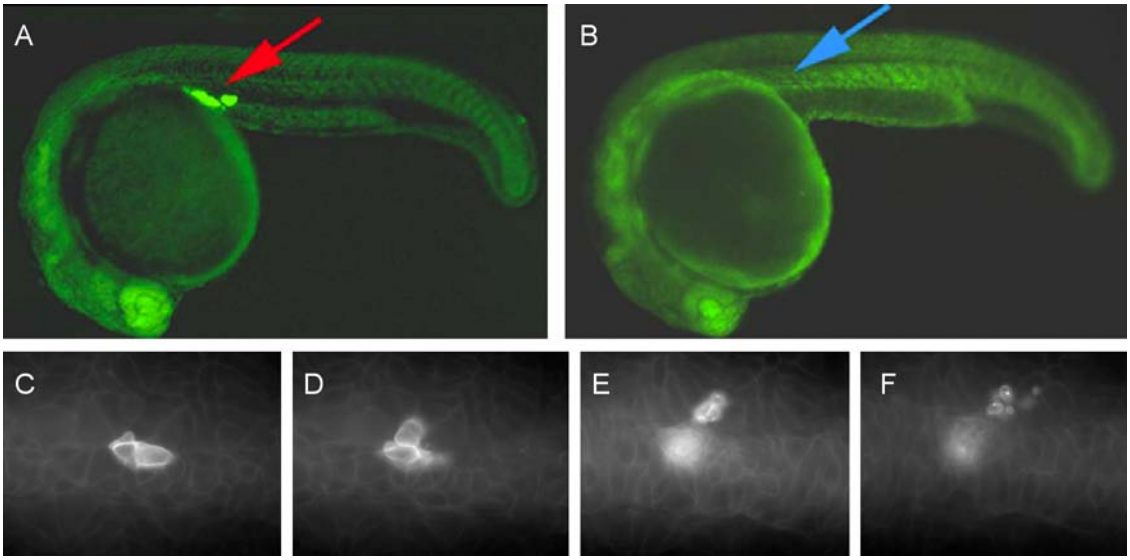

Figure 9.2: The Dead end knockdown phenotype in zebrafish.
A-B. Zebrafish embryos 24 h after fertilisation. The germ cells are labelled in green in wild-type embryos (red arrow in A), but are absent in embryos in which the function of Dead end is knocked down (blue arrow in B).
C-F. Germ cells knocked down for Dead end function undergo cell death within the first day of development (pictures from a 45 minutes long time-lapse movie).

which requires an intact RNA binding domain, inhibits access of the miRNA mi-221 to the mRNA UTR. Luciferase reporter assays revealed that uridine rich regions in the 3'UTR of p27 are sufficient and required for this process and thereby mediate the function of HDnd1. The proposed mechanism whereby Dead end protects mRNA and consequently alleviates miRNA repression was also confirmed *in vivo*, as the zebfrafish isoform (ZDnd) is able to counteract the inhibition of the developmental determinant Nanos by miR-430.

Other functions, such as counteracting RISC activity or involvement in subcellular sequestration, can not be ruled out at this time. In fact, the putative RRM motifs in different Dead end isoforms only account for about one third of the protein's length, thus suggesting the possibility of further interactions, both on the protein-RNA as well as on the protein-protein side. Indeed, recent studies focused on the function of mouse Dead end suggest that Dnd1 interacts with APOBEC3,

both *in vitro* and *in vivo* as shown by immuno-precipitation experiments [145]. Apobec proteins contain zinc-finger domains with homology to bacterial cytidine deaminases. The function of many members of this protein family has remained unknown, however specifically APOBEC3 has been shown to inhibit viral replication through hypermutation of DNA by cytidine deamination [158, 159]. However, to date it is neither known whether the aforementioned interaction of Dnd1 and APOBEC3 is of direct nature, or whether it is mediated by other proteins or cofactors, nor what functional significance this interaction bears.

9.4 Structural data

Currently, no structural data on *dnd* is available, however homology searches predict an RNA recognition motif (RRM) as found in pre- and mRNA associated proteins. This finding is in agreement with the sub-cellular location of Dead end, as studies in different organisms have shown that germ granules contain multiple putative RNA-binding proteins. An even stronger indication that Dead end contains a canonical RNA recognition motif, stems from immuno-precipitation experiments that demonstrated binding of Dead end to mRNA UTRs [149]. Furthermore the mouse Dead end protein is most similar to the apobec complementation factor [6], which is part of the cytidine to uridine RNA-editing complex.

The following chapters describe design, cloning, expression and purification studies of several Dead end isoforms, thereby providing the basis for structural studies by NMR or X-ray crystallography. Said studies provide an important step in the elucidation of the molecular basis of Dead end function, both as a developmental determinant and as potential (proto-)oncogene with direct implications in human cancer.

Chapter 10

Materials & Methods

10.1 Materials

10.1.1 Plasmids and *E. coli* strains

Plasmids

Table 10.1: Names, features and references of plasmids used

Plasmid	Resistance	Features	Reference
Raz495	Amp	ZDnd 1-411	E. Raz ZMBE Münster
pCMV-sport6	Amp	HDnd1 including UTRs	R. Agami NKI, Amsterdam
pETM30	Kan	XDnd	T. Pieler GZMB, Göttingen
pMal-c2x	Amp	XDnd	T. Pieler GZMB, Göttingen
pET32a	Amp	MCS	Stratagene
pET16b-TEV	Amp	derived from pET16b, including an additional TEV protease cleavage site allowing removal of the N-termial His_7-tag	K. Saxena, JW Goethe Universität Frankfurt
pGEX-2T	Amp	N-terminal GST tag, MCS	Stratagene
pGEX-2TEV	Amp	derived from pGEX-2T, including an additional TEV protease cleavage site	S. Becker, MPI-bpc, Göttingen
pMal-TEV	Amp	derived from pMal, replacing the Thrombin by a TEV cleavage site	K. Saxena, JW Goethe Universität Frankfurt
pFastBac HTa	Amp	MCS, His-tag, TEV site	Invitrogen GmbH Karlsruhe

E. Coli strains

Strain	Genotype	Reference
Arctic Express^{TM}(DE3)	F^-, *ompT*, *hsdS*$_B$, (r_{B^-}, m_{B^-}), *dcm*, *gal* λ(DE3) *endA* Hte [*cpn10 cpn60* Gent^r	Stratagene
Bl21 (DE3)	F^-, *ompT*, *hsdS*$_B$, (r_{B^-}, m_{B^-}), *dcm*, *gal* λ(DE3)	[69]
Rosetta2 (DE3)	F^-, *ompT*, *hsdS*$_B$, (r_{B^-}, m_{B^-}), *dcm*, *gal* λ(DE3) pLacIRARE2 (Cam^R)	Novagen
XL2-Blue	recA1 endA1 garA96 thi-1 hsdR17 supE44 relA1 lac [$FproABlac^qZ\triangle M15Tn10(Tet^r)AmyCam^r$]	Stratagene

Table 10.2: Names, genotypes and references of bacterial strains used

10.1.2 Oligonucleotides

All oligonucleotides were synthesised on a 25 nmol scale by Invitrogen GmbH, Karlsruhe. For primer design considerations, please refer to Section 10.2.1 on page 105.

Oligonucleotides for cloning of ZDnd and XDnd constructs

Name	Sequence	Features
Dnd1	CGCTCAGACATATGGTAGGCGATATGGACGCAC	N-term fwd, NdeI
Dnd2	GTCGTCGACGGATCCTTAAGTCCATCTGACGGTGATGGAAATG	218 rev, BamHI
Dnd3	GTCGTCGGATCCAAGCTTTTAGGTTACGCAGCTGTCTTCCTG	237 rev, HindIII, BamHI
Dnd4	GTCGTCGGATCCAAGCTTTTAACCCCCAACAGCCCGGAAG	269 rev, HindIII, BamHI
Dnd5	GTCGTCGGATCCAAGCTTTTAGAAAGGCCGTAAATTTGAGACTC	C-term rev, HindIII, BamHI
Dnd6	GTCGTCGGATCCAAGCTTTTAAGTCCATCTGACGGTGATGGAAATG	218 rev, HindIII, BamHI
Dnd7	CTGCCCGGGATCCATGGTAGGCGATATGGACGCAC	N-term fwd, BamHI, SmaI
Dnd10	GTCGTCGGATCCATGGGTTCGGGCTGTGAGGTTTTC	57 fwd, BamHI
Dnd11	GTCGTCAAGCTTGAATTCTTAGCTCCTGCGCACGGTCAG	136 rev, EcoRI, HindIII
Dnd12	CGCTCAGACATATGGGTTCGGGCTGTGAGGTTTTC	57 fwd, NdeI
Dnd13	GTCGTCGACGGATCCTTAGCTCCTGCGCACGGTCAG	136 rev, BamHI
Dnd14	GCTAAGAAAGTGCTCGTGGAGGCTTTTCGGAACCGGTACG	mutagenesis fwd (-HindIII)
Dnd15	CGTACCGGTTCCGAAAAGCCTCCACGAGCACTTTCTTAGC	mutagenesis rev (-HindIII)
Dnd17	GTCGGATCCGGCGCCATGGTAGGCGATATGGACGCAC	N-term fwd, NarI
Dnd18	GGTTCTAGAGAATTCTTAGAAAGGCCGTAAATTTGAGACTCG	C-term rev, EcoRI, XbaI

Table 10.3: Sequences and features of oligonucleotides for cloning of ZDnd and XDnd constructs[1]

Oligonucleotides for cloning of HDnd1 constructs

Table 10.4: Sequences and features of oligonucleotides for cloning of HDnd1 constructs

Name	Sequence	Features
Dnd19	CGCTCAGACATATGCAGTCCAAGCGGGATTGTG	N-term fwd, NdeI
Dnd20	CGCTCAGACATATGGGGTCAGAGGTGTTCATCGGG	57 fwd, NdeI
Dnd21	CGCTCAGACATATGGGCTGGGTGGGCAGCCC	48 fwd, NdeI
Dnd22	GTCGTCGGATCCAAGCTTTTACTGTTTAACCATGGTACCTGCCTC	C-term rev, BamHI, HindIII
Dnd23	GTCGTCGGATCCAAGCTTTTACTTCTCGGTGCTGCGGCAC	137 rev, BamHI, HindIII
Dnd24	GTCGTCGGATCCAAGCTTTTACTTGAGCCACTCCACAGCCAC	217 rev, BamHI, HindIII
Dnd25	GTCGTCGGATCCAAGCTTTTAACCCACAAGCTGCTGGCGAAG	230 rev, BamHI, HindIII
Dnd26	CGCTCAGACATATGTCCCCACAGCCAGAGGGC	234 fwd, NdeI
Dnd27	CTGAACCGCGGCTACGCCTATGCCCG	mutagenesis fwd, F100 → Y
Dnd28	CGGGCATAGGCGTAGCCGCGGTTCAG	mutagenesis rev, F100 → Y

10.1.3 Growth media and antibiotics

Standard growth media and antibiotics were prepared as described in Chapter 3. Please refer to Table 3.3 on page 25.

Auto-inducing media

Complex, auto-inducing media were prepared as described in [160]. In particular, the ZYM-5052 medium was used routinely when expression of soluble protein in conventional LB medium was unsuccessful.

10.1.4 Chemicals and enzymes

For a list of chemicals and enzymes used in this work, please refer to Table 3.4 on page 26. Standard chemicals not listed therein were supplied by Merck KGaA (Darmstadt, Germany) and Fluka (Neu-Ulm, Germany).

[1]The attentive reader will notice that the oligonucleotide Dnd16 is not included in the table. This specific primer was not used in the work described herein, however the original nomenclature of oligonucleotides was left unaltered to ease cross-referencing with original laboratory journals.

TEV protease production

Recombinant TEV protease was expressed and purified as described in [161].

Restriction enzymes

Restriction enzymes are listed in Table 3.4 on page 26.

10.1.5 Laboratory equipment

For a list of laboratory equipment used in this work, please refer to Table 3.5 on page 28.

10.1.6 Bioinformatical tools and programs

The ExPASy proteomics server [162] offers a wide range of useful tools to analyze protein sequences and to aid in the design of suitable constructs. Table 10.5 lists tools that have been used in this work. Most programs, and many others, are available or linked through the ExPASy site (`http://www.expasy.org`).

Table 10.5: Bioinformatical tools

Name	Application	Reference
BLAST	Similarity searches	[163, 164, 165]
ScanProsite	Pattern and profile searches	[166]
Pfam HMM search	Pattern and profile searches	[167, 168]
DisEMBL	Prediction of disordered regions	[169]
GlobPlot	Protein disorder/globularity/domain predictor	[170]
ProtParam	Primary sequence analysis	[162]
JPred	Secondary structure prediction	[171]
PredictProtein	Secondary structure prediction	[172]
PSIpred	Secondary structure prediction	[173, 174]
SOPMA	Secondary structure prediction	[175]
3D-PSSM	Tertiary structure prediction	[176]
Geno3d	Tertiary structure prediction	[177]
Phyre	Tertiary structure prediction	[178]
SWISS-MODEL	Tertiary structure prediction	[179]
OligoCalc	Oligonucleotide properties	[180]
RaCC	Rare codon calculator	`http://nihserver.mbi.ucla.edu/RACC`
Mfold	DNA/RNA secondary structure prediction	[181]

10.2 Methods

Unless stated otherwise, all molecular-biological procedures were performed according to [70] or, in the case of commercial kits, according to the manufacturers' instructions.

10.2.1 Cloning

Primer design

Primers were designed with a salt adjusted T_m of 61-62 °C, aiming for a 'GC' clamp at the 5' and 3' end whilst avoiding self-complementarity. Melting temperatures and self-complementarity can be conveniently calculated with the web based tool *OligoCalc* [180] which is based on the formulae proposed in [182] and [183].

Polymerase chain reaction

Polymerase chain reaction (PCR) was employed to clone target fragments into expression vectors. Phusion™ polymerase Master Mix with high fidelity buffer (Finnzymes, Espoo, Finland) was used in all reactions according to the manufacturer's instructions. In contrast to other polymerases, this entails a 30 s denaturation step at 98 °C, annealing at 3 °C above the primer T_m for 30 s, and using 25 s/kb for extension at 72 °C. Reactions were carried out in a HYBAID PCR sprint cycler (Perkin Elmer, Wellesley, USA).

Phusion™ polymerase was found to be superior to *Pfu* polymerase commercially available from other sources, not only with respect to speed, but also fidelity, efficiency and overall performance. Consequently, Phusion™ polymerase was used exclusively throughout this work (with the exception of site directed mutagenesis, see below).

Subcloning

Subcloning was performed according to [70].

Site directed mutagenesis

In order to remove an unwanted restriction site whilst retaining the same translated amino acid sequence, or in order to introduce a point mutation on the amino acid level, site directed mutagenesis was performed on the plasmid level using the QuickChange™ mutagenesis kit. This kit is based on the *Site-specific Mutagenesis by Overlap Extension* principle as described in [70].

Primers containing the desired mutation and complementary regions around the target site were designed according to the QuickChange™ kit using the following formula

$$T_m = 81.5 + 0.41(\%GC) - 675/N - \%mismatch$$

where N is the primer length and $\%GC$ and $\%mismatch$ are whole numbers. The resulting primers to remove a HindIII site in XDnd are listed in Table 10.1.2 (Dnd14 and Dnd15), whilst the primers to introduce the F100 → Y point mutation are listed in Table 10.4 (Dnd27 and Dnd28).

The PCR reaction was carried out as described in the manufacturer's instructions, before digesting the (non-mutated) template plasmid with DpnI (37 °C, 1 h). 1 μL of the reaction was then transformed into XL2-blue cells (see page 107). The sequence of plasmid DNA prepared from resulting clones was verified using the *HotShot* sequencing procedure described below.

DNA preparation and purification

Plasmid preparations were performed on two scales. For typical yields of 25 μg and 200 μg plasmid DNA, *E. coli* (XL2-blue, see Table 10.1.1) overnight cultures were grown in 10 mL and 100 mL LB-medium respectively. Small scale preparations employed the NucleoSpin Plasmid kit (Machery-Nagel, Düren, Germany), large scale preparations the Nucleobond AX PC 100 kit (Machery-Nagel, Düren, Germany) according to the manufacturer's instructions.

PCR products were purified using the NucleoSpin Extract II kit (Machery-Nagel, Düren, Germany) according to the manufacturer's instructions.

DNA concentrations were determined from the optical density at 260 nm (A_{260}) and calculated using the Beer-Lambert law [70]. Briefly, an optical density of 1 meas-

ured in a cuvette of path length 1 cm, corresponds to a concentration of 50 μg/mL and 40 μg/mL of double-stranded and single-stranded DNA, respectively.

Sequencing

To verify cloning procedures, purified plasmid DNA was sequenced commercially by the company SeqLab (Göttingen, Germany). The *HotShot*-procedure follows a specialized *Cycle Sequencing* protocol and is based on the original method proposed by [184].

10.2.2 Transformation in *E. coli*

A 50 μL aliquot of competent *E. coli* cells stored at -80 °C was thawed on ice and mixed with 1 μL plasmid DNA. The suspension was incubated for 30 min on ice, before being subjected to a heat shock treatment for 45 sec at 42 °C. Subsequently the cells were cooled on ice for 2 min, suspended in 950 μL 2x YT-medium (37 °C warm) and incubated for 1 h at 37 °C. Following centrifugation (3000g, 5 min), the cell pellet was gently resuspended in approximately 120 μL 2x YT-medium and plated on agar plates containing the desired antibiotic for selection.

10.2.3 Protein expression in *E. coli*

Different strains were screened to optimize expression, including the all-purpose expression strain BL21 (DE3), the rare-codon compensating strain Rosetta2, and the low-temperature, chaperone carrying strain Arctic Express^{TM} (for genotypes and suppliers please refer to Table 10.1.1 on page 102).

A single culture from an agar plate incubated overnight (37 °C) was used to inoculate 10 mL LB medium containing an antibiotic for selection of the expression plasmid (Ampicillin or Kanamycin). This culture was incubated for at least 5 h and then diluted (1:1000) into the growth medium of choice (including the selection antibiotic). To optimize expression and solubility, several media were tested for each construct (LB, 2x YT, ZYM-5052 (see below), LB + 1% ethanol to induce *E. coli* heat shock responses [185]). This large culture was then grown at 37 °C and monitored by regular measurements of the optical density at 600 nm (OD_{600}). Upon

reaching an OD_{600} of 0.4, if expression at a temperature different than 37 °C was desired, the culture was shifted to this temperature. In addition to 37 °C, the following temperatures were tested for many constructs: 30 °C, 25 °C and 20 °C. Expression at even lower temperatures was tested using the specialized strain Arctic Express™ (see below).

Upon reaching the desired expression temperature, expression was induced in the logarithmic growth phase (OD_{600} 0.8) by the addition of 1 mM IPTG. To assess the efficiency of expression, samples were taken before, 1 h and 3 h after induction. All cells were finally harvested by centrifugation (5000*g*, 30 min, 4 °C), and frozen in N_2 (l) if necessary.

Alternative growth media & auto induction

To improve expression levels of poorly expressing constructs, and to improve solubility of insoluble constructs, a complex, auto-inducing medium was used. The ZYM-5052 medium [160] contains a more standardized amount of cofactors and transition metals than found in yeast extract and tryptone alone. It also inhibits early induction and allows large numbers of constructs of conditions to be tested simultaneously whilst bearing the potential for significantly higher yields than standard expression media. Expression in ZYM-5052 was performed as described in [160].

Expression in Arctic Express™ cells

To improve solubility, constructs were expressed in commercially available Arctic Express™ cells. The activity of the *E. coli GroEL/ES* complex is known to decrease dramatically at low temperatures [186]. Therefore, to enable low temperature expression whilst retaining chaperone activity, Arctic Express™ cells contain a pACYC based plasmid encoding the cold-adapted chaperonins *Cpn10* and *Cpn60* from the psychrophilic bacterium *O. antarctica*.

The expression protocol closely followed the supplier's instructions. Briefly, a single colony was picked from an agar plate and used to inoculate 10 mL LB medium containing antibiotics for selection of the expression plasmid and the *Cpn10/Cpn60* coding plasmid. This culture was grown overnight at 37 °C and subsequently diluted

(1:1000) into LB medium containing no antibiotics. This culture was incubated at 30 °C for 3 h, before being transferred to 11 °C. As soon as the temperature of the medium had equilibrated, protein expression was induced with 1 mM IPTG. Subsequently, expression continued for 24 h at 11 °C, before harvesting the cells by centrifugation (5000g, 30 min, 4 °C).

10.2.4 Protein expression in insect cells

The construct of interest was cloned into the pFastBac HTa vector (Invitrogen GmbH, Karlsruhe, Germany). Protein expression performed using the *Bac-to-Bac* Baculovirus Expression System (Invitrogen GmbH, Karlsruhe, Germany) according to the manufacturer's instructions. Briefly, the bacmid was transfected into sf9 insect cells using the Cellfectin® transfection reagent (Invitrogen GmbH, Karlsruhe, Germany) to achive expression. Five days after transfection, cells were harvested for analysis.

10.2.5 Protein purification

Lysis buffer screens

The solubility of a protein can be strongly affected by the lysis buffer used. For initial trials to assess expression levels and solubility, a standard lysis buffer was used (see Table 10.6). Following harvesting of the bacterial cells as described above, the cell pellet was resuspended in standard lysis buffer at a ratio of 10 mL lysis buffer per 1 g of cells (wet weight). The cells were lysed by sonication (8 cycles, 20 s, 10-80% power, TT13 spike) before separating soluble and insoluble fractions by centrifugation (40000g, 20 min, 4 °C). The pellet was resuspended in lysis buffer, and samples from the soluble fraction and the resuspended pellet were analysed by SDS-PAGE.

If lysis in standard buffer did not yield satisfactory levels of soluble protein, a range of buffers was screened to determine optimal conditions for cell lysis. A large number of buffers may be conveniently screened using a procedure based on the

Table 10.6: Standard lysis buffer

Compound	Concentration
Tris-HCl pH 7.4	50 mM
NaCl	250 mM
DTT	5 mM

solubility screening protocol of Jeanne Perry (Molecular Biology Institute, UCLA, Los Angeles, USA).

Cell growth and induction of protein expression followed the protocol described in section Section 10.2.3. After induction, one 1 mL sample at OD_{600} of 1 was taken for each buffer condition to be tested. If the OD of the cell culture was not 1, the sample volume was adjusted to obtain an equivalent size pellet. Cells were harvested by centrifugation in a microfuge (6000 rpm, 5 min) before discarding the supernatant and adding 100 μL of lysis buffer to each pellet. The first round of screening always employed the buffers listed in Table 10.7. After resuspending the cells (vortex 5 s), the cells were lysed by sonication (3 cycles, 10 s, 10% power, MS 72 spike). If a suitable sonicator is not available, common Freeze-thaw protocols may be substituted instead. Soluble and insoluble protein fractions were separated by centrifugation (maximum speed, 5 min, microfuge). Proteins were extracted from the soluble fraction by acetone precipitation: 1 mL acetone was added to each supernatant, vortexed (5 s) and frozen at -20 °C for 20 min, before removing acetone by centrifugation (maximum speed, 5 min, microfuge). Residual acetone was removed by drying the protein extract at 37 °C (15 min). The cell and acetone extract pellets were both resuspended in SDS-PAGE loading buffer, boiled (98 °C, 5 min), vortexed, centrifuged (maximum speed, 2 min, microfuge) and subsequently analysed by SDS-PAGE.

Based on the results of the basic lysis buffer screen, further screens were used to refine lysis conditions. These follow-up screens focus on pH, salt concentration, denaturant (urea), detergents and stabilizing additives. Buffers for screening each aforementioned parameter are listed in Tables 10.8 - 10.12.

Table 10.7: Basic lysis buffer screen

Buffer	Composition
7.4N	50 mM Tris-HCl pH 7.4
	50 mM NaCl
	50 mM EDTA
	1 mg/mL lysozyme
2S	50 mM Tris-HCl pH 7.4
	2 M NaCl
	5 mM EDTA
	1 mg/mL lysozyme
0.5U	50 mM Tris-HCl pH 7.4
	50 mM NaCl
	5 mM EDTA
	0.5 M urea
	1 mg/mL lysozyme
X	50 mM Tris-HCl pH 7.4
	50 mM NaCl
	5 mM EDTA
	0.2% triton X-100
	1 mg/mL lysozyme

Table 10.8: pH solubility buffer screen

Name	Buffer	Other components
5N	50 mM Na-Acetate pH 5.0	50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme
6N	50 mM MES pH 6.0	50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme
7N	50 mM Tris-HCl pH 7.0	50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme
8N	50 mM Tris-HCl pH 8.0	50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme
9N	50 mM Tris-HCl pH 9.0	50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme

Table 10.9: Salt solubility buffer screen

Name	Salt	Other components
0.1S	0.1 M NaCl	50 mM Tris-HCl pH 7.4, 5 mM EDTA, 1 mg/mL lysozyme
0.5S	0.5 M NaCl	50 mM Tris-HCl pH 7.4, 5 mM EDTA, 1 mg/mL lysozyme
1S	1.0 M NaCl	50 mM Tris-HCl pH 7.4, 5 mM EDTA, 1 mg/mL lysozyme
0.1KS	0.1 M KCl	50 mM Tris-HCl pH 7.4, 5 mM EDTA, 1 mg/mL lysozyme
1KS	1.0 M KCl	50 mM Tris-HCl pH 7.4, 5 mM EDTA, 1 mg/mL lysozyme

Table 10.10: Urea solubility buffer screen

Name	Urea	Other components
0.5U	0.5 M urea	50 mM Tris-HCl pH 7.4, 50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme
1U	1 M urea	50 mM Tris-HCl pH 7.4, 50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme
2U	2 M urea	50 mM Tris-HCl pH 7.4, 50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme
3U	3 M urea	50 mM Tris-HCl pH 7.4, 50 mM NaCl, 5 mM EDTA, 1 mg/mL lysozyme
4U	4 M urea	50 mM Tris-HCl pH 7.4, 50 mM NaCl, 5 mM EDTA
5U	5 M urea	50 mM Tris-HCl pH 7.4, 50 mM NaCl, 5 mM EDTA
6U	6 M urea	50 mM Tris-HCl pH 7.4, 50 mM NaCl, 5 mM EDTA

Table 10.11: Detergent solubility buffer screen

Name	Detergent	Other components
D	0.2% NP 40	20 mM Tris-HCl pH 7.4, 50 mM NaCl, 1 mg/mL lysozyme
X	0.2% triton X-100	20 mM Tris-HCl pH 7.4, 50 mM NaCl, 1 mg/mL lysozyme
T	0.2% Tween-20	20 mM Tris-HCl pH 7.4, 50 mM NaCl, 1 mg/mL lysozyme
M	0.2% dodecylmaltoside	20 mM Tris-HCl pH 7.4, 50 mM NaCl, 1 mg/mL lysozyme

Table 10.12: Stabilizer solubility buffer screen

Name	Stabilizer	Other components
10G	10% glycerol	20 mM Tris-HCl pH 7.4, 50 mM NaCl, 1 mg/mL lysozyme
20G	20% glycerol	20 mM Tris-HCl pH 7.4, 50 mM NaCl, 1 mg/mL lysozyme

10.2.6 Protein refolding

Refolding screens

Protein refolding screens were conducted using the iFold™Protein Refolding System 1 (Novagen, Darmstadt, Germany) according to the manufacturer's instructions. Briefly, inclusion bodies are isolated and washed using a detergent (Triton X-100). Upon detergent removal, inclusion bodies are solubilised using a mild detergent (N-laurylsarcosine) and subsequently refolded by rapid dilution into a 96-well matrix containing a selection of refolding conditions.

Large scale refolding procedures

Large scale refolding trials either followed rapid dilution by drop-wise addition to a buffer reservoir using a pump (LKB-Pump P1, Pharmacia, Uppsala, Sweden) or removed the detergent through dialysis.

Chapter 11

Results

11.1 The Zebrafish Dead end isoform

11.1.1 Prerequisite considerations

The first step towards determining a protein structure at atomic resolution consists of establishing a protocol to obtain a suitable sample of the protein of interest. If the protein is to be investigated by NMR spectroscopy, it is further required that the protein can be uniformly ^{13}C-^{15}N labelled and remains stable at concentrations of at least 0.3 mM. If this proves impossible to achieve with the wild-type protein, during the pursuit of a protocol it may be desirable to work with the smallest functional unit available for a number of reasons.

With the increase of molecular masses beyond 20 kDa, proteins become progressively more difficult to study with established NMR techniques, owing both to the increased number of resonances as well as broader linewidths due to faster transverse relaxation rates caused by longer rotational correlation times [187]. Consequently, spectral complexity escalates rapidly, whilst increased transverse relaxation rates also lead to a decrease in overall sensitivity [187]. With respect to X-ray crystallography, smaller fragments often display less dynamics and conformational variability, therefore increasing chances of successful crystallization and reducing the chance of missing flexible parts of the structure. Consequently, it may be necessary to seek the right balance between small size and retaining function if initial trials with the full length protein are not successful.

11.1.2 Expression of wild-type ZDnd and XDnd

Sequence information and cDNA clones are currently available for several Dead end isoforms (african clawed frog, chick, human, mouse and zebrafish). Out of these, the zebrafish isoform (ZDnd) was chosen as initial target as most functional knowledge at the molecular level was available for this isoform at the time (10/2005). Additionally, *Danio rerio* provides a powerful, yet fast system to test structure derived hypotheses *in vivo*. Direct collaboration with the group of Erez Raz made this a very attractive strategy, with mutual benefits to both structural as well as developmental biologists.

Full length ZDnd was cloned into two *E. coli* expression vectors based on the pET system (pET32a and pET16b-TEV), allowing protein expression under the control of a T7 promoter regulated by the *lac* operon. The pET32a vector allows expression of wild-type protein, whilst the pET16b-TEV vector is engineered to introduce an N-terminal His_7 affinity tag. The affinity tag opens up more purification strategies rather than solely relying on the basic pI of the untagged protein, but it may negatively impact solubility and hence both variants were tested.

The expression vectors were subsequently transformed into *E. coli* host strains to assess expression levels and solubility of the full length construct.

Expression in *E. coli* - BL21 (DE3)

Initial tests in the standard expression strain work horse BL21 (DE3) [69] showed high levels of expression, however upon lysis all target protein was found to be in inclusion bodies (IB) in the insoluble cell fraction. This observation was independent of the Histidine affinity tag and applied to expression trials from both vectors. Possible causes for this behaviour are manifold and need to be systematically assessed.

A well known parameter known to affect protein aggregation in *E. coli* expression is temperature [188, 189]. Whilst optimum growth conditions for the host, and concurrently often maximum protein yields, are reached at 37 °C, high cytoplasmic protein concentrations and rapid synthesis may lead to aggregation. Critical concentrations of folding intermediates or high concentrations of folded protein exposing hydrophobic patches has been postulated as the main reason for this effect.

A common remedy in this case is to lower the expression temperature with the aim to find a compromise between yield and solubility [188, 189]. Whilst this negatively impacts bacterial growth and target protein yield, the reduced speed of translation allows more time for correct folding and critical concentrations leading to aggregation are reached more slowly. This strategy was applied to the expression of full length ZDnd, screening several temperatures ranging from 20 °C and 30 °C. Samples were taken as early as 1 h after induction with IPTG to exclude adverse effects of increased target protein concentration. However, whilst the expected decrease in yield was observed, solubility could not be improved by these means.

Another important factor influencing the solubility of recombinant protein is the action of chaperones [190]. These proteins, whether endogenous or engineered, are able to assist folding and unfolding of macromolecular structures and may even be capable of reverting aggregation [191, 192]. Endogenous *E. coli* chaperones may be taken advantage of in a very straightforward fashion, without the need for additional expression vectors, as supplementing the growth medium with ethanol has been shown to trigger bacterial heat shock responses [185]. Part of the bacterial stress response, the expression of chaperones, most notably the *DnaJ/DnaK* and *GroEL/ES* complexes are strongly up regulated[1]. In the case of expression of recombinant ZDnd protein in *E. coli* however, the addition of ethanol to the growth medium did not lead to improved solubility. A second take on co-expression of chaperones in combination with low temperature expression is described in Section 11.2.2.

A final factor that is often neglected when expressing recombinant protein is the composition of the growth medium. Media based on different digests of casein have been found to impact solubility, although the reason for this has not yet been elucidated. A possible explanation could be the variation of trace element content in different lots. Additionally, early induction of expression in the non-logarithmic growth phase may negatively impact solubility and expression efficiency. The most frequent cause for this is traces of lactose in casein digests that relieve repression of the T7 promotor through the *lac* operon. These issues were addressed by expression in the complex, auto-inducing medium ZYM-5052 [160]. Amongst other benefits,

[1] for a review on heat shock responses, please refer to [193]

this medium contains standardized amounts of trace metals, and suppresses early induction by providing glucose, a repressor of the *lac* operon, as the primary carbon source during the early growth phases. Once the glucose has been consumed, expression is induced by the lactose content of the medium, and lactose as well as glycerol continue to serve as carbon sources. Whilst this approach did not lead to soluble protein either, during the course of this work it was noted that the use of complex, auto-inducing medium significantly simplifies the logistics of expression large numbers of constructs simultaneously, as well as improving possible yields by allowing higher culture densities to be reached.

Lysis condition screens

Beyond expression parameters, lysis conditions are another important aspect to take into account during the quest for soluble protein. This includes both buffer conditions during cellular disruption as well as the disruption method itself. With respect to the latter, both sonication and traditional French press were evaluated, yet neither showed improved solubility.

Buffer conditions during cell breakage are a much more complex topic that has been shown to be able to affect solubility of recombinant protein to a great effect [194]. To this end, a systematic lysis buffer screen was employed to determine optimal conditions covering the following parameters: pH, salt concentration, detergents, denaturant and stabilizers. The screen itself is a modified version of the solubility screening protocol of Jeanne Perry of the Molecular Biology Institute, UCLA, LA, which to the author's knowledge has not been published through peer review and is therefore described in detail in Section 10.2.5. Despite the fact that this approach proved to be an efficient way to cover many buffer parameters, the results differed little from the trials described so far and yielded only insoluble target protein in the form of inclusion bodies. As with all trials so far, this behaviour was observed for both the Histidine tagged, as well as the untagged protein (see Fig. 11.1).

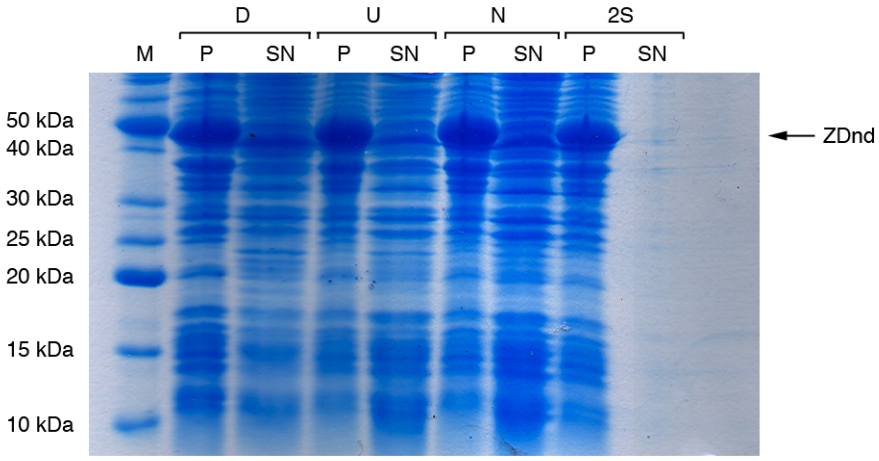

Figure 11.1: SDS-PAGE analysis of an initial lysis buffer screen of full-length ZDnd expressed in *E. coli*.
Lane 1, labelled **M**, contains the molecular weight marker, the molecular weight of the ZDnd protein is indicated by the arrow on the right. Lanes 2-9 contain insoluble (**P**) and soluble (**SN**) fractions of four different buffer conditions: detergent (**D**), denaturant (**U**), standard (**N**) and high salt (**2S**).

Purification trials

Depending on the background of soluble endogenous *E. coli* proteins it may not always be possible to deduce with certainty, from an SDS-PAGE experiment alone, whether a heterologously expressed protein is 100% insoluble. Whilst the bulk of the target protein may have found its way into inclusion bodies, it is still possible that a small proportion is indeed in the soluble fraction under certain lysis conditions. This is the case also with the SDS-PAGE analysis of the lysis buffer screen shown in Fig. 11.1.

The target protein, full-length ZDnd is this example, can be enriched by affinity or ion exchange chromatography to ascertain the presence of the suspected species in the soluble cell fraction. As described in Section 10.2.5, the high affinity of the His_7-tag to Nickel ions was exploited to attempt to enrich the target protein. This process is shown in Fig. 11.2. From the two step elution protocol carried out, it is evident however, that the soluble cell fraction did not contain detectable amounts of

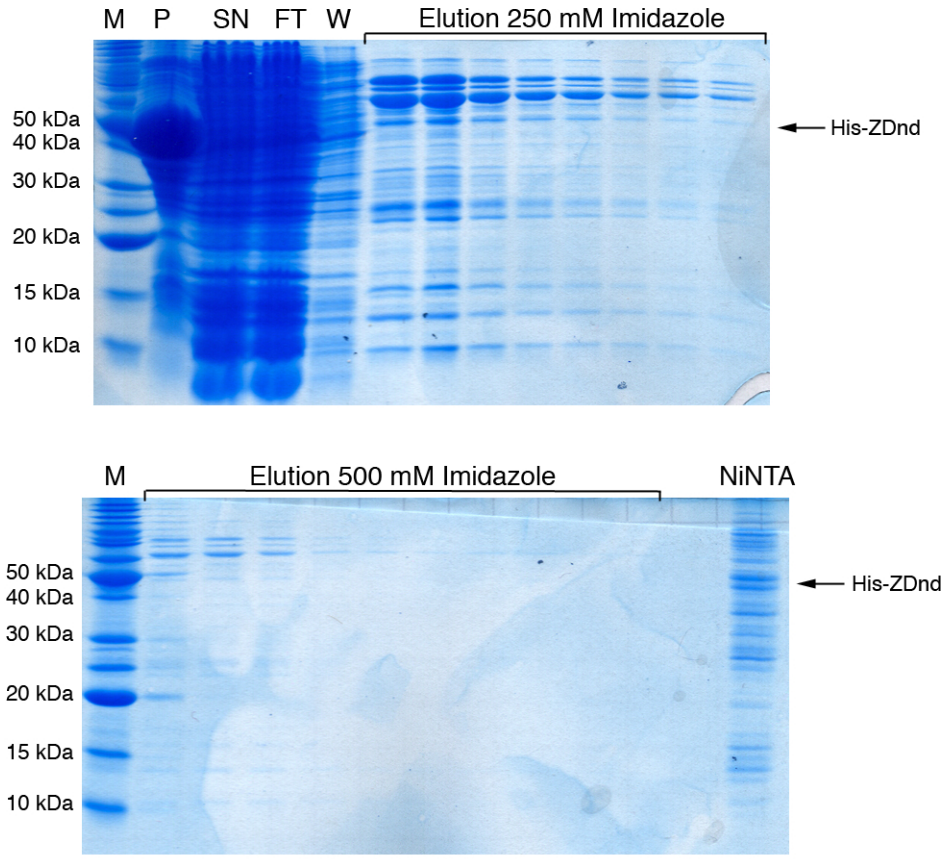

Figure 11.2: SDS-PAGE analysis of NiNTA affinity chromatography of His_7 tagged full length ZDnd.
Lane 1 in both gels, labelled **M**, contains the molecular weight marker. The insoluble (**P**) cell fraction is shown for reference whilst the soluble (**SN**) cell fraction was incubated with NiNTA agarose. Proteins that did not bind the resin are removed by the flow through (**FT**) and the washing step (**W**). Elution was carried out in two steps: 250 mM imidazole and 500 mM imidazole (2 mL fractions). The final lane (**NiNTA**) shows the NiNTA resin after elution. The target ZDnd protein is only present in the cell pellet and can not be enriched by NiNTA affinity chromatography. The molecular weight of the His_7-tagged ZDnd protein is indicated by the arrow on the right.

target protein. Like the imidazole eluted fractions, the NiNTA resin did not contain the target protein either.

Expression in *E. coli* - Rosetta2

In addition to the factors mentioned above, codon bias has also been shown to influence the efficiency of heterologous protein expression in *E. coli* as well as the ability to be purified [195]. So called rare codons, for which the *E. coli* genome either does not encode corresponding tRNAs, or codons whose corresponding tRNAs are produced in insufficient levels, may not only decrease yields, but also hamper correct folding. A number of bioinformatics tools are available that allow sequences to be analysed towards codon bias, and indeed the Dead end mRNA sequence does contain an above average number of rare codons.

Two principal ways have been proposed to overcome codon related issues. Either through the design of synthetic genes with optimized codon distributions or through the co-expression of additional tRNAs recognizing rare codons. For the purpose of screening, the latter approach is significantly more efficient, especially if several isoforms of a protein may have to be tested. Hence, the commercially available *E. coli* strain Rosetta2 (DE3) was chosen to test the effect of codon bias, as it carries the pLacIRARE2 plasmid encoding seven rare tRNAs. However, expression in this strain did neither increase overall expression levels nor lead to observable increases in solubility of the Dead end protein.

Expression in insect cells

The lack of Eukaryotic cofactors and/or the Eukaryotic machinery for post-translational modifications in *E. coli* might be one reason for the observed insolubility of the Dead end protein. Without detailed knowledge of required factors it is nearly impossible to overcome an obstacle of this kind in the realms of the *E. coli* expression system. To overcome the inherent limitations of Prokaryotic expression, Eukaryotic expression systems have been developed. Whilst they do suffer from a number of drawbacks, such as reduced efficiency, increased costs and labour intensity, they are able to provide Eukaryotic cofactors and chaperones, as well as post-translational modifications and disulphide bridge formation, each of which may be a deciding factor with respect to solubility.

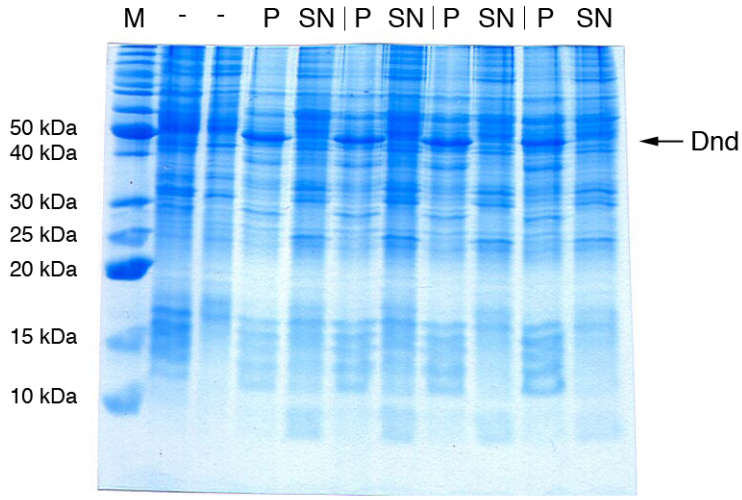

Figure 11.3: Expression of ZDnd in insect cells through the baculovirus system. Lane 1, labelled **M**, contains the molecular weight marker. Lanes 4 - 11 show the insoluble (**P**) and soluble (**SN**) fractions of four successfully transfected clones of Sf9 insect cells five days after transfection. The Dead end target protein (**Dnd**) is exclusively present in the insoluble cell fraction, its molecular weight is indicated by the arrow on the right.

One of the most versatile and well-established Eukaryotic expression systems is the Baculovirus system. Utilizing a viral transfer vector encoding the target protein, heterologous expression of recombinant protein is achieved in *S. frugiperda* Sf9 insect cells[2].

Expression tests of wild-type ZDnd in insect cells were conducted on four successfully transfected clones, but the baculovirus expression system was not able to deliver the much desired soluble protein either (Fig. 11.3).

Expression as a fusion construct

Beyond cultivation conditions (medium, temperature, rare codon compensation, etc.) the nature of the heterologous protein itself is a common obstacle in many cases. A common strategy to improve solubility whilst maintaining the sequence of one's target protein, is the use of solubility enhancers in the form of fusion part-

[2]for reviews on heterologous expression in the baculovirus system see [196] and [197].

ners. Maltose binding protein (MBP) has emerged as a particularly useful tool over the last decade [198, 199]. In addition to being a highly soluble protein itself, it was shown to also positively influence the solubility of fusion partners it is conjugated to[3]. Further to its solubility enhancement, MBP fusion constructs also lead to higher yields in many cases, which is attributed to the juxtaposition of the N-terminal MBP domain to the ribosome binding site. Finally, MBP itself can be exploited as an affinity tag in conjunction with amylose columns.

Consequently, the zebrafish *dnd* coding sequence was cloned into the pMal-TEV vector, which allows expression with an N-terminal MBP fusion protein separated by a tobacco etch virus (TEV) protease cleavage site. To assess the effect of the primary sequence on solubility, the *Xenopus laevis* Dead end isoform was also expressed as a MBP fusion protein containing a FactorXa cleavage site.

Expression in *E. coli* produced a familiar picture - high amounts of insoluble protein in inclusion bodies. But, behold, the soluble fraction also contained small amounts of MBP-Dnd fusion protein. This could be observed for both isoforms, *D. rerio* and *X. laevis*. Similarly as to the His_7-ZDnd construct described above, affinity chromatography was employed to purify the target protein, through which both fusion proteins could be successfully purified on amylose resin.

Unfortunately however, the resulting proteins were very prone to aggregation and degradation. During the purification, the fusion constructs were eluted in fractions of 1 mL, resulting in a concentration gradient of the target protein across different fractions. Fractions strongly enriched in MBP-fusion protein showed severe aggregation, degradation and precipitation, irrespective of whether the purification was carried out at room temperature or at 4 °C. Fractions containing lower concentrations of the target protein did show less aggregation, but were not stable over periods exceeding several hours either. Fig. 11.4-A shows purification of the MBP-XDnd construct, Fig. 11.4-D of the MBP-ZDnd construct.

Continuing with less concentrated fractions, the efficiency of protease cleavage in the fusion construct linker was assessed. TEV protease cleavage for the ZDnd

[3]Whilst glutathione S-transferase (GST) is known to be highly soluble itself, has been shown to have very little ability to enhance the solubility of a fusion partner [198, 199]. Contrary to common notion, a highly soluble protein does not automatically constitute a powerful solubility enhancer.

construct, as well as FactorXa cleavage for the XDnd isoform were successful, but did lead to increased aggregation of the Dead end protein in both cases. SDS-PAGE analysis showed that the majority of the precipitate that forms upon cleavage is aggregated Dead end protein, however small amounts of it remain in solution together with its former fusion partner, MBP (see Fig. 11.4-B, C and E).

Upon successful protease cleavage, purification trials were undertaken for both Dead end isoforms. After initial difficulties had been overcome in bringing the MBP/Dead end mixture into a buffer system suitable for ion exchange chromatography, both anion and cation exchange chromatography were employed in order to purify the target protein. However, irrespective of which chromatographic method was used, it has proven impossible to separate the Dead end protein from its former fusion partner. This was the case both for standard buffer systems, as well as detergent supplemented buffers.

11.1.3 Construct design

Using a variety of approaches as laid out above, it has not been possible to obtain and purify soluble full length protein, the reasons for which may be manifold. A possible cause lies in the nature of the protein containing domains that require specific cofactors for correct folding, or hydrophobic surfaces that require an interaction partner. The next logical step to take in this case, is to abandon the full length protein and to seek out a suitable construct that is not subject to the aforementioned limitations. As discussed in Section 11.1.1 (p. 115), the ideal construct contains the functional site and is of small size, in addition to being amenable to sample production for use with structural biological methods. Consolidating all information available, including biological data as well as *in silico* bioinformatics tools, a rational approach to construct design was taken to maximise the chances of success.

Functional data

First clues towards a minimal yet functional construct stemmed from deletion and mutation studies conducted in the laboratory of Erez Raz [200, 201]. The majority of single amino acid mutations that affect Dead end localization and impair its

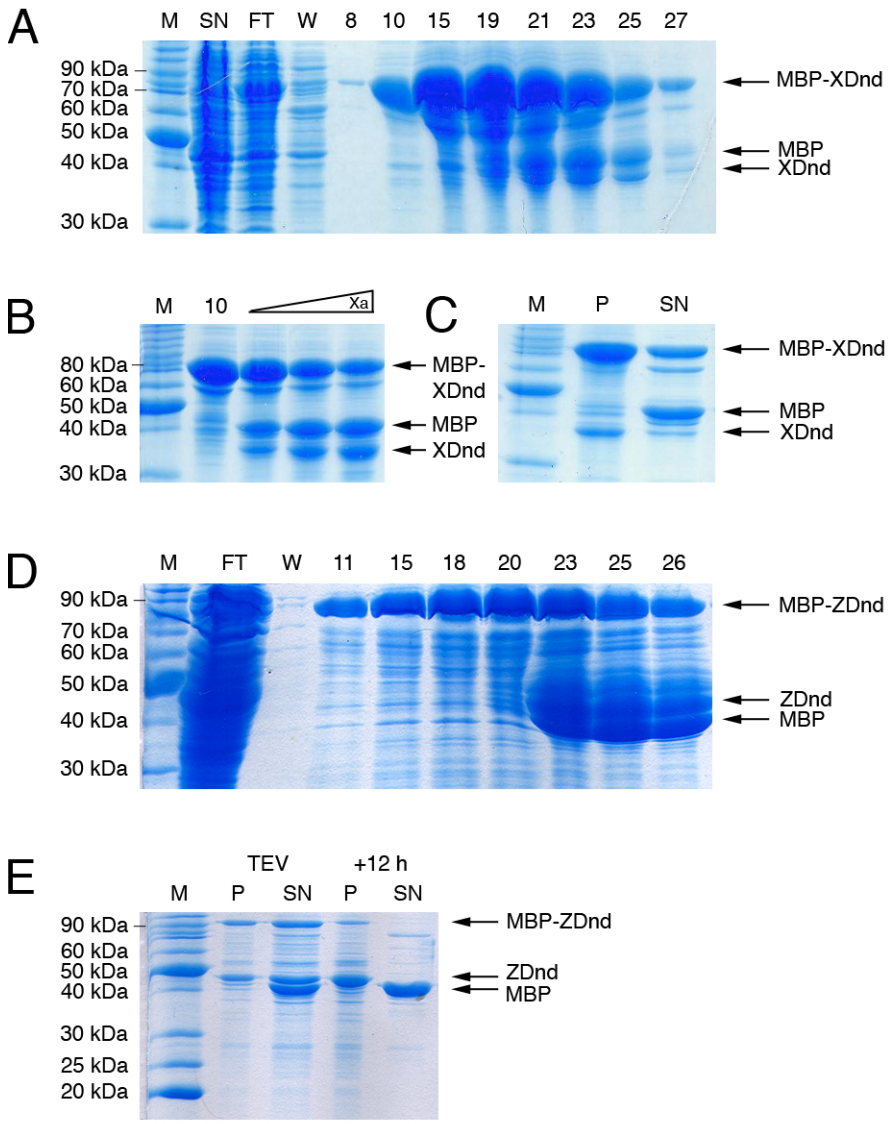

Figure 11.4: SDS-PAGE analysis of expression, purification and protease cleavage of MBP fusion constructs.

A. Amylose affinity chromatography of the MBP-XDnd fusion protein.

B. Protease cleavage of the MBP-XDnd fusion with different concentrations of FactorXa

C. Soluble (SN) and insoluble (P) fractions after FactorXa cleavage

D. Amylose affinity chromatography of the MBP-ZDnd fusion protein.

E. Soluble (SN) and insoluble (P) fractions after TEV cleavage of MBP-ZDnd.

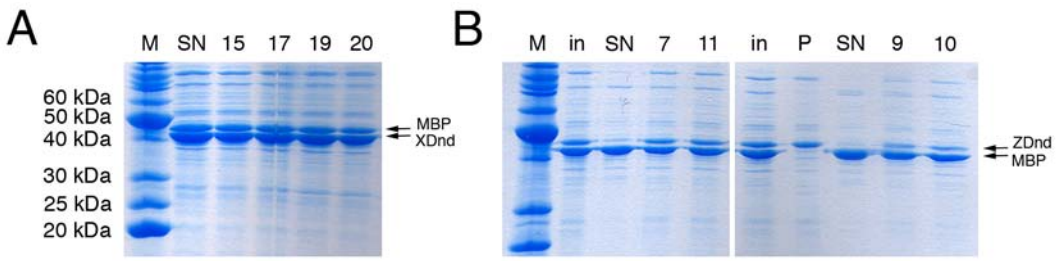

Figure 11.5: SDS-PAGE analysis of ion exchange chromatography to separate ZDnd from its MBP fusion partner.
A. SDS-PAGE analysis of a cation exchange chromatography of cleaved MBP-XDnd on a HiTrap^{TM}SP XL column. **M** denotes the molecular weight marker, **SN** the starting material before chromatography, 15-20 denote eluted fractions (1 mL). It was not possible to separate the target protein (XDnd) from its former fusion partner, MBP.
B. SDS-PAGE analysis of a cation exchange chromatography of cleaved MBP-ZDnd on a HiTrap^{TM}SP XL column. The gel on the left shows fractions of a chromatographic run under standard buffer conditions, the gel on the right shows fractions from a detergent supplemented buffer. In each gel, **in** denotes the starting material before dialysis into the chromatography buffer, **P** and **SN** refer to the insoluble and soluble fractions after dialysis, whilst the numerically labelled lanes contain 1 mL fractions after elution from the column. It was not possible to separate the target protein (ZDnd) from its former fusion partner, MBP.

function *in vivo* map to the putative RRM (amino acids 61-134 in the zebrafish isoform, c.f. Fig. 11.6-A). C-terminal deletion constructs strengthened the importance of this domain, as N-terminal fragments spanning as little as amino acids 1-220 and 1-285 showed significant rescue of the *dnd* phenotype in zebrafish embryos. Further C-terminal truncation, however, abolished Dead end function yielding the first boundary to keep in mind.

In order to assess whether the entire N-terminus is required *in vivo* to rescue the *dnd* phenotype, a further construct encompassing amino acids 45-220 was cloned. This construct was co-injected into zebrafish embryos at the one cell stage together with Dead end morpholino. The morpholino reagent is a complementary degradation resistant RNA analogue that pairs with the wild-type mRNA of the target protein

thereby inhibiting translation and abolishing expression. Co-injection of *in vitro* transcribed RNA containing a number of point mutations such that this RNA is not complementary to the morpholino, yet yields the correct protein sequence when translated, thus enabling one to assess the effect of a certain construct in the context of a live zebrafish embryo during its early stages of development[4]. However, the 45-220 construct failed to show significant rescue in zebrafish, thus suggesting that a minimal, functional construct ought to contain amino acids 1-220, as indicated by the mutational screen and N- and C-terminal deletion constructs.

Structural data and predictive approaches

Structural data on the Dead end protein is extremely sparse, and in effect limited to *in silico* prediction tools. Nevertheless, the accuracy of such tools has been improving, especially with respect to secondary structure prediction. Both secondary structure prediction, as well as disorder prediction can provide useful clues in finding and optimising boundaries for suitable expression constructs.

Four different programs (JPred, PredictProtein, PSIpred, SOPMA) predicting secondary structures were used to analyse the ZDnd protein sequence giving rise to the 'consensus' predicted secondary structure shown in Fig. 11.6-B. When defining construct boundaries it should be chiefly avoided to set them within secondary structure elements, as this may have disastrous consequences on folding and solubility. Information drawn from secondary structure prediction can be complemented by its counterpart, disorder prediction. To this effect the two, themselves complementary, programs DisEMBL and GlobPlot were used to predict regions of high disorder, which often occurs in loops linking secondary structure elements or even more so, protein domains (see Fig. 11.6-C). On a larger scale, pattern and profile searches can identify known folds and domains within a sequence on the basis of homology to known structures and thereby significantly contribute to the rational design of expression constructs. In the case of Dead end (*D. rerio*) however, only one homologous domain, an RRM fold spanning amino acids 61-134 is predicted,

[4]As the embryo grows and cell division ensues, the morpholino injected at the one cell stage is eventually diluted out, thereby abolishing its effect.

although an additional domain boundary is suggested in the center of the sequence (see Fig. 11.6-A).

Combining all this information whilst keeping two additional nuances in mind, namely that poly-proline motifs often constitute breaks in secondary structure elements and therefore lend themselves to construct boundaries, and the fact that experience has shown hydrophilic amino acids at construct termini to yield more successful constructs, the constructs listed in Fig. 11.7 were cloned. As with the full-length protein, each construct was cloned into a tag-less vector and a vector encoding a His_7 affinity tag.

Due to the strong indications that the Dead end protein harbours a canonical RRM motif, a literature review was conducted to identify solved RRM structures and the underlying constructs and purification procedures. The construct boundaries of RRMs in the literature were found to agree well with the boundaries determined by the aforementioned *in silico* approaches. The majority of purification protocols used standard ion exchange chromatography (e.g. [202, 203, 204, 205]) or Histidine affinity chromatography (e.g. [204, 206, 207]), with the exception of one RRM core domain that has been successfully expressed and purified as a GST fusion construct ([208]). Accordingly, in addition to the untagged and His_7-tagged constructs, the minimal RRM construct of ZDnd was also cloned into two different GST fusion vectors (pGEX-2T and pGEX-2TEV).

11.1.4 Expression of Zebrafish Dead end constructs

With the new constructs in hand, expression tests were carried out in *E. coli* to identify soluble candidates. As initial tests failed, the alternative strategies discussed in Section 11.1.2 were also applied to each of the new constructs. However, despite testing different cultivation temperatures, growth media and lysis buffers, no soluble protein could be obtained in this way.

Of note, in the case of both GST fusion constructs, translational abortions were observed leading to GST protein in the soluble fraction without being conjugated to the RRM motif. The GST-RRM fusion protein was also expressed, but absent from the soluble fraction and instead solely present in inclusion bodies.

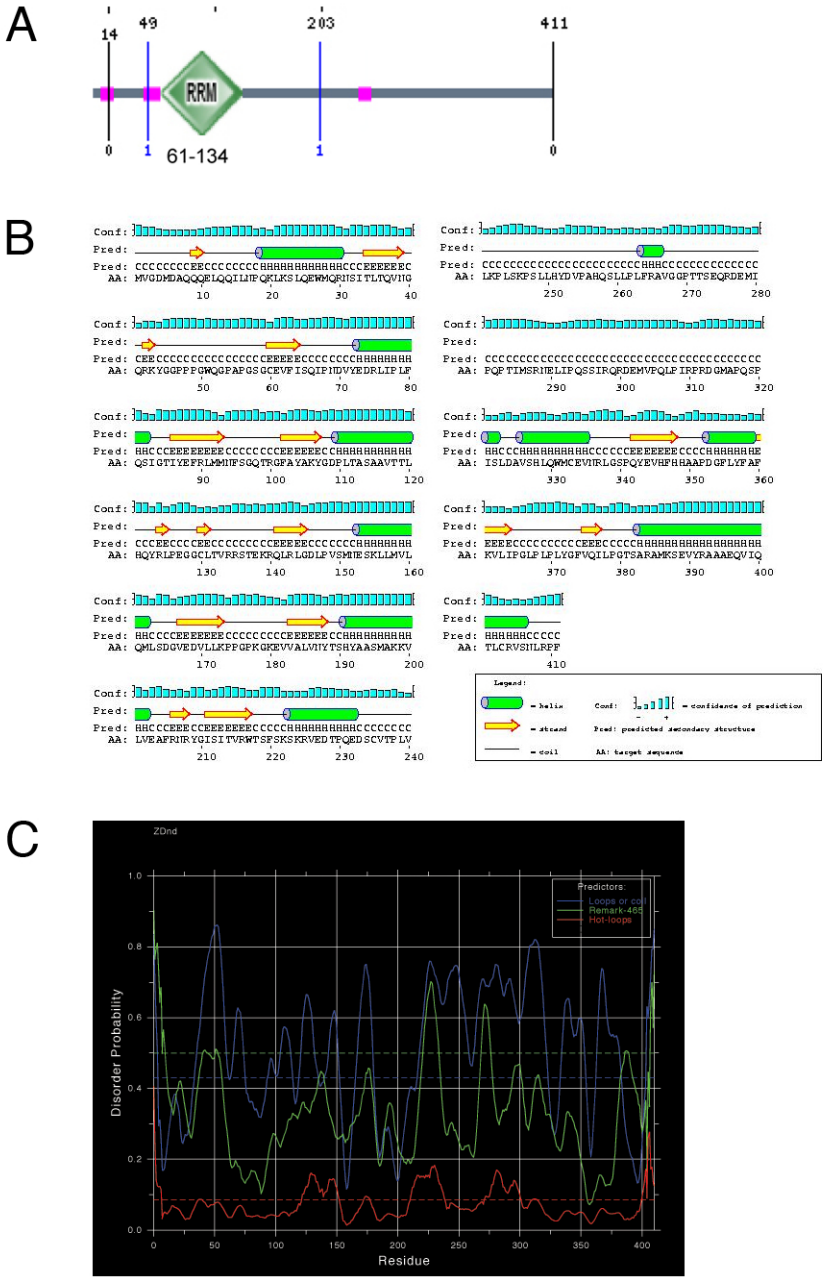

Figure 11.6: Profile searches, secondary structure and disorder prediction on ZDnd.

A. Pfam profile search.

B. Consensus secondary structure prediction.

C. DisEMBL disorder prediction.

Figure 11.7: ZDnd constructs designed and cloned.
The RRM motif (aa 61-134) is shown in green, a possible second RRM, not identified in ZDnd, however present in other Dnd isoforms, is also indicated. The N- and C-terminal boundaries are indicated by their amino acid position. Affinity tags are indicated in maroon.

Purification attempts

As with the full-length ZDnd protein, SDS-PAGE analysis of the soluble cell fraction could show with certainty that a given construct was 100% insoluble. If the molecular weight of a given construct is close to that of a soluble, endogenous *E. coli* protein, it is necessary to verify the presence or absence of the target protein either through chromatographic enrichment or western blotting.

Due to the lack of a reliable antibody at the time, purification trials using affinity or ion exchange chromatography were conducted. However, as in the case of full-length Dead end, no target protein could be detected in any of the soluble fractions.

11.2 The Human Dead end isoform

Due to the lack of success in obtaining soluble protein of the zebrafish isoform, the focus of the work was shifted to a different isoform, human Dead end (HDnd). Whilst

the function of both isoforms seems to be conserved, they differ on the primary sequence level, especially in a large non-conserved region in the center of ZDnd. Experience in the laboratory has shown that expression efficiency and solubility in *E. coli* can differ vastly between different isoforms, and that this behaviour is unpredictable. Utilizing strategies analogous to the ones described for the Dead end isoform, in addition to the full-length protein, a number of HDnd constructs were designed and expression and purification trials conducted in order to obtain soluble protein.

11.2.1 Construct design

The rationale behind designing the HDnd constructs was identical to the one laid out above in Section 11.1.3 dealing with construct design of the ZDnd isoform. Consolidating functional data, secondary structure and disorder predictions, pattern and profile searches, as well as homologous structures and procedures identified through literature research, three N-terminal and three C-terminal boundaries were identified. Secondary structure information on the basis of which the boundaries were chosen, are presented in Fig. 11.8 and Fig. 11.9. No additional functional data, other than for the ZDnd isoform as discussed above, was available at this time. Homologous structures in the literature and corresponding purification strategies are identical to the ones drawn upon in the design of the ZDnd constructs (see Section 11.1.3). The resulting HDnd constructs are shown in Fig. 11.10.

11.2.2 Expression

E. coli **- BL21 (DE3)**

The full length HDnd protein, as well as the other constructs listed in Fig. 11.10 were expressed in the standard expression strain BL21 (DE3). As initial tests did not yield soluble protein, a systematic screen of expression conditions was conducted. Following the rationale laid out in Section 11.1.2, low temperature cultivation, alternative growth media and heat shock inducing supplements were tested, however

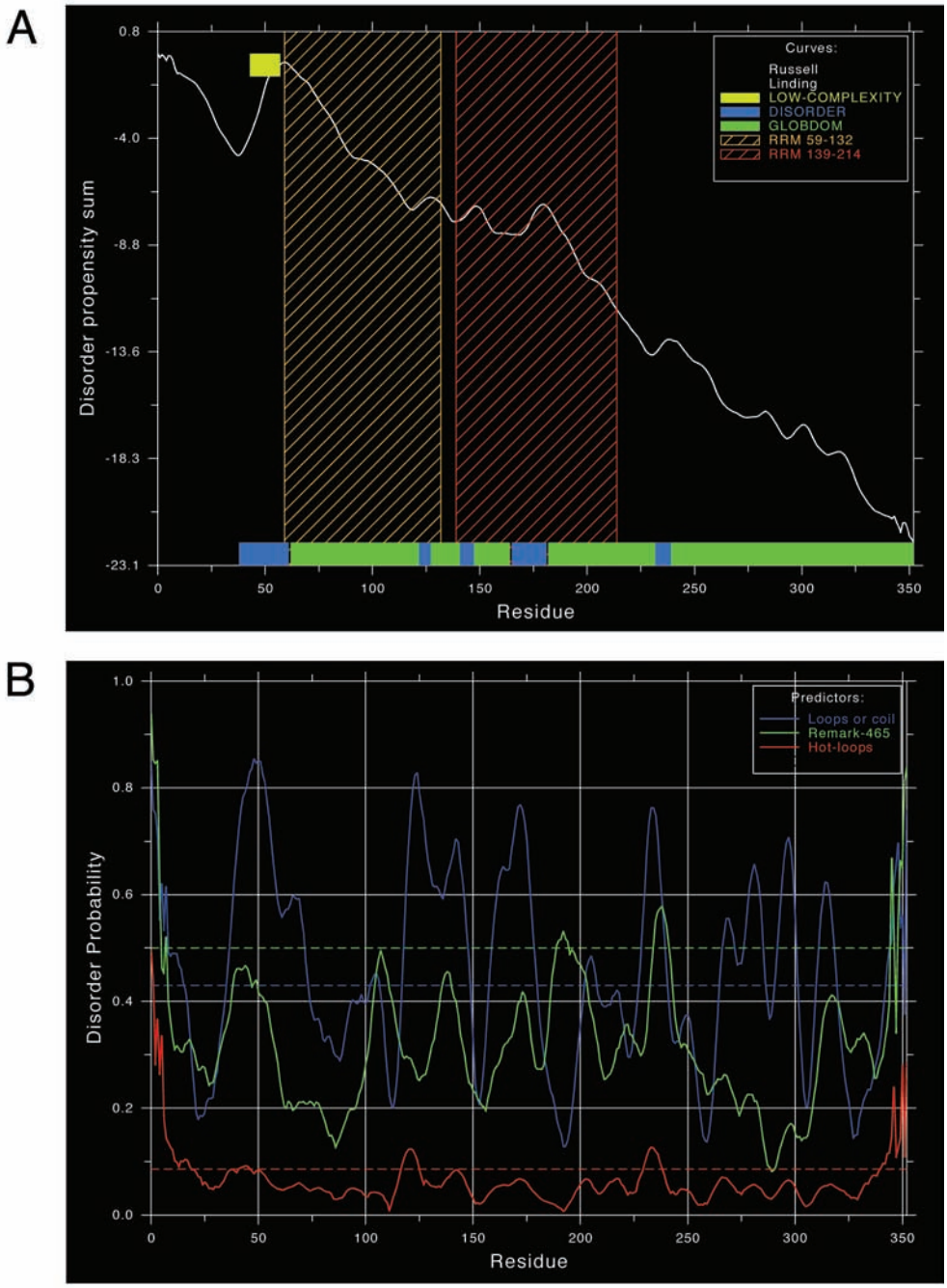

Figure 11.8: Order and disorder predictions for the HDnd isoform.

A. GlopPlot order prediction.

B. DisEMBL disorder prediction.

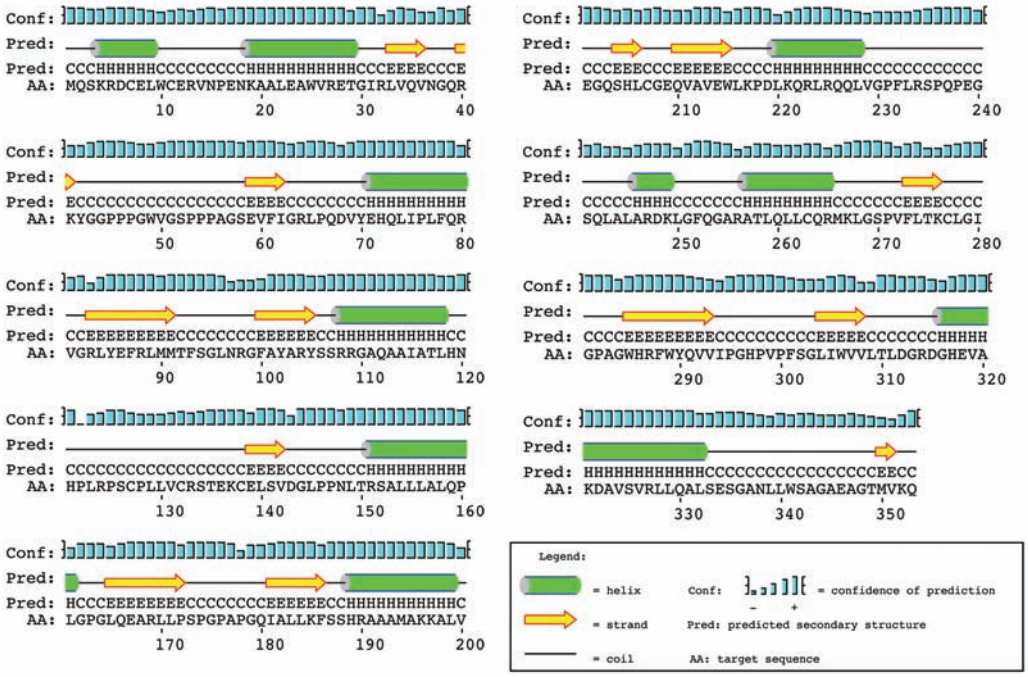

Figure 11.9: Consensus secondary structure prediction of the HDnd isoform.

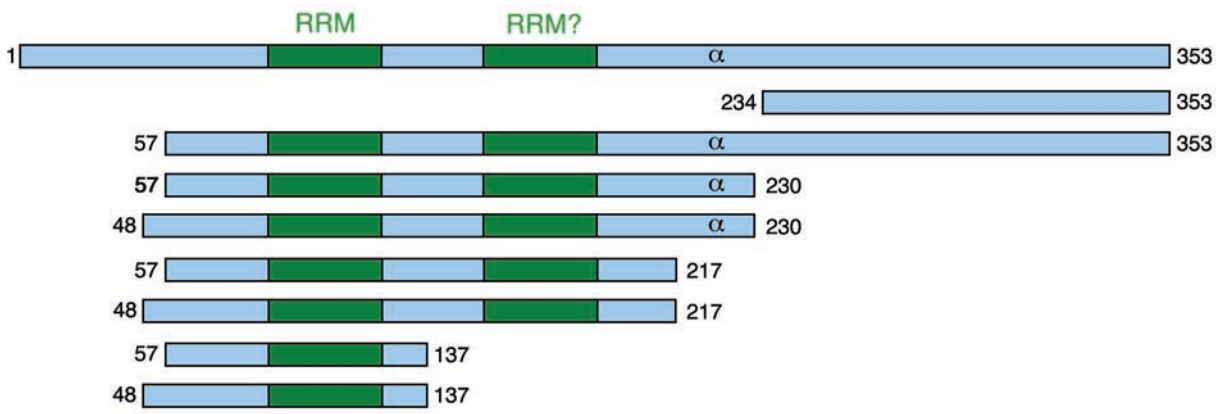

Figure 11.10: HDnd constructs designed and cloned.
RRM motifs (aa 59-132 and 139-214) are shown in green. The N- and C-terminal boundaries are indicated by their amino acid position. Each construct was cloned twice, once without affinity tag, and once with an N-terminal His_7 tag.

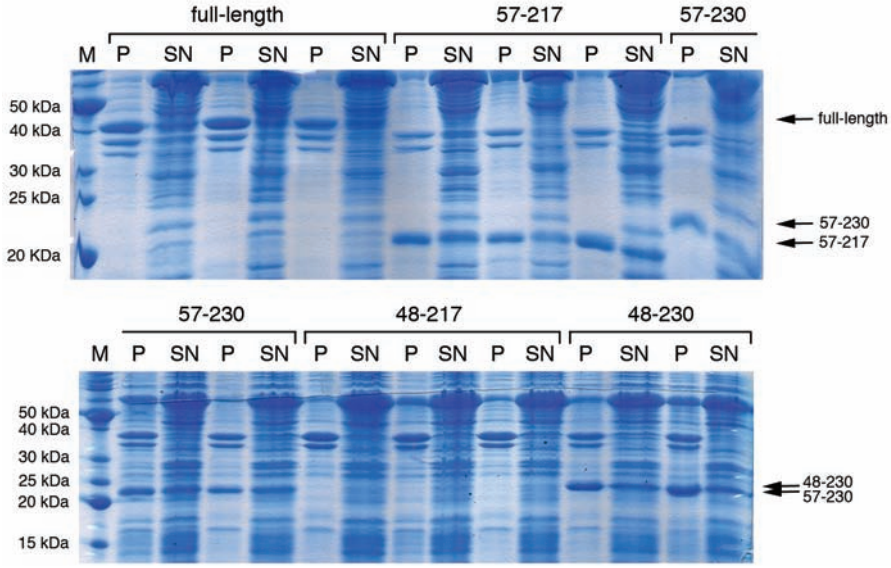

Figure 11.11: SDS-PAGE analysis of HDnd expression in the Arctic Express^{TM} strain. Representative samples from lysis buffer screens of five different HDnd constructs are shown. Molecular weight markers are indicated by **M**, insoluble and soluble cell fractions by **P** and **SN** respectively, the constructs are indicated on the top. The majority of heterologously expressed protein is in the insoluble fraction, but part of it is also contained in the soluble fraction. Expression of the 48-217 construct could not be detected. Molecular weights of the expressed constructs are indicated by the arrows on the right.

to no avail. None of the constructs could be expressed in a soluble form using any of the aforementioned measures.

E. coli - Rosetta2

Like the zebrafish isoform, *h. sapiens* Dead end constructs were also expressed in the special strain Rosetta2 that carries the pLacIRARE2 plasmid encoding seven rare tRNAs. Codon bias could be excluded as sole cause of insoluble protein, as none of the constructs showed improvement solubility or expression levels under these conditions.

Expression in *E. coli* - Arctic Express™

As laid out in Section 11.1.2, low temperature cultivation represents a widely employed strategy to overcome insolubility of heterologously expressed proteins in *E. coli* [188, 189]. A drawback of this method however, is the resulting decreased activity of endogenous *E. coli* chaperonins. Facilitating correct folding at their optimum temperature of 30 °C, endogenous chaperonins suffer significant reductions in activity at lower temperatures, resulting in the opposite effect as the one amplified by the addition of ethanol as described above [186]. Recently, the new *E. coli* strain Arctic Express™ has become available that combines the strengths of both strategies by providing chaperonins that retain their activity at temperatures as low as 10 °C. This is achieved by the co-expression of the cold-adapted chaperonins *Cpn10* and *Cpn60* from the psychrophilic bacterium *Oleispira antarctica*. The fact that these foldases are constitutively co-expressed from their own dedicated plasmid also has the additional positive side effect of highly elevated expression levels when compared to endogenous chaperones. Together with the ultra-low cultivation temperature this provides a much improved environment for correct folding of heterologously expressed proteins.

Following the procedure described in Section 10.2.3 on p. 108, all His_7-tagged HDnd constructs were expressed in Arctic Express™ cells at 11 °C. Additionally, the full-length zebrafish isoform was also tested in this strain. As expected due to severely reduced growth rates at low cultivation temperatures, protein yields were low compared to other expression strategies. With the exception of one construct however, protein production was successful in this system.

When subjected to a lysis buffer screen, no impact of different buffer systems could be observed. The majority of heterologously expressed proteins had accumulated in inclusion bodies, however small amounts of target protein were present in every lysis condition tested. Fig. 11.11 shows representative SDS-PAGE analyses of five different constructs with three lysis buffer conditions each.

Subsequently, purification trials were conducted to isolate the recombinant target protein produced in Arctic Express™. Independent of the construct tested, this led to two recurring observations. Firstly, only low yields of the target protein can

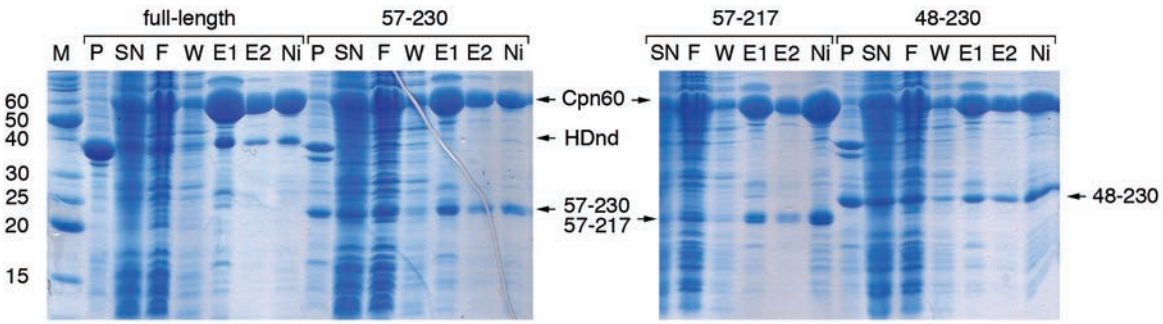

Figure 11.12: SDS-PAGE analysis of Ni affinity chromatography trials of constructs expressed in Arctic Express™
Lanes are labelled as follows: **M** - molecular weight marker, **P** - insoluble cell fraction, **SN** - soluble cell fraction, **F** - flow-through, **W** - fraction of the washing step, **E1** - fraction of elution step 1 (200 mM imidazole), **E2** - fraction of elution step 2 (400 mM imidazole), **Ni** - NiNTA resin after elution. Molecular weights are indicated in kDa on the left, arrows indicate molecular weights of the target proteins and the *O. Oleispira* charperone Cpn60.

be recovered as large amounts can not be eluted from the column. It has remained unclear whether the protein is prone to aggregation during the purification procedure, or whether it strongly associates with the resin matrix. The second observation is even more striking. Whilst it was possible to conduct a first purification step on the target protein removing the majority of endogenous *E. coli* proteins, it proved impossible to seperate it from the coexpressed chaperonin *Cpn60*. Neither affinity chromatography, nor the addition of ATP to trigger the release from the foldase were able to overcome this strong interaction. Fig. 11.12 shows representative SDS-PAGE analyses of such purification trials.

11.2.3 Point mutations

In the course of extensive literature review conducted in the construct design stage (see Section 11.1.3), the case of one particular expression strategy caught attention. Troubled by the limited solubility of RNA binding domain of the well-known *Drosophila* determinant Sex-lethal (Sxl), Inoue and coworkers undertook a point mutation screen exchanging hydrophobic putative surface residues for more hydrophilic ones

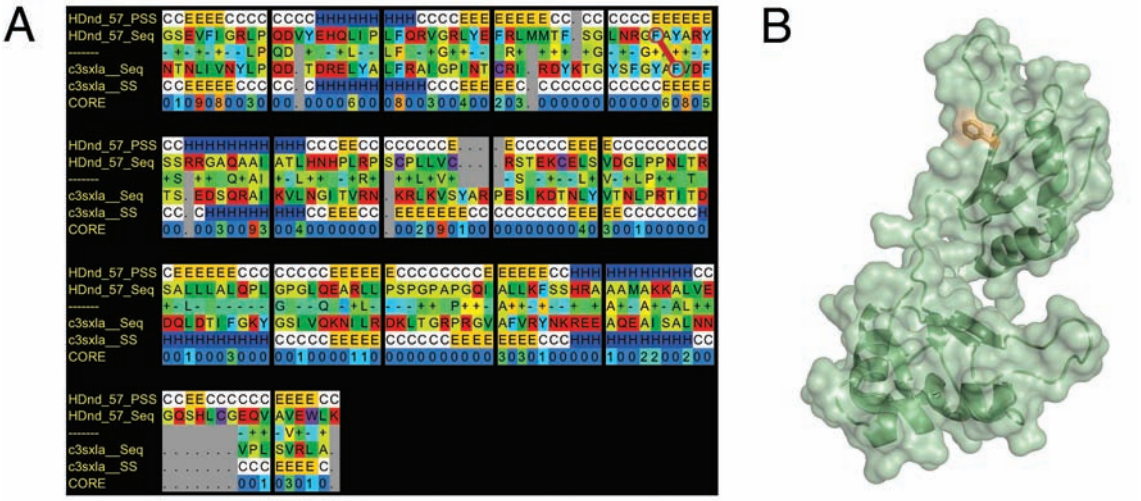

Figure 11.13: Dead end structure model and basis for point mutations.
A. The basis of the structural homology model. Shown are the secondary structure prediction for HDnd (HDnd_57_PSS), the HDnd aa sequence (HDnd_57_Seq), the sequence of sex-lethal (c3sxl_Seq), the known secondary structure of sxl (c3sxl_SS), a measure of surface accessibility/hydrophobic contacts (CORE, 0 – surface/no contacts, 9 – buried/many contacts). A consensus amino acid sequence is shown in the centre, colour coded from good (orange) to bad (blue) agreement. The residues subjected to point mutations are highlighted in red.
B. Structural model of the HDnd fragment 57-217 based on the PDB structure 3sxl of sex-lethal. The target of the point mutation, F100, is highlighted in orange.

in order to engineer a more soluble protein [209]. One particular mutation, 166 Phe → Tyr, dramatically raised the to date existing, yet limited solubility of Sxl about ten-fold.

Given the homology between the RRMs of the Dead end protein and other RNA binding proteins, among them Sxl, there is a chance that this strategy could also be successful in the quest for soluble Dead end. Using both classical sequence alignment algorithms and a number of sophisticated tertiary structure prediction approaches, the Dead end protein sequence was searched for similar residues amenable to point mutation. Looking for hydrophobic residues, that are likely to be surface accessible in the tertiary structure of the protein and that occur in a similar environment or

position as Phe 166 in Sxl, Phe 100 was identified as the most promising candidate. Fig. 11.13 shows sequence alignment of Dnd with Sxl, secondary structure elements as well as buried and surface accessible residues in the context of a tertiary structure prediction assuming Sxl as a model for the Dnd RRM.

Expression of point mutation constructs

Using site-specific mutagenesis by overlap extension, a F100 → Y was introduced into six HDnd constructs[5] that are of similar size than the Sxl construct used in [209]. Analogous to all expression tests carried out so far, these constructs were tested under several cultivation conditions and subjected to a lysis buffer screen. Unlike the example in the literature however, none of the tested constructs showed any improvements in solubility.

11.3 Refolding

11.3.1 Inclusion body anatomy

Despite the variety of different strategies employed thus far, the deposition of insoluble protein in the form of inclusion proteins has presented a common obstacle to all experiments. The classical view of inclusion body anatomy describes them as aggregates formed by unspecific hydrophobic interactions between disorderly deposited polypeptides, often coated by chaperones like DnaK [210]. Several studies however, have challenged this view of a 'molecular dust-ball' and suggested that proteins aggregate mainly with themselves due to specific interactions or folding intermediates, rather than non-specific coaggregation [211], thereby providing a reservoir of alternative conformational states [212]. Other reports have characterised inclusion bodies as amyloid-like and comprised of amino-acid sequence-specific cross-beta structures [213, 214]. Whilst the identified processes do lead to relatively high purity of the target proteins within the inclusion body reservoir, in its aggregated state it is of little use to the structural study at hand.

[5]The following constructs were mutated:
57-137, 48-137, 57-217, 48-217, 57-230, 47-230; c.f. Fig. 11.10.

Structural biologists have been facing this situation repeatedly, not limited to, but very evident through recent advances in structural genomics. Most strategies to counteract this problem focus on the prevention of inclusion formation, or the use of chaperones to revert aggregation, as discussed above [215]. The accumulation of insoluble protein, not only in bacterial cytoplasm, but also in the hands of many scientists however, has led to the development of alternative approaches, most notably refolding[6].

11.3.2 Important parameters

Refolding strategies taking advantage of high target protein purity within the aggregates generally follow these steps: inclusion body isolation, solublilisation/denaturation and subsequent refolding. The first two steps, isolation and denaturation, are relatively straightforward to accomplish. The actual refolding process however is of more delicate virtue as aggregation and folding continue competing with each other.

One key aspect is to keep the protein concentration to a minimum during the refolding procedure in order to minimise the chances of recurring aggregation. Various techniques achieving this have been developed, the most popular ones among them being direct dilution and membrane controlled denaturant removal, although chromatographic and matrix-assisted methods have also evolved. Other important factors contributing to refolding efficiency have been identified as physical parameters such as temperature and pressure, as well as a plethora of chemicals. The long list of commonly used additives is somewhat reminiscent of crystallisation conditions and covers low molecular weight non-detergent zwitterionic agents, polymers, organic ligands, chaotrophs, hydrophobic shields, micelles, buffer systems, disulphide bridge reshuffling systems and chaperones to name a few [218, 219]. Knowledge of these parameters in conjunction with recently introduced screening systems have made this approach viable.

[6] for reviews see [194, 216, 217]

11.3.3 Refolding screens

Given the large number of factors influencing refolding efficiency, it is necessary to employ a systematic screen. This was reconfirmed by initial small scale refolding trials which were unsuccessful. Therefore, a large scale refolding screen was conducted, which in the meantime had become commercially available in the form of the iFold^{TM}system (Novagen). In this system, following inclusion body purification and solubilization, the target protein is refolded by rapid dilution into a 96-well matrix. Refolding efficiency may subsequently be estimated by recording the absorbance at 340 nm (A_{340}), where polypeptide UV spectra exhibit a minimum in absorption. In contrast, protein aggregates approaching or beyond the size of this wavelength scatter light irrespective of their spectral properties. Consequently, low readings at A_{340} correspond to little insoluble protein in a given refolding condition.

However, as small amounts of precipitate do not necessarily go hand in hand with correct folding of the protein of interest, a more direct way of identifying successful conditions is required. Ideally this is an activity assay that can be coupled to a UV-spectroscopic or luminescence assay in 96-well format, thereby constituting a very powerful and efficient screening system. Given the lack of any kind of activity assay for the Dead end protein, its solubility was assessed by SDS-PAGE analysis instead, in order to identify candidate conditions. Upon detection of stable target protein in soluble form, these conditions can then be investigated in detail using spectroscopic methods such as CD or NMR spectroscopy. An initial reduction to several candidates is essential however, as otherwise these methods are not suitable for high-throughput screening.

Although CD and NMR spectroscopy can not replace an activity assay, the structural information gained in conjunction with knowledge of structural features in the target protein enables one to assess the likelihood that correct refolding took place.

11.3.4 ZDnd RRM refolding

The first construct to be tested in the refolding screen was the minimal zebrafish construct spanning amino acids 57 to 137 and containing the putative RRM, as

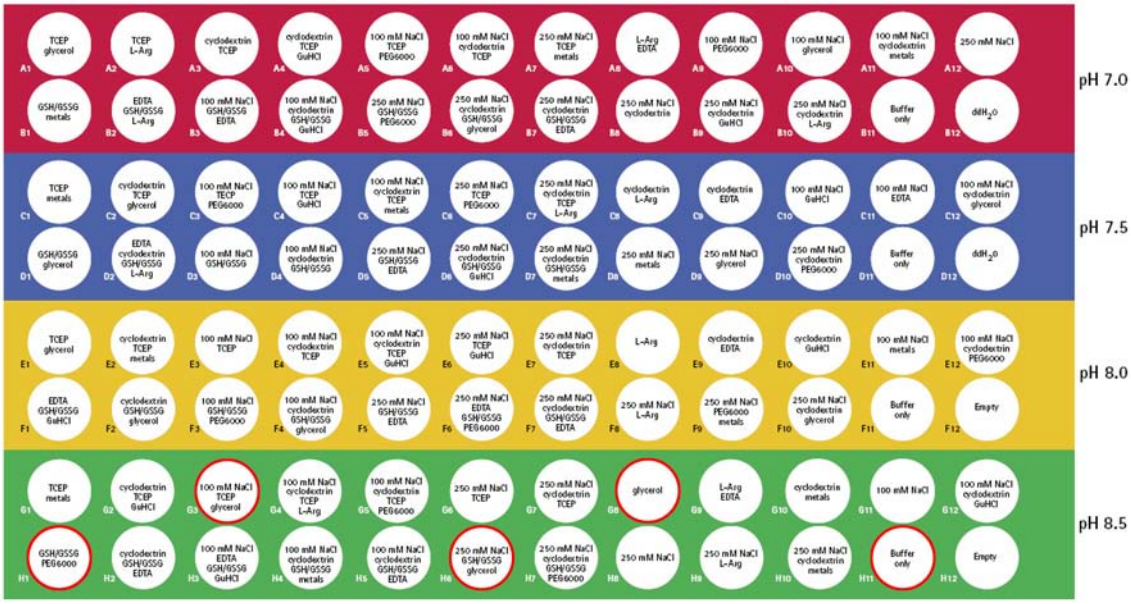

Figure 11.14: Refolding conditions in the iFold™ Refolding System 1. Conditions yielding soluble protein as assessed by SDS-PAGE are indicated as red circles.

single domains are thought to be more likely to be capable of autonomous refolding than large multidomain proteins. Only few conditions showed comparatively little protein precipitate as assessed both visually and by readings at A_{340} (c.f. Table C.4). This observation was confirmed after SDS-PAGE analysis of the soluble fraction of all buffer conditions in the screen. Fig. 11.14 shows the positive hits in the context of all conditions contained in the refolding matrix along with their corresponding with SDS-PAGE analyses.

The five identified conditions were scaled up to provide material for structural characterisation of the Dead end RRM. This revealed higher molecular weight impurities that persisted after refolding, clearly visible by SDS-PAGE at concentrations of 1 mg/mL or greater. Subsequent purification trials however failed as the Dead end construct consistently adhered to the column resin, indepent of if the purification was carried out prior to or after refolding.

Structural characterisation was further impeded by many of the refolding buffer components. Most of the components (Tris, glycerol, salt, GSH/GSSG, PEG) are highly undesirable for CD spectroscopy as they show strong absorptions at around

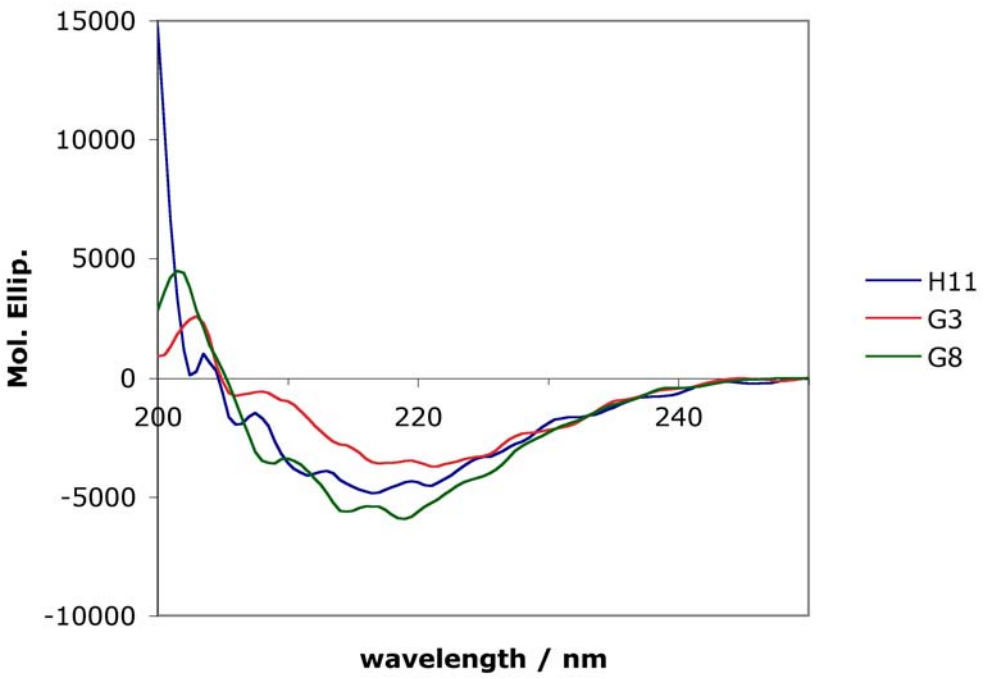

Figure 11.15: CD spectra of refolded ZDnd RRM fragments. H11 – pH 8.5; G3 - pH 8.5, 100 mM NaCl, TCEP 1 mM, glycerol 20% (v/v); G8 – pH 8.5, glycerol 20% (v/v).

200 nm, one of the most important regions of CD spectra. All attempts to bring the refolded protein into other buffer systems failed as the protein strongly aggregated in the process. Nevertheless, it was possible to record spectra of three of the conditions as shown in Fig. 11.15. The spectra show signs both of secondary structure elements, as well as substantial random coil components preventing firm conclusions about the structure under these conditions.

Furthermore, it was noted that the refolded protein showed high aggregation tendencies that increased with increasing concentration of the protein.

11.3.5 ZDnd full length refolding

The refolding screen of the full length zebrafish isoform showed a very suprising result. As indicated by the readings at A_{340} (c.f. Table C.5), the majority of the conditions led to soluble protein. This striking observation was reinforced by SDS-PAGE analysis, revealing only two conditions (C4 and F1, c.f. Fig. 11.14)

in which no target protein could be detected in the soluble fraction after refolding (Fig. C.1). All remaining conditions showed soluble target protein. As with the minimal RRM construct however, this was accompanied by a large molecular weight range of contaminats that proved impossible to remove. Purification attempts by affinity and ion exchange chromatography failed, as also the full length protein precipitated on the column. Furthermore, the full length protein displayed very severe aggregation tendencies after refolding, such that stable working conditions could not be established.

11.3.6 HDnd tandem RRM refolding

The construct spanning amino acids 57 to 217 of the human isoform, and thereby encompassing both putative RRM modules, was tested in a third refolding screen. The result of both the readings at A_{340} (c.f. Table C.6) and the SDS-PAGE analysis of all refolding conditions showed a very similar picture to that of the full length zebrafish protein. Again, soluble protein could be detected in the majority of conditions after refolding, but the freshly refolded protein proved resilient to further purification or buffer exchange, and aggregated again over the course of several hours.

11.4 Interaction partners

Recombinant proteins expressed in *E. coli* are often found to be insoluble, as they lack their natural interaction partners in the heterologous environment, thus exposing hydrophobic interaction sites which in turn lead to aggregation. A common remedy in this situation is co-expression with the missing interaction partner, which allows one to purify a soluble complex.

In the case of Dead end however, no direct protein interaction partners are known to date. As a yeast-two-hybrid screen was unable to identify interaction candidates (Erez Raz, personal communication), a large scale immuno-precipitation of Dead end expressed in human MCF9 cells was conducted in the laboratory of Reuven Agami (NKI, Amsterdam). The immuno-precipitation sampes were analysed by

mass spectrometry[7] in conjunction with the group of Henning Urlaub. In contrast to the initial yeast-two-hybrid screen, a list of approximately 100 putative interaction candidates were identified which is currently being validated for direct interactions in the laboratories of Erez Raz and Reuven Agami. Independently, APOBEC3 has been reported to interact with Dead end, however it remains to be determined if this interaction is of direct nature.

[7]for principles of mass spectrometry, see Section 6.2.1 on page 70.

Chapter 12

Discussion

12.1 Discussion

'We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover up all the tracks, to not worry about the blind alleys or describe how you had the wrong idea first, and so on. So there isn't any place to publish, in a dignified manner, what you actually did in order to get to do the work.'

—Richard Feynman, Nobel Lecture, 1966

12.1.1 Protein expression & purification

Common to many proteomics endeavors, protein expression and purification has also constituted the major bottle-neck in the advancement of structural studies of Dead end. Whilst a large number of parameters have been investigated, no strategy for obtaining a stable soluble sample could be identified. Owing to the large variety of parameters involved, the clear need for a systematic and efficient screening method has become apparent. Thus the combination of recently developed auto-induction media together with lysis buffer screens has been established. This approach, although ultimately unsuccessful, has proven much superior to the classical approach of time-consuming individual small scale trials.

Although a large number of constructs and conditions have been surveyed, it can not be excluded that there exists a combination of factors that allow successful

sample preparation of the Dead end protein. Further tests could investigate the behaviour of a fourth and fifth, yet untested isoforms, namely mouse and chick[1]. With respect to the constructs that have been investigated, it can also not be excluded that a similar construct, differing by as little as one amino acid, will behave differently and yield the desired results [220]. In this respect the search for successful expression constructs and conditions, and nondeterministic polynomial time (NP) complete computational problems are very much alike, once the rational basis for construct design has been exhausted.

The gravity of this time-consuming bottle neck has been widely recognized in the fields of structural genomics and proteomics, leading to increasing development of high throughput methods to overcome this obstacle [221, 222, 223]. Whilst initially only of interest to structural genomics, many of these methods are beginning to find their ways into classical laboratories, as to take full advantage of a variety of strategies developed to improve the expression of soluble protein in *E. coli*, an easy, rapid means to test many growth parameters is necessary. Utilizing increasing automatisation it is thus possible to test a large range of conditions such as growth temperature, growth media, host strain and affinity tag type and placement.

Whilst the aforementioned techniques cover the majority of factors affecting expression, they do not address the problem of genetic construct design. Using PCR based methods, it is possible to generate small libraries of truncation and point mutation libraries by hand, as demonstrated in this work. If larger libraries are required however, this quickly become unfeasible. To overcome this limitation, several methods have been developed in recent years to overcome this obstacle. This includes directed molecular evolution and combinatorial methodologies to generate large gene libraries through random point mutations [224] or gene shuffling [225]. A yet more powerful method has recently been proposed by Darren Hart and coworkers, called 'expression of soluble proteins by random incremental truncation (ESPRIT)' [226, 227]. This combinatorial library approach allows the generation and high throughput screening of very large truncation libraries covering the entire protein sequence.

[1]To be precise, in mouse two isoforms have been identified, sequenced and are available es EST clones

Returning to the results at hand, expression as a MBP fusion protein did result in limited solubility, rather than out right insolubility. Whilst proteolytic susceptibility, another common problem, was a non-issue, the limited solubility and aggregation proneness rendered this approach unsuitable with regard to structural investigation by X-ray crystallography or NMR, due to the minimum requirements of protein concentration and stability. Nevertheless, the developed protocol for the purification of limited solubility constructs may prove useful in the generation of reliable antibodies to the Dead end protein.

12.1.2 Protein refolding

As protein production using classical techniques had proven unsuccessful, a refolding strategy was adopted instead. Finding optimal refolding conditions is also considered to exhibit NP-completeness due to the multitude of factors involved, and therefore requires an efficient high throughput approach [218, 219]. This has become feasible with the development of commercial refolding systems that have become available during the time of this work.

During the first screen focusing on the minimal ZDnd RRM construct, five successful conditions were identified. Common to all conditions is the elevated pH, although it is difficult to draw conclusions about the nature of the protein from this fact. Furthermore, five out of six conditions contain high amounts of stabilizers (glycerol, PEG) indicating a limited stability of the protein alone. Although two conditions were identified that contain a disulphide bridge reshuffling system, two others do not whilst the final condition contains a reducing agent preventing disulphide bridge formation, which indicates that disulphide bridges are unlikely to play a role in refolding of the RRM. Owing to the limited stability of the soluble form and its aggregation susceptibility, it proved impossible to purify the construct further or exchange the refolding buffer hampering a conclusive structural characterisation by NMR, CD spectroscopy or X-ray crystallography. Consequently, the nature of the soluble species obtained after refolding has remained unknown, and it is unclear whether it represents the native structure of the RRM, or whether it

is a refolding artefact. This underlines the need for an activity assay, which could provide strong indications whether the native structure is present or not.

The subsequent refolding screens of the longer Dead end constructs revealed a very surprising result. Although the behaviour of the refolded proteins was very similar to the putative RRM, in contrast to it almost all conditions indicated soluble species after refolding. Given the large variability throughout the tested conditions, it appears rather unlikely that a multidomain protein, as is the case for these constructs, could be capable of sponaneous refolding across the whole screen. Although correct refolding can not be ruled out, it indicates that at least a number of the conditions tested yielded incorrectly refolded protein, refolding intermediates or partially soluble oligomers. Whilst it was possible to assess the amount of soluble protein in solution, both through UV spectroscopy as well as SDS-PAGE analysis, it has remained unclear if and which conditions contain correctly refolded protein. Chromatographic and spectroscopic methods have been employed to assess selected conditions. As this does not constitute a feasible screening method, given the large number of potential candidates in the screen, an activity assay amenable to the 96-well format is required.

12.1.3 Hydrophobic interactions

Throughout all of the work conducted on the Dead end protein, one very promiment characteristic has been recurring. All tested isoforms and constructs (all containing the putative RRM), have displayed a very strong tendency to interact with other proteins or compounds. This applies both to the auto-aggregation of the protein itself, but also to the strong association with the MBP fusion partner, coexpressed chaperones or even chromatography resin. As high salt or pH had little effect on this observation, the interaction is presumably of a hydrophobic nature, leading to the hypothesis that a specific cofactor is required for correct folding, or that the Dead end protein requires an interaction partner to be stable. This scenario is backed by numerous reports in the literature where co-expression of an endogenous interaction partner has drastically altered shifted the picture from insoluble inclusion bodies to a stable, soluble complex.

Despite repeated efforts, no direct interaction partners of Dead end has been identified today. The large scale mass spectrometric analysis of the cell culture pull-down experiment has provided a number of candidates which are currently being verified *in vivo* through functional genomics, colocalization experiments, mutational studies and co-immunoprecipication experiments. The most promising candidate so far is APOBEC3 which has been identified by co-immunoprecipiation, however it remains to be shown whether this interaction is of direct nature.

Finally, the role of mRNA as an interaction partner should be investigated further. Although, to date the literature has not reported cases where the solubility of a complex could be increased through the presence of RNA, it does constitute a novel valuable option to pursue.

Independet of a protein or nucleic acid substrate to Dead end, the constructs and protocols established in this work provide a solid basis upon which newly identified interaction partners may be easily tested. Furthermore, the partially soluble MBP fusion constructs may provide the required synergy to achieve this, through their value in generating antibodies to the Dead end protein.

Part V

Summary

Chapter 13

Summary and conclusions

The work described within presents an investigation of protein function using a selection of state of the art methods. A wide range of protein expression and purification strategies are discussed in the context of the Dead end protein. This first step towards structure determination or biochemical characterisation presents a common obstacle in many structural genomics studies. Although this obstacle was met by an in-depth, systematic approach soluble and stable sample amenable to structural characterisation could be obtained. However, the work has laid the foundation of future biophysical and biochemical investigations of this disease relevant protein, and has established a protocol for the preparation of dilute recombinant protein solutions that may prove valuable in the generation of antibodies to the Dead end protein. Putative interaction partners of Dead end could be identified which may provide a promising approach to obtaining soluble protein as soon as direct interaction can be confirmed for one of the candidates.

Mass spectrometry has been recognised as a powerful tool in the identification of putative interaction partners of the Dead end protein. Existing analysis schemes for the analysis of protein interaction studies using cross-linking reagents, as well as analysis schemes to identify post-translational modifications through ubiquitin-like modifiers have been extended and integrated into the ChopTools web service. This tool creates peptide fragment libraries that can easily be integrated in existing analysis schemes in a high-throughput compatible manner, allowing efficient identification of large peptide based modifications. The software is available as a browser

based web interface, as well as a machine-to-machine compatible web service and a stand alone command line version licensed under GPLv3.

NMR spectroscopy has been employed to investigate the structural model of a calmodulin-peptide complex. Following the establishment of sample preparation protocols for isotope labelled and paramagnetically aligned NMR samples, triple resonance experiments were employed to assign resonances in the diamagnetic and paramagnetic samples of the complex. Through the measurement of RDCs, which provide orientational and motionally averaged probes of domain motions, the conformational model proposed by an X-ray structure could be extended. It was shown that the confomers in the X-ray structure alone, do not adequately describe the state of the complex in solution. Instead, both conformations are likely to be a subset of the conformational space sampled by calmodulin. The work etablishes the basis for a future N-state analysis of the motion, which will be able to define this conformational space in a geometric manner.

Part VI

Appendix

Appendix A

Calmodulin Dynamics (Supplement)

A.1 Primary sequences

Table A.1: Amino acid sequence of *Xenopus* calmodulin (wild-type).

```
        10         20         30         40         50
ADQLTEEQIA EFKEAFSLFD KDGDGTITTK ELGTVMRSLG QNPTEAELQD

        60         70         80         90        100
MINEVDADGN GTIDFPEFLT MMARKMKDTD SEEEIREAFR VFDKDGNGYI

       110        120        130        140        150
SAAELRHVMT NLGEKLTDEE VDEMIREADI DGDGQVNYEE FVQMMTAK
```

Table A.2: Amino acid sequence of *Xenopus* calmodulin N60D mutant.

```
        10         20         30         40         50
ADQLTEEQIA EFKEAFSLFD KDGDGTITTK ELGTVMRSLG QNPTEAELQD

        60         70         80         90        100
MINEVDADGD GTIDFPEFLT MMARKMKDTD SEEEIREAFR VFDKDGNGYI

       110        120        130        140        150
SAAELRHVMT NLGEKLTDEE VDEMIREADI DGDGQVNYEE FVQMMTAK
```

Table A.3: Amino acid sequence of *Xenopus* calmodulin T146C mutant.

```
        10         20         30         40         50
ADQLTEEQIA EFKEAFSLFD KDGDGTITTK ELGTVMRSLG QNPTEAELQD

        60         70         80         90        100
MINEVDADGN GTIDFPEFLT MMARKMKDTD SEEEIREAFR VFDKDGNGYI

       110        120        130        140        150
SAAELRHVMT NLGEKLTDEE VDEMIREADI DGDGQVNYEE FVQMMCAK
```

Appendix B

ChopTools (Supplement)

B.1 Protease rules

Table B.1 lists the details of *in silico* cleavage rules implemented in Chop'N'Spice and Chop'N'Stitch. In this table, "X" stands for any amino acid, "—" denotes the cleavage site. Most cleavage rules were adapted from [228].

Table B.1: Protease cleavage rules

Protease	Rule
Arg-C proteinase	R | X
Asp-N endopeptidase	X | D
Asp-N endopeptidase (+N-terminal Glu)	X | D X | E
Chymotrypsin high specificity	F | not P Y | not P W | not M or P
Chymotrypsin low specificity	F | not P Y | not P L | not P W | not M or P M | not P or Y H | not D, M, P or W
CNBr - high excess	M | X
CNBr - low excess	M | not S or T
Enterokinase	D or N, D or N, D or N | K
Factor Xa	any of AFGILTVM, D or E, G, R | X

Formic acid	D | X
Glu-C endopeptidase	E | X
Glu-C endopeptidase (N-term D or E)	D | X E | X
Granzyme B	I, E, P, D | X
Hydroxylamine	N | G
Iodosobenzoic acid	W | X
Lys-C proteinase	K | not P
Lys-C proteinase + C-term P	K | X
NTCB	X | C
Pepsin - pH 1.3	(not H, K, R), (not P), (not R) | (F or L or W or Y), (not P) (not H, K, R), (not P), (F or L or W or Y) | X, (not P)
Pepsin - pH > 2	(not H, K, R), (not P), (not R) | (F or L), (not P) (not H, K, R), (not P), (F or L) | X, (not P)
Proteinase K	any of AEFILTVWY | X
TEV protease	E, N, L, Y, F | G
Thermolysin	(not D,E) | (any of AFILMV)
Thrombin	G, R | G (any of AFGILTVM), (any of AFGILTVWA), P, R | (not D,E), (not DE)
Trypsin N-term K or R	K | X R | X
Trypsin N-term K or R, not C-term P	K | not P R | not P
Trypsin advanced model with the exceptions: (no cleavage!)	K or R | not P W, K | P M, R | P C or D, K | D C, K | H or Y C, R | K R, R | H or R
Trypsin + Chymotrypsin/CNBr	Cleavage rules for both enzymes apply

B.2 Standard modifiers

Table B.2 lists Ubiquitin and Ubiquitin like modifier sequences that are implemented as "standard spices" within the Chop'N'Spice module. In a local installation of Chop'N'Spice these spices may be changed by editing the main configuration file.

Table B.2: Standard spices

Species	Modifier	Sequence
A. thaliana	Rub1	MQIFVKTLTG KTITLEVESS DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGGTMIK VKTLTGKEIE IDIEPTDTID RIKERVEEKE GIPPVQQRLI YAGKQLADDK TAKDYNIEGG SVLHLVLALR GG
	SUMO1	MSANQEEDKK PGDGGAHINL KVKGQDGNEV FFRIKRSTQL KKLMNAYCDR QSVDMNSIAF LFDGRRLRAE QTPDELDMED GDEIDAMLHQ TGG
	SUMO2	MSATPEEDKK PDQGAHINLK VKGQDGNEVF FRIKRSTQLK KLMNAYCDRQ SVDFNSIAFL FDGRRLRAEQ TPDELEMEDG DEIDAMLHQT GG
	SUMO3	MSNPQDDKPI DQEQEAHVIL KVKSQDGDEV LFKNKKSAPL KKLMYVYCDR RGLKLDAFAF IFNGARIGGL ETPDELDMED GDVIDACRAM SGG
	SUMO5	MVSSTDTISA SFVSKKSRSP ETSPHMKVTL KVKNQQGAED LYKIGTHAHL KKLMSAYCTK RNLDYSSVRF VYNGREIKAR QTPAQLHMEE EDEICMVMEL GG
	Ubq	MQIFVKTLTG KTITLEVESS DTIDNVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLADYN IQKESTLHLV LRLRGG
H. sapiens	Fat10	MAPNASCLCV HVRSEEWDLM TFDANPYDSV KKIKEHVRSK TKVPVQDQVL LLGSKILKPR RSLSSYGIDK EKTIHLTLKV VKPSDEELPL FLVESGDEAK RHLLQVRRSS SVAQVKAMIE TKTGIIPETQ IVTCNGKRLE DGKMMADYGI RKGNLLFLAS YCIGG
	FUBI	MQLFVRAQEL HTFEVTGQET VAQIKAHVAS LEGIAPEDQV VLLAGAPLED EATLGQCGVE ALTTLEVAGR MLGG
	ISG15	MGWDLTVKML AGNEFQVSLS SSMSVSELKA QITQKIGVHA FQQRLAVHPS GVALQDRVPL ASQGLGPGST VLLVVDKCDE

		PLSILVRNNK GRSSTYEVRL TQTVAHLKQQ VSGLEGVQDD LFWLTFEGKP LEDQLPLGEY GLKPLSTVFM NLRLRGG
	Nedd8	MLIKVKTLTG KEIEIDIEPT DKVERIKERV EEKEGIPPQQ QRLIYSGKQM NDEKTAADYK ILGGSVLHLV LALRGG
	SUMO1	MSDQEAKPST EDLGDKKEGE YIKLKVIGQD SSEIHFKVKM TTHLKKLKES YCQRQGVPMN SLRFLFEGQR IADNHTPKEL GMEEEDVIEV YQEQTGG
	SUMO2	MADEKPKEGV KTENNDHINL KVAGQDGSVV QFKIKRHTPL SKLMKAYCER QGLSMRQIRF RFDGQPINET DTPAQLEMED EDTIDVFQQQ TGG
	SUMO3	MSEEKPKEGV KTENDHINLK VAGQDGSVVQ FKIKRHTPLS KLMKAYCERQ GLSMRQIRFR FDGQPINETD TPAQLEMEDE DTIDVFQQQT GG
	Ubq	MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
	URM1	MAAPLSVEVE FGGGAELLFD GIKKHRVTLP GQEEPWDIRN LLIWIKKNLL KERPELFIQG DSVRPGILVL INDADWELLG ELDYQLQDQD SVLFISTLHG G
M. musculus	Fat10	MASVRTCVVR SDQWRLMTFE TTENDKVKKI NEHIRSQTKV SVQDQILLLD SKILKPHRKL SSYGIDKETT IHLTLKVVKP SDEELPLFLV ESKNEGQRHL LRVRRSSSVA QVKEMIESVT SVIPKKQVVN CNGKKLEDGK IMADYNIKSG SLLFLTTHCT GG
	FUBI	SVANMQLFVR AQELHTLEVT GQETVAQIKD HVASLEGIAP EDQVVLLAGS PLEDEATLGQ CGVEALTTLE VAGRMLGG
	ISG15	MAWDLKVKML GGNDFLVSVT NSMTVSELKK QIAQKIGVPA FQQRLAHQTA VLQDGLTLSS LGLGPSSTVM LVVQNCSEPL SILVRNERGH SNIYEVFLTQ TVDTLKKKVS SGTSHEDQFW LSFEGRPMED KELLGEYGLK PQCTVIKHLR LRGG
	Nedd8	MLIKVKTLTG KEIEIDIEPT DKVERIKERV EEKEGIPPQQ QRLIYSGKQM NDEKTAADYK ILGGSVLHLV LALRGG
	SUMO1	MSDQEAKPST EDLGDKKEGE YIKLKVIGQD SSEIHFKVKM TTHLKKLKES YCQRQGVPMN SLRFLFEGQR IADNHTPKEL GMEEEDVIEV YQEQTGG
	SUMO2	MADEKPKEGV KTENNDHINL KVAGQDGSVV QFKIKRHTPL SKLMKAYCER QGLSMRQIRF RFDGQPINET DTPAQLEMED EDTIDVFQQQ TGG
	SUMO3	MSEEKPKEGV KTENDHINLK VAGQDGSVVQ FKIKRHTPLS KLMKAYCERQ GLSMRQIRFR FDGQPINETD TPAQLEMEDE

		DTIDVFQQQT GG
	Ubq	MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
	URM1	MAAPLCVKVE FGGGAELLFD GVKKHQVALP GQEEPWDIRN LLVWIKKNLL KERPELFIQG DSVRPGILVL INDADWELLG ELDYQLQDQD SILFISTLHG G
R. norvegicus	Fat10	MASCVCVVRS EQWPLMTFDT TMSDKVKKIN EHIRSQTKVS VQDQILLLDS KILKPHRALS SYGIDKENTI HLTLKVVKPS DEELPLSLVE SGDEGQRHLL RVRRSSSVAQ VKEMIENVTA VPPKKQIVNC NGKRLEDGKI MADYNIKSGS LLFLTAHCIG G
	FUBI	MQLFVRAQEL HTLEVTGQET VAQIKAHVAS LEGIAPEDQV VLLAGSPLED EATLGQCGVE ALTTLEVAGR MLGG
	ISG15	MTWNLKVKML GGKEFLVSMT NSMMLSELKK QVAQKSGVPA FQQRLAHQSG EMLQDGVALI RQGLSSGSTV MLMVENCSHP LSILVRNERG RSNVYEVQLT QTVEVLMRQV SQHEQVSQDQ FWLSFNGRPM EDKEPLGEYG LTPHCTVIMN LRLRGG
	Nedd8	MLIKVKTLTG KEIEIDIEPT DKVERIKERV EEKEGIPPQQ QRLIYSGKQM NDEKTAADYK ILGGSVLHLV LALRGG
	SUMO1	MSDQEAKPST EDLGDKKEGE YIKLKVIGQD SSEIHFKVKM TTHLKKLKES YCQRQGVPMN SLRFLFEGQR IADNHTPKEL GMEEEDVIEV YQEQTGG
	SUMO2	MADEKPKEGV KTENNDHINL KVAGQDGSVV QFKIKRHTPL SKLMKAYCER QGLSMRQIRF RFDGQPINET DTPAQLEMED EDTIDVFQQQ TGG
	SUMO3	MSEEKPKEGV KTENDHINLK VAGQDGSVVQ FKIKRHTPLS KLMKAYCERQ GLSMRQIRFR FDGQPINETD TPAQLEMEDE DTIDVFQQQT GG
	Ubq	MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
	URM1	MAAPLCVEVE FGGGAELLFD GVKKHQVTLP GQEEPWDIRN LLVWIKTNLL KERPELFIQG DSVRPGILVL INDADWELLG ELDYQLQDQD SILFISTLHG G
S. cerevisiae	Rub1	MIVKVKTLTG KEISVELKES DLVYHIKELL EEKEGIPPSQ QRLIFQGKQI DDKLTVTDAH LVEGMQLHLV LTLRGG
	SMT3	MSDSEVNQEA KPEVKPEVKP ETHINLKVSD GSSEIFFKIK KTTPLRRLME AFAKRQGKEM DSLRFLYDGI RIQADQTPED LDMEDNDIIE AHREQIGG
	Ubq	MQIFVKTLTG KTITLEVESS DTIDNVKSKI QDKEGIPPDQ

		QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG
	URM1	MVNVKVEFLG GLDAIFGKQR VHKIKMDKED PVTVGDLIDH
		IVSTMINNPN DVSIFIEDDS IRPGIITLIN DTDWELEGEK
		DYILEDGDII SFTSTLHGG
S. pombe	Nedd8	MLIKVKTLTG KEIELDIDPN DKVSRIKERV EEKEGIPPSQ
		QRLIYAGKQM ADDKNAESYH LEGGSVLHLV LALRGG
	SUMO	MSESPSANIS DADKSAITPT TGDTSQQDVK PSTEHINLKV
		VGQDNNEVFF KIKKTTEFSK LMKIYCARQG KSMNSLRFLV
		DGERIRPDQT PAELDMEDGD QIEAVLEQLG G
	Ubq	MQIFVKTLTG KTITLEVESS DTIDNVKSKI QDKEGIPPDQ
		QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG

B.3 Molecular weights

Table B.3 lists the molecular weights used in the m/z calculation modes "monoisotopic" and "average" of both Chop'N'Spice and Chop'N'Stitch. These values may be changed by editing the main configuration files of both modules.

Table B.3: Monoisotopic and average molecular weights

Entity	Monoistopic MW	Average MW
A	71.037114	71.0779
C	103.009185	103.1429
D	115.026943	115.0874
E	129.042593	129.1140
F	147.068414	147.1739
G	57.021464	57.0513
H	137.058912	137.1393
I	113.084064	113.1576
J	0.0000000	0.0000
K	128.094963	128.1723
L	113.084064	113.1576
M	131.040485	131.1961
N	114.042927	114.1026
P	97.052764	97.1152
Q	128.058578	128.1292
R	156.101111	156.1857
S	87.032028	87.0773
T	101.047679	101.1039
V	99.068414	99.1311
W	186.079313	186.2099
Y	163.063329	163.1733
HYDROGEN	1.007825035	1.00794
CARBON	12.0	12.0107
NITROGEN	14.003074	14.0067
OXYGEN	15.99491463	15.9994
ELECTRON	0.000549	0.000549
PROTON	1.007276035	1.007391
H_2O	18.0105647	18.015280

B.4 Technical details

B.4.1 Specifications

ChopTools is written in PHP 5.2. The software has no external dependencies, but requires the following extensions that are distributed together with the PHP core, and which may be dynamically linked:

- BCMath – arbitrary precision mathematics
- iconv – character set conversion facility
- XML Parser – toolkit to build XML parsers

ChopTools

B.4.2 Architecture

ChopTools follows a loose interpretation of the Model-View-Controller design pattern

B.4.3 Data & resources

Protease cleavage rules were adopted from [228]. A complete list of all proteases and cleavage rules is given in Table B.1 in the Appendix. A list of all ubiquitin-like modifiers (Ubl) and their corresponding sequences that are implemented in ChopNSpice is given in Table B.2. Molecular weights used in the calculation of monoisotopic or average peptide molecular weights are listed in Table B.3 in the Appendix. Dynamic Doctype Definitions (DTD) of XML input files for the web service are available through the ChopTools website.

B.4.4 License and availability

The developed software tools presented in this thesis are freely available online (`http://choptools.gwdg.de`) as a web server. The source code is released as open source under the terms of the General Public License v3 (GPLv3) and may be obtained at the same website.

PHP is distributed under the PHP license. The software, license and documentation are freely available through `http://www.php.net`.

Appendix C

Structural Basis of Dead end Function (Supplement)

C.1 Primary sequences

C.1.1 Zebrafish

Genebank accession number AAP22373, 411 aa, MW: 46020.1 Da, predicted pI: 8.96

Table C.1: Amino acid sequence of *Danio rerio* Dead end (wild-type).

```
        10         20         30         40         50
MVGDMDAQQQ ELQQILNPQK LKSLQEWMQR NSITLTQVNG QRKYGGPPPG

        60         70         80         90        100
WQGPAPGSGC EVFISQIPND VYEDRLIPLF QSIGTIYEFR LMMNFSGQTR

       110        120        130        140        150
GFAYAKYGDP LTASAAVTTL HQYRLPEGGC LTVRRSTEKR QLRLGDLPVS

       160        170        180        190        200
MNESKLLMVL QMLSDGVEDV LLKPPGPKGK EVVALVNYTS HYAASMAKKV

       210        220        230        240        250
LVEAFRNRYG ISITVRWTSF SKSKRVEDTP QEDSCVTPLV LKPLSKPSLL

       260        270        280        290        300
HYDVPAHQSL LPLFRAVGGP TTSEQRDEMI PQPTIMSRNE LIPQSSIRQR

       310        320        330        340        350
DEMVPQLPIR PRDGMAPQSP ISLDAVSHLQ WMCEVNRLGS PQYEVHFHHA

       360        370        380        390        400
APDGFLYFAF KVLIPGLPLP LYGFVQILPG TSARAMKSEV YRAAAEQVIQ

       410        420        430        440        450
TLCRVSNLRP F
```

C.1.2 Human

NCBI accession number NP_919225, 353 aa, MW: 38686.9 Da, predicted pI: 9.74

Table C.2: Amino acid sequence of *Homo sapiens* Dead end (wild-type).

```
        10         20         30         40         50
MQSKRDCELW CERVNPENKA ALEAWVRETG IRLVQVNGQR KYGGPPPGWV

        60         70         80         90        100
GSPPPAGSEV FIGRLPQDVY EHQLIPLFQR VGRLYEFRLM MTFSGLNRGF

       110        120        130        140        150
AYARYSSRRG AQAAIATLHN HPLRPSCPLL VCRSTEKCEL SVDGLPPNLT

       160        170        180        190        200
RSALLLALQP LGPGLQEARL LPSPGPAPGQ IALLKFSSHR AAAMAKKALV

       210        220        230        240        250
EGQSHLCGEQ VAVEWLKPDL KQRLRQQLVG PFLRSPQPEG SQLALARDKL

       260        270        280        290        300
GFQGARATLQ LLCQRMKLGS PVFLTKCLGI GPAGWHRFWY QVVIPGHPVP

       310        320        330        340        350
FSGLIWVVLT LDGRDGHEVA KDAVSVRLLQ ALSESGANLL WSAGAEAGTM

       360
VKQ
```

C.1.3 African clawed frog

Genebank accession number AAX84947, 354 aa, MW: 39602.0 Da, predicted pI: 8.06

Table C.3: Amino acid sequence of *Xenopus laevis* Dead end (wild type).

```
        10         20         30         40         50
MELSDEQKSQ MWINSVNSEN KEALLAWVKE TGIELVQING QRKYGGPPPG

        60         70         80         90        100
WIGNAPVSGS EVFIGKIPQD IYEDKLIPLF QSVGKLYEFR LMMTFSGLNR

       110        120        130        140        150
GFAYARYISR RQAISAIMSL NGFEITKGCC IVVCRSTEKS ELALDGLPGN

       160        170        180        190        200
FDENMLKNVL DEVTSGVSSI SLHPSPTKES QVLAVVKYDS HRAAAMAKKT

       210        220        230        240        250
LCEGSPILPG LPLTVNWLKT DMRQKLRSTD KLQQTKDLSP LPLLYTDRPD

       260        270        280        290        300
LPKETLLSAV GCLNMLCQEM KLGRPVFLIK LFSVTSFGWI RFWYQVVIPT

       310        320        330        340        350
YPTPFCGYAW MIGENLELNE KYEHARQVVA MKILSALGYI PDFSLGDVTA

       360
RNAL
```

C.2 A_{340} measurements (refolding screen)

Table C.4: Absorbance at 340 nm in the iFold 1 screen of ZDnd 57-137

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.080	0.110	0.090	0.099	0.098	0.113	0.127	0.115	0.105	0.112	0.079	0.114
B	0.159	0.155	0.099	0.084	0.163	0.088	0.078	0.103	0.109	0.117	0.086	0.086
C	0.098	0.099	0.116	0.193	0.161	0.123	0.126	0.111	0.091	0.145	0.090	0.091
D	0.086	0.114	0.142	0.108	0.109	0.153	0.080	0.106	0.100	0.115	0.102	0.074
E	0.075	0.158	0.104	0.132	0.105	0.251	0.100	0.112	0.092	0.169	0.090	0.076
F	0.762	0.088	0.102	0.085	0.094	0.098	0.306	0.093	0.138	0.101	0.086	0.813
G	0.067	0.127	0.079	0.135	0.110	0.103	0.124	0.073	0.114	0.136	0.090	0.080
H	0.067	0.102	0.359	0.144	0.092	0.081	0.084	0.081	0.107	0.134	0.073	0.078

Table C.5: Absorbance at 340 nm in the iFold 1 screen of full-length ZDnd

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.060	0.036	0.165	0.089	0.037	0.083	0.037	0.039	0.039	0.066	0.138	0.035
B	0.122	0.037	0.035	0.085	0.293	0.076	0.118	0.094	0.111	0.073	0.037	0.035
C	0.046	0.072	0.038	0.408	0.122	0.036	0.076	0.075	0.113	0.369	0.036	0.070
D	0.050	0.059	0.034	0.060	0.037	0.148	0.087	0.037	0.052	0.089	0.036	0.036
E	0.042	0.138	0.032	0.097	0.116	0.116	0.107	0.039	0.120	0.091	0.038	0.089
F	0.936	0.065	0.038	0.056	0.037	0.038	0.081	0.040	0.036	0.078	0.037	0.097
G	0.038	0.145	0.044	0.074	0.117	0.042	0.103	0.044	0.048	0.083	0.041	0.070
H	0.038	0.069	0.054	0.107	0.066	0.045	0.111	0.046	0.048	0.117	0.040	0.026

Table C.6: Absorbance at 340 nm in the iFold 1 screen of HDnd 57-217

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.037	0.039	0.183	0.216	0.034	0.227	0.038	0.042	0.040	0.035	0.200	0.035
B	0.228	0.040	0.036	0.247	0.540	0.138	0.171	0.143	0.240	0.087	0.037	0.034
C	0.035	0.098	0.041	0.293	0.137	0.036	0.085	0.095	0.076	0.459	0.037	0.116
D	0.035	0.098	0.035	0.169	0.036	0.175	0.125	0.038	0.040	0.162	0.034	0.035
E	0.051	0.202	0.042	0.140	0.190	0.288	0.164	0.039	0.058	0.276	0.036	0.136
F	0.498	0.049	0.035	0.105	0.036	0.037	0.140	0.040	0.037	0.135	0.038	0.045
G	0.036	0.182	0.040	0.094	0.144	0.038	0.159	0.042	0.044	0.139	0.038	0.287
H	0.037	0.102	0.671	0.145	0.086	0.042	0.181	0.045	0.047	0.157	0.038	0.039

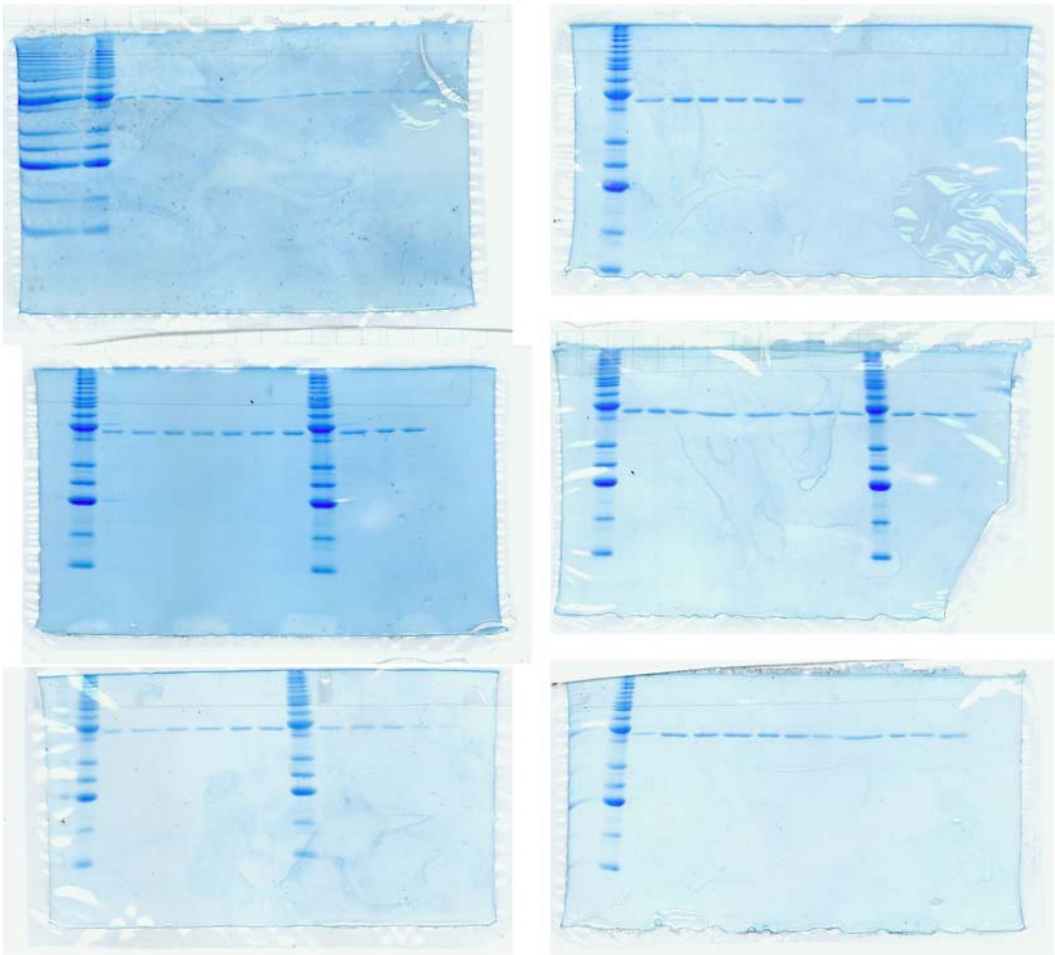

Figure C.1: SDS-PAGE analysis of full-length ZDnd in the iFold^{TM} Refolding System 1. As indicated by the A_{340} screen, all conditions display protein in solution, with the exception of C4 and F1.

List of Symbols and Abbreviations

APS	Ammoniumpersulfate
$\boldsymbol{B}$	Magnetic field (magnetic flux density)
$\boldsymbol{B}_0$	Static external magnetic field
$\boldsymbol{B}_1$	Magnetic field generated by r.f. irradiation
BMRB	Biological Magnetic Resonance Data Bank (`http://www.bmrb.wisc.edu`)
CaM	Calmodulin
CARA	Computer Aided Resonance Assignment (program) `http://www.nmr.ch/`
CD	Circular dichroism
CDF	Calcium dependant facilitation
CDI	Calcium dependant inactivation
D_{IS}	Dipolar coupling constant for two spins I, S
δ	Chemical shift
E	Energy
ΔE	Energy difference
DNA	desoxyribonucleic acid
Dnd	Dead End
E. coli	*Escherichia coli*
Da	Dalton (molecular weight measure for biopolymers)
DTD	Dynamic Doctype Definition
EDTA	Ethylenediaminetetraacetate
EM	electron microscopy
FID	Free induction decay
Fmoc	9-Fluorenylmethoxycarbonyl
γ	Gyromagnetic ratio
$\hbar$	Planck's constant (divided by 2π)
$\hat{H}$	Hamilton operator
HDnd	Dead End, *human* isoform
HSQC	heteronuclear single quantum coherence

$\boldsymbol{I}$	Operator corresponding to (nuclear) spin
I_x, I_y, I_z	Operators corresponding to (nuclear) spin components
INEPT	Insensitive nuclei enhanced by polarization transfer
IPTG	Isopropyl-β-D-thiogalactopyranoside
k_B	Boltzmann constant
L	litre
LB	Luria-Bertani broth, also known as Luria broth or Lysogeny broth
LC	Liquid chromatography
Ln	Lanthanide
M	mol per litre
m/z	mass-to-charge ratio
MCS	Multiple cloning site
MD	Molecular dynamics
mRNA	messenger RNA
miRNA	MicroRNA
MBP	Maltose binding protein
MOL	Molecular axis frame
MS	Mass spectrometry
MWCO	molecular weight cut-off
NMR	Nuclear magnetic resonance
NMRdev	NMRdev computer program `http://galileo.usc.es/~armando/software/nmrdev/`
NMRpipe	NMRpipe computer program `http://spin.niddk.nih.gov/NMRPipe/`
NOESY	Nuclear Overhauser effect spectroscopy
NP	non deterministic polynomial time
ω	Resonance frequency
ω_0	Nuclear Larmor frequency of Zeeman interaction
ω_1	Nuclear Larmor frequency of r.f. interaction
Ω	Resonance frequency offset with respect to carrier frequency
OD_{600}	Optical density at 600 nm
PAGE	Polyacrylamide-Gelelectrophoresis
Pales	Pales computer program
PCS	Pseudo contact shift
PCR	Polymerase chain reaction
PDB	Protein Data Bank, `http://www.pdb.org`
PEG	Poly(ethylene glycol)
PGC	Primordial germ cell

PHP	'PHP: Hypertext processor'
	a programming language
ppm	Parts per million (relative unit of nuclear chemical shift)
PRE	Paramagnetic relaxation enhancement
PTM	Post-translational modification
RDC	Residual dipolar coupling
relax	relax computer program, `http://www.nmr-relax.com`
r.f.	Radio frequency
$\hat{\rho}$	Density operator
RISC	RNA-induced silencing complex
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
RRM	RNA recognition motif
SDS	sodium dodecylsulfate
s.e.m.	standard error of the mean
Sparky	Sparky computer program
	`http://www.cgl.ucsf.edu/home/sparky/`
Sxl	Sex-lethal, a *Drosophila* RNA binding protein
	and developmental sex determinant
t	Time
TCA	Trichloroacetic acid
TEV	Tobacco etch virus
Ubl	Ubiquitin-like modifier
Ubq	Ubiquitin
UTR	untranslated region
Vim	Vim computer program (editor), `http://www.vim.org`
	undeniably superior to Emacs
XCaM	Calmodulin, *Xenopus* isoform
XDnd	Dead End, *Xenopus* isoform
XML	eXtensible Markup Language
YT	Yeast & tryptone, bacterial growth medium
ZDnd	Dead End, *Zebrafish* isoform
ZYM-5052	Complex, auto-inducing bacterial growth medium

Furthermore, references to and abbreviations of amino acids follow the recommendations of IUPAC (International Union of Pure and Applied Chemistry) and IUB (International Union of Biochemistry and Molecular Biology).

Trivial name	**Symbol**	**Abb.**	**Systematic name**
Alanine	Ala	A	2-Aminopropanoic acid
Arginine	Arg	R	2-Amino-5-guanidinopentanoic
Asparagine	Asn	N	2-Amino-3-carbamoylpropanoic acid
Aspartic acid	Asp	D	2-Aminobutanedioic acid
Cysteine	Cys	C	2-Amino-3-mercaptopropanoic acid
Glutamine	Gln	Q	2-Amino-4-carbamoylbutanoic acid
Glutamic acid	Glu	E	2-Aminopentanedioic acid
Glycine	Gly	G	Aminoethanoic acid
Histidine	His	H	2-Amino-3-(1*H*-imidazol-4-yl)-propanoic acid
Isoleucine	Ile	I	2-Amino-3-methylpentanoic acid
Leucine	Leu	L	2-Amino-4-methylpentanoic acid
Lysine	Lys	K	2,6-Diaminohexanoic acid
Methionine	Met	M	2-Amino-4-(methylthio)butanoic acid
Phenylalanine	Phe	F	2-Amino-3-phenylpropanoic acid
Proline	Pro	P	Pyrrolidine-2-carboxylic acid
Serine	Ser	S	2-Amino-3-hydroxypropanoic acid
Threonine	Thr	T	2-Amino-3-hydroxybutanoic acid
Tryptophan	Trp	W	2-Amino-3-(l*H*-indol-3-yl)-propanoic acid
Tyrosine	Tyr	Y	2-Amino-3-(4-hydroxyphenyl)-propanoic acid
Valine	Val	V	2-Amino-3-methylbutanoic acid
Unspecified amino acid	Xaa	X	

Bibliography

[1] J. D. WATSON and F. H. CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[2] J. C. KENDREW, G. BODO, H. M. DINTZIS, R. G. PARRISH, H. WYCKOFF, and D. C. PHILLIPS. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, Mar 1958.

[3] H. MUIRHEAD and M. F. PERUTZ. Structure of haemoglobin. a three-dimensional fourier synthesis of reduced human haemoglobin at 5-5 a resolution. *Nature*, 199:633–638, Aug 1963.

[4] Filip Van Petegem, Franck C Chatelain, and Daniel L Minor. Insights into voltage-gated calcium channel regulation from the structure of the cav1.2 iq domain-ca2+/calmodulin complex. *Nat Struct Mol Biol*, 12(12):1108–1115, Dec 2005.

[5] Gilbert Weidinger, Jürg Stebler, Krasimir Slanchev, Karin Dumstrei, Clare Wise, Robin Lovell-Badge, Christine Thisse, Bernard Thisse, and Erez Raz. dead end, a novel vertebrate germ plasm component, is required for zebrafish primordial germ cell migration and survival. *Curr Biol*, 13(16):1429–1434, Aug 2003.

[6] Kirsten K Youngren, Douglas Coveney, Xiaoning Peng, Chitralekha Bhattacharya, Laura S Schmidt, Michael L Nickerson, Bruce T Lamb, Jian Min Deng, Richard R Behringer, Blanche Capel, Edward M Rubin, Joseph H Nadeau, and Angabin Matin. The ter mutation in the dead end gene causes germ cell loss and testicular germ cell tumours. *Nature*, 435(7040):360–364, May 2005.

[7] J Robison and R.J. Colbran. *Encyclopedia of Biological Chemistry*, volume 1. Elsevier, 2004.

[8] Mitsuhiko Ikura and James B Ames. Genetic polymorphism and protein conformational plasticity in the calmodulin superfamily: two ways to promote multifunctionality. *Proc Natl Acad Sci U S A*, 103(5):1159–1164, Jan 2006.

[9] Y. S. Babu, J. S. Sack, T. J. Greenhough, C. E. Bugg, A. R. Means, and W. J. Cook. Three-dimensional structure of calmodulin. *Nature*, 315(6014):37–40, 1985.

[10] Shibani Bhattacharya, Christopher G Bunick, and Walter J Chazin. Target selectivity in ef-hand calcium binding proteins. *Biochim Biophys Acta*, 1742(1-3):69–79, Dec 2004.

[11] M. Walsh, F. C. Stevens, K. Oikawa, and C. M. Kay. Circular dichroism studies of native and chemically modified ca2+-dependent protein modulator. *Can J Biochem*, 57(3):267–278, Mar 1979.

[12] J. R. Dedman, J. D. Potter, R. L. Jackson, J. D. Johnson, and A. R. Means. Physicochemical properties of rat testis ca2+-dependent regulator protein of cyclic nucleotide phosphodiesterase. relationship of ca2+-binding, conformational changes, and phosphodiesterase activity. *J Biol Chem*, 252(23):8415–8422, Dec 1977.

[13] M. C. Kilhoffer, J. G. Demaille, and D. Gérard. Tyrosine fluorescence of ram testis and octopus calmodulins. effects of calcium, magnesium, and ionic strength. *Biochemistry*, 20(15):4407–4414, Jul 1981.

[14] M. C. Kilhoffer, D. Gerard, and J. G. Demaille. Terbium binding to octopus calmodulin provides the complete sequence of ion binding. *FEBS Lett*, 120(1):99–103, Oct 1980.

[15] Jenny J Yang, Amy Gawthrop, and Yiming Ye. Obtaining site-specific calcium-binding affinities of calmodulin. *Protein Pept Lett*, 10(4):331–345, Aug 2003.

[16] Rhonda A Newman, Wendy S Van Scyoc, Brenda R Sorensen, Olav R Jaren, and Madeline A Shea. Interdomain cooperativity of calmodulin bound to melittin preferentially increases calcium affinity of sites i and ii. *Proteins*, 71(4):1792–1812, Jun 2008.

[17] M. Zhang, T. Tanaka, and M. Ikura. Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. *Nat Struct Biol*, 2(9):758–767, Sep 1995.

[18] D. Chin and A. R. Means. Calmodulin: a prototypical calcium sensor. *Trends Cell Biol*, 10(8):322–328, Aug 2000.

[19] A. Crivici and M. Ikura. Molecular and structural basis of target recognition by calmodulin. *Annu Rev Biophys Biomol Struct*, 24:85–116, 1995.

[20] J. Trewhella, D. K. Blumenthal, S. E. Rokop, and P. A. Seeger. Small-angle scattering studies show distinct conformations of calmodulin in its complexes with two peptides based on the regulatory domain of the catalytic subunit of phosphorylase kinase. *Biochemistry*, 29(40):9316–9324, Oct 1990.

[21] E. W. Small and S. R. Anderson. Fluorescence anisotropy decay demonstrates calcium-dependent shape changes in photo-cross-linked calmodulin. *Biochemistry*, 27(1):419–428, Jan 1988.

[22] E. R. Chapman, K. Alexander, T. Vorherr, E. Carafoli, and D. R. Storm. Fluorescence energy transfer analysis of calmodulin-peptide complexes. *Biochemistry*, 31(51):12819–12825, Dec 1992.

[23] G. Barbato, M. Ikura, L. E. Kay, R. W. Pastor, and A. Bax. Backbone dynamics of calmodulin studied by 15n relaxation using inverse detected two-dimensional nmr spectroscopy: the central helix is flexible. *Biochemistry*, 31(23):5269–5278, Jun 1992.

[24] N. Tjandra, H. Kuboniwa, H. Ren, and A. Bax. Rotational dynamics of calcium-free calmodulin studied by 15n-nmr relaxation measurements. *Eur J Biochem*, 230(3):1014–1024, Jun 1995.

[25] Jennifer L Fallon and Florante A Quiocho. A closed compact structure of native ca(2+)-calmodulin. *Structure*, 11(10):1303–1307, Oct 2003.

[26] Klaus P Hoeflich and Mitsuhiko Ikura. Calmodulin in action: diversity in target recognition and activation mechanisms. *Cell*, 108(6):739–742, Mar 2002.

[27] W. A. Catterall. Structure and regulation of voltage-gated ca2+ channels. *Annu Rev Cell Dev Biol*, 16:521–555, 2000.

[28] Diane Lipscombe, Thomas D Helton, and Weifeng Xu. L-type calcium channels: the low down. *J Neurophysiol*, 92(5):2633–2641, Nov 2004.

[29] Yoshiro Saimi and Ching Kung. Calmodulin as an ion channel subunit. *Annu Rev Physiol*, 64:289–311, 2002.

[30] E. Bourinet, P. Charnet, W. J. Tomlinson, A. Stea, T. P. Snutch, and J. Nargeot. Voltage-dependent facilitation of a neuronal alpha 1c l-type calcium channel. *EMBO J*, 13(21):5032–5039, Nov 1994.

[31] N. F. Lim, M. C. Nowycky, and R. J. Bookman. Direct measurement of exocytosis and calcium currents in single vertebrate nerve terminals. *Nature*, 344(6265):449–451, Mar 1990.

[32] N. Dascal, T. P. Snutch, H. Lübbert, N. Davidson, and H. A. Lester. Expression and modulation of voltage-gated calcium channels after rna injection in xenopus oocytes. *Science*, 231(4742):1147–1150, Mar 1986.

[33] J. P. Imredy and D. T. Yue. Mechanism of ca(2+)-sensitive inactivation of l-type ca2+ channels. *Neuron*, 12(6):1301–1318, Jun 1994.

[34] R. D. Zühlke, G. S. Pitt, K. Deisseroth, R. W. Tsien, and H. Reuter. Calmodulin supports both inactivation and facilitation of l-type calcium channels. *Nature*, 399(6732):159–162, May 1999.

[35] R. D. Zühlke, G. S. Pitt, R. W. Tsien, and H. Reuter. Ca2+-sensitive inactivation and facilitation of l-type ca2+ channels both depend on specific amino acid residues in a consensus calmodulin-binding motif in the(alpha)1c subunit. *J Biol Chem*, 275(28):21121–21129, Jul 2000.

[36] G. S. Pitt, R. D. Zühlke, A. Hudmon, H. Schulman, H. Reuter, and R. W. Tsien. Molecular basis of calmodulin tethering and ca2+-dependent inactivation of l-type ca2+ channels. *J Biol Chem*, 276(33):30794–30802, Aug 2001.

[37] B. Z. Peterson, C. D. DeMaria, J. P. Adelman, and D. T. Yue. Calmodulin is the ca2+ sensor for ca2+ -dependent inactivation of l-type calcium channels. *Neuron*, 22(3):549–558, Mar 1999.

[38] Michael G Erickson, Haoya Liang, Masayuki X Mori, and David T Yue. Fret two-hybrid mapping reveals function and location of l-type ca2+ channel cam preassociation. *Neuron*, 39(1):97–107, Jul 2003.

[39] C. D. DeMaria, T. W. Soong, B. A. Alseikhan, R. S. Alvania, and D. T. Yue. Calmodulin bifurcates the local ca2+ signal that modulates p/q-type ca2+ channels. *Nature*, 411(6836):484–489, May 2001.

[40] Badr A Alseikhan, Carla D DeMaria, Henry M Colecraft, and David T Yue. Engineered calmodulins reveal the unexpected eminence of ca2+ channel inactivation in controlling heart excitation. *Proc Natl Acad Sci U S A*, 99(26):17185–17190, Dec 2002.

[41] Jennifer L Fallon, D. Brent Halling, Susan L Hamilton, and Florante A Quiocho. Structure of calmodulin bound to the hydrophobic iq domain of the cardiac ca(v)1.2 calcium channel. *Structure*, 13(12):1881–1886, Dec 2005.

[42] James Keeler. *Understanding NMR Spectroscopy*. Wiley, 2005.

[43] Malcolm H. Levitt. *Spin Dynamics*. Wiley, 2001.

[44] John Cavanagh, Wayne N. Fairbrother, Arthur G. Palmer III, Mark Rance, and Nicholas J. Skelton. *Protein NMR Spectroscopy - Principles and Practice*. Academic Press, Elsevier, 2 edition, 2007.

[45] Robert Schneider. *Crystalline, membrane-embedded, and fibrillar proteins investigated by solid-state NMR spectroscopy*. PhD thesis, Georg August University Göttingen, 2009.

[46] Ad Bax, James J. Chou, and Benjamin E. Ramirez. *Liquid crystalline samples: application to macromolecular structure determination*, volume 9 of *Methods in Enzymology*. Academic Press, Elsevier, 2001.

[47] W. Y. Choy, M. Tollinger, G. A. Mueller, and L. E. Kay. Direct structure refinement of high molecular weight proteins against residual dipolar couplings and carbonyl chemical shift changes upon alignment: an application to maltose binding protein. *J Biomol NMR*, 21(1):31–40, Sep 2001.

[48] J. Meiler, J. J. Prompers, W. Peti, C. Griesinger, and R. Brüschweiler. Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins. *J Am Chem Soc*, 123(25):6098–6107, Jun 2001.

[49] Wolfgang Peti, Jens Meiler, Rafael Brüschweiler, and Christian Griesinger. Model-free analysis of protein backbone motion from residual dipolar couplings. *J Am Chem Soc*, 124(20):5822–5833, May 2002.

[50] Guillaume Bouvignies, Pau Bernadó, and Martin Blackledge. Protein backbone dynamics from n-hn dipolar couplings in partially aligned systems: a comparison of motional models in the presence of structural noise. *J Magn Reson*, 173(2):328–338, Apr 2005.

[51] Guillaume Bouvignies, Phineus R L Markwick, and Martin Blackledge. Characterization of protein dynamics from residual dipolar couplings using the three dimensional gaussian axial fluctuation model. *Proteins*, 71(1):353–363, Apr 2008.

[52] Oliver F Lange, Nils-Alexander Lakomek, Christophe Farès, Gunnar F Schröder, Korvin F A Walter, Stefan Becker, Jens Meiler, Helmut Grubmüller, Christian Griesinger, and Bert L de Groot. Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–1475, Jun 2008.

[53] Ivano Bertini, Ioannis Gelis, Nikolaos Katsaros, Claudio Luchinat, and Alessandro Provenzani. Tuning the affinity for lanthanides of calcium binding proteins. *Biochemistry*, 42(26):8011–8021, Jul 2003.

[54] Ivano Bertini, Cristina Del Bianco, Ioannis Gelis, Nikolaus Katsaros, Claudio Luchinat, Giacomo Parigi, Massimiliano Peana, Alessandro Provenzani, and Maria Antonietta Zoroddu. Experimentally exploring the conformational space sampled by domain reorientation in calmodulin. *Proc Natl Acad Sci U S A*, 101(18):6841–6846, May 2004.

[55] Takahisa Ikegami, Laurent Verdier, Peyman Sakhaii, Susanne Grimme, Barbara Pescatore, Krishna Saxena, Klaus M Fiebig, and Christian Griesinger. Novel techniques for weak alignment of proteins in solution using chemical tags coordinating lanthanide ions. *J Biomol NMR*, 29(3):339–349, Jul 2004.

[56] Andrei Leonov, Brigitte Voigt, Fernando Rodriguez-Castañeda, Peyman Sakhaii, and Christian Griesinger. Convenient synthesis of multifunctional edta-based chiral metal chelates substituted with an s-mesylcysteine. *Chemistry*, 11(11):3342–3348, May 2005.

[57] Jens Wöhnert, Katherine J Franz, Mark Nitz, Barbara Imperiali, and Harald Schwalbe. Protein alignment by a coexpressed lanthanide-binding tag for the measurement of residual dipolar couplings. *J Am Chem Soc*, 125(44):13338–13339, Nov 2003.

[58] Langdon J Martin, Martin J Hähnke, Mark Nitz, Jens Wöhnert, Nicholas R Silvaggi, Karen N Allen, Harald Schwalbe, and Barbara Imperiali. Double-lanthanide-binding tags: design, photophysical properties, and nmr applications. *J Am Chem Soc*, 129(22):7106–7113, Jun 2007.

[59] L.C. Snyder. Analysis of molecular magnetic resonance spectra of molecules in liquid-crystal solvents. *J Chem Phys*, 43:4041–4050, 1965.

[60] N. Tjandra and A. Bax. Direct measurement of distances and angles in biomolecules by nmr in a dilute liquid crystalline medium. *Science*, 278(5340):1111–1114, Nov 1997.

[61] A. Saupe. Recent results in the field of liquid crystals. *Angewandte Chemie*, 7:97–112, 1968.

[62] Ad Bax, Georg Kontaxis, and Nico Tjandra. *Methods in Enzymology*, volume 339. Academic Press, Elsevier, 2001.

[63] Maurice Guéron. Nuclear relaxation in macromolecules by paramagnetic ions - novel mechanism. *J Magn Reson*, 19:58–66, 1975.

[64] J. C. Hus, D. Marion, and M. Blackledge. De novo determination of protein structure by nmr using orientational and long-range order restraints. *J Mol Biol*, 298(5):927–936, May 2000.

[65] P. K. Madhu, Pravat K Mandal, and Norbert Müller. Cross-correlation effects involving curie spin relaxation in methyl groups. *J Magn Reson*, 155(1):29–38, Mar 2002.

[66] Marco Longinetti, Claudio Luchinat, Giacomo Parigi, and Luca Sgheri. Efficient determination of the most favoured orientations of protein domains from paramagnetic nmr data. *Inverse Problems*, 22:1485–1502, 2006.

[67] Richard J. Garnder, Marco Longinetti, and Luca Sgheri. Reconstruction of orientations of a moving protein domain from paramagnetic data. *Inverse Problems*, 21:879–898, 2005.

[68] Ivano Bertini, Yogesh K Gupta, Claudio Luchinat, Giacomo Parigi, Massimiliano Peana, Luca Sgheri, and Jing Yuan. Paramagnetism-based nmr restraints provide maximum allowed probabilities for the different conformations of partially independent protein domains. *J Am Chem Soc*, 129(42):12786–12794, Oct 2007.

[69] J. Grodberg and J. J. Dunn. ompt encodes the escherichia coli outer membrane protease that cleaves t7 rna polymerase during purification. *J Bacteriol*, 170(3):1245–1253, Mar 1988.

[70] J. Sambrook, D. Russell, and D.W. Russell. *Molecular cloning: A laboratory manual.* Cold Spring Harbor Laboratory Press, 3rd edition, 2001.

[71] Peter Haberz, Fernando Rodriguez-Castañeda, Jochen Junker, Stefan Becker, Andrei Leonov, and Christian Griesinger. Two new chiral edta-based metal chelates for weak alignment of proteins in solution. *Org Lett*, 8(7):1275–1278, Mar 2006.

[72] Fernando Rodriguez-Castañeda, Peter Haberz, Andrei Leonov, and Christian Griesinger. Paramagnetic tagging of diamagnetic proteins for solution nmr. *Magn Reson Chem*, 44 Spec No:S10–S16, Jul 2006.

[73] M. Ottiger, F. Delaglio, and A. Bax. Measurement of j and dipolar couplings from simplified two-dimensional nmr spectra. *J Magn Reson*, 131(2):373–378, Apr 1998.

[74] W.L. DeLano. The pymol molecular graphics system, 2002.

[75] Rochus Keller. *The Computer Aided Resonance Assignment Tutorial*. CANTINA Verlag, 2004.

[76] Christophe Schmitz, Michael John, Ah Young Park, Nicholas E Dixon, Gottfried Otting, Guido Pintacuda, and Thomas Huber. Efficient chi-tensor determination and nh assignment of paramagnetic proteins. *J Biomol NMR*, 35(2):79–87, Jun 2006.

[77] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. Nmrpipe: a multidimensional spectral processing system based on unix pipes. *J Biomol NMR*, 6(3):277–293, Nov 1995.

[78] T.D. Goddard and D.G. Kneller. Sparky 3. University of California, San Francisco.

[79] Christophe Schmitz, Mitchell J Stanton-Cook, Xun-Cheng Su, Gottfried Otting, and Thomas Huber. Numbat: an interactive software tool for fitting deltachi-tensors to molecular coordinates using pseudocontact shifts. *J Biomol NMR*, 41(3):179–189, Jul 2008.

[80] Markus Zweckstetter. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by nmr. *J. Am. Chem. Soc.*, 122:3791–3792, 2000.

[81] Markus Zweckstetter. Nmr: prediction of molecular alignment from structure using the pales software. *Nat Protoc*, 3(4):679–690, 2008.

[82] Edward J d'Auvergne and Paul R Gooley. Optimisation of nmr dynamic models i. minimisation algorithms and their performance within the model-free and brownian rotational diffusion spaces. *J Biomol NMR*, 40(2):107–119, Feb 2008.

[83] I. Bertini, M. B. Janik, Y. M. Lee, C. Luchinat, and A. Rosato. Magnetic susceptibility tensor anisotropies for a lanthanide ion series in a fixed protein matrix. *J Am Chem Soc*, 123(18):4181–4188, May 2001.

[84] Guido Pintacuda, Michael John, Xun-Cheng Su, and Gottfried Otting. Nmr structure determination of protein-ligand complexes by lanthanide labeling. *Acc Chem Res*, 40(3):206–212, Mar 2007.

[85] Fernando Rodriguez-Castañeda. *NMR spectroscopic studies of calmodulin plasticity in calcium singalling*. PhD thesis, IMPRS, Georg August University Göttingen, 2007.

[86] Najl V Valeyev, Pat Heslop-Harrison, Ian Postlethwaite, Nicolai V Kotov, and Declan G Bates. Multiple calcium binding sites make calmodulin multifunctional. *Mol Biosyst*, 4(1):66–73, Jan 2008.

[87] Najl V Valeyev, Declan G Bates, Pat Heslop-Harrison, Ian Postlethwaite, and Nikolay V Kotov. Elucidating the mechanisms of cooperative calcium-calmodulin interactions: a structural systems biology approach. *BMC Syst Biol*, 2:48, 2008.

[88] M. Ikura, L. E. Kay, and A. Bax. A novel approach for sequential assignment of 1h, 13c, and 15n spectra of proteins: heteronuclear triple-resonance three-dimensional nmr spectroscopy. application to calmodulin. *Biochemistry*, 29(19):4659–4667, May 1990.

[89] Michael Sattler, Jürgen Schleucher, and Christian Griesinger. Heteronuclear multidimensional nmr experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. NMR Spectrosc.*, 34:93–158, 1999.

[90] J.C. Lindon and A. G. Ferrige. Nmr data processing. *Prog. NMR Spectrosc.*, 14:27, 1980.

[91] Giovanni Lipari and Attila Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *J. Am. Chem. Soc.*, 104:4546–4559, 1982.

[92] Edward J d'Auvergne and Paul R Gooley. Model-free model elimination: a new step in the model-free dynamic analysis of nmr relaxation data. *J Biomol NMR*, 35(2):117–135, Jun 2006.

[93] J. J. Chou, S. Li, C. B. Klee, and A. Bax. Solution structure of ca(2+)-calmodulin reveals flexible hand-like properties of its domains. *Nat Struct Biol*, 8(11):990–997, Nov 2001.

[94] Elodie Laine, Julliane D Yoneda, Arnaud Blondel, and Thérèse E Malliavin. The conformational plasticity of calmodulin upon calcium complexation gives a model of its interaction with the oedema factor of bacillus anthracis. *Proteins*, 71(4):1813–1829, Jun 2008.

[95] M. A. Schumacher, A. F. Rivard, H. P. Bächinger, and J. P. Adelman. Structure of the gating domain of a ca2+-activated k+ channel complexed with ca2+/calmodulin. *Nature*, 410(6832):1120–1124, Apr 2001.

[96] Anne Houdusse, Jean-François Gaucher, Elena Krementsova, Suet Mui, Kathleen M Trybus, and Carolyn Cohen. Crystal structure of apo-calmodulin bound to the first

two iq motifs of myosin v reveals essential recognition features. *Proc Natl Acad Sci U S A*, 103(51):19326–19331, Dec 2006.

[97] Jennifer L Fallon, Mariah R Baker, Liangwen Xiong, Ryan E Loy, Guojun Yang, Robert T Dirksen, Susan L Hamilton, and Florante A Quiocho. Crystal structure of dimeric cardiac l-type calcium channel regulatory domains bridged by ca2+* calmodulins. *Proc Natl Acad Sci U S A*, 106(13):5135–5140, Mar 2009.

[98] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, Dec 1999.

[99] J. R. Yates, J. K. Eng, A. L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, 67(8):1426–1436, Apr 1995.

[100] Simone Cristoni and Luigi Rossi Bernardi. Development of new methodologies for the mass spectrometry study of bioorganic macromolecules. *Mass Spectrom Rev*, 22(6):369–406, 2003.

[101] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, Oct 1989.

[102] E. J. Zaluzec, D. A. Gage, and J. T. Watson. Matrix-assisted laser desorption ionization mass spectrometry: applications in peptide and protein characterization. *Protein Expr Purif*, 6(2):109–123, Apr 1995.

[103] Gary J Van Berkel. An overview of some recent developments in ionization methods for mass spectrometry. *Eur J Mass Spectrom (Chichester, Eng)*, 9(6):539–562, 2003.

[104] Francesco L Brancia. Recent developments in ion-trap mass spectrometry and related technologies. *Expert Rev Proteomics*, 3(1):143–151, Feb 2006.

[105] Richard H Perry, R. Graham Cooks, and Robert J Noll. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom Rev*, 27(6):661–699, 2008.

[106] J. Throck Watson and O. David Sparkman. *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation.* Wiley, 4 edition, 2007.

[107] Jürgen H. Gross. *Mass Spectrometry: A Textbook.* Springer, 2006.

[108] N. L. Anderson and N. G. Anderson. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19(11):1853–1861, Aug 1998.

[109] W. P. Blackstock and M. P. Weir. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol*, 17(3):121–127, Mar 1999.

[110] Ida Chiara Guerrera and Oliver Kleiner. Application of mass spectrometry in proteomics. *Biosci Rep*, 25(1-2):71–93, 2005.

[111] Xuemei Han, Aaron Aslanian, and John R Yates. Mass spectrometry for proteomics. *Curr Opin Chem Biol*, 12(5):483–490, Oct 2008.

[112] Andrew W Stoker. Protein tyrosine phosphatases and signalling. *J Endocrinol*, 185(1):19–33, Apr 2005.

[113] C. M. Pickart. Mechanisms underlying ubiquitination. *Annu Rev Biochem*, 70:503–533, 2001.

[114] Joerg Herrmann, Lilach O Lerman, and Amir Lerman. Ubiquitin and ubiquitin-like proteins in protein regulation. *Circ Res*, 100(9):1276–1291, May 2007.

[115] Helle D Ulrich. The sumo system: an overview. *Methods Mol Biol*, 497:3–16, 2009.

[116] Anne-Claude Gingras, Ruedi Aebersold, and Brian Raught. Advances in protein complex analysis using mass spectrometry. *J Physiol*, 563(Pt 1):11–21, Feb 2005.

[117] Anne-Claude Gingras, Matthias Gstaiger, Brian Raught, and Ruedi Aebersold. Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Biol*, 8(8):645–654, Aug 2007.

[118] Ignat V Shilov, Sean L Seymour, Alpesh A Patel, Alex Loboda, Wilfred H Tang, Sean P Keating, Christie L Hunter, Lydia M Nuwaysir, and Daniel A Schaeffer. The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol Cell Proteomics*, 6(9):1638–1655, Sep 2007.

[119] K. R. Clauser, P. Baker, and A. L. Burlingame. Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing ms or ms/ms and database searching. *Anal Chem*, 71(14):2871–2882, Jul 1999.

[120] Ronald Beavis and David Fenyö. Finding protein sequences using prowl. *Curr Protoc Bioinformatics*, Chapter 13:Unit 13.2, Oct 2004.

[121] W. R. Pearson. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the smith-waterman and fasta algorithms. *Genomics*, 11(3):635–650, Nov 1991.

[122] Shonda A Leonard, Timothy G Littlejohn, and Andreas D Baxevanis. Common file formats. *Curr Protoc Bioinformatics*, Appendix 1:Appendix 1B, Jan 2007.

[123] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–214, Mar 2007.

[124] Andrea Sinz. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom Rev*, 25(4):663–682, 2006.

[125] Stanley M Jeram, Tharan Srikumar, Patrick G A Pedrioli, and Brian Raught. Using mass spectrometry to identify ubiquitin and ubiquitin-like protein conjugation sites. *Proteomics*, 9(4):922–934, Feb 2009.

[126] Yu Xue, Fengfeng Zhou, Chuanhai Fu, Ying Xu, and Xuebiao Yao. Sumosp: a web server for sumoylation site prediction. *Nucleic Acids Res*, 34(Web Server issue):W254–W257, Jul 2006.

[127] C.J. Friedline, X.P. Zhang, Z.E. Zehne, and Z.M. Zhao. Findsumo: A pssm-based method for sumoylation site prediction. *Advanced Intelligent Computing Theories and Applications, Proceedings*, 5227:1004–1011, 2008.

[128] Matthew Knuesel, Hiu Tom Cheung, Micah Hamady, Kristen K B Barthel, and Xuedong Liu. A method of mapping protein sumoylation sites by mass spectrometry using a modified small ubiquitin-like modifier 1 (sumo-1) and a computational program. *Mol Cell Proteomics*, 4(10):1626–1636, Oct 2005.

[129] Patrick G A Pedrioli, Brian Raught, Xiang-Dong Zhang, Richard Rogers, John Aitchison, Michael Matunis, and Ruedi Aebersold. Automated identification of sumoylation sites using mass spectrometry and summon pattern recognition software. *Nat Methods*, 3(7):533–539, Jul 2006.

[130] Feixia Chu, Shu ou Shan, Demetri T Moustakas, Frank Alber, Pascal F Egea, Robert M Stroud, Peter Walter, and Alma L Burlingame. Unraveling the interface of signal recognition particle and its receptor by using chemical cross-linking and tandem mass spectrometry. *Proc Natl Acad Sci U S A*, 101(47):16454–16459, Nov 2004.

[131] Yong Tang, Yingfeng Chen, Cheryl F Lichti, Roger A Hall, Kevin D Raney, and Steven F Jennings. Clpm: a cross-linked peptide mapping algorithm for mass spectrometric analysis. *BMC Bioinformatics*, 6 Suppl 2:S9, Jul 2005.

[132] Oliver Rinner, Jan Seebacher, Thomas Walzthoeni, Lukas N Mueller, Martin Beck, Alexander Schmidt, Markus Mueller, and Ruedi Aebersold. Identification of cross-linked peptides from large sequence databases. *Nat Methods*, 5(4):315–318, Apr 2008.

[133] Alessio Maiolica, Davide Cittaro, Dario Borsotti, Lau Sennels, Claudio Ciferri, Cataldo Tarricone, Andrea Musacchio, and Juri Rappsilber. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol Cell Proteomics*, 6(12):2200–2211, Dec 2007.

[134] He-Hsuan Hsiao, Erik Meulmeester, Benedikt T.C. Frank, Frauke Melchior, and Henning Urlaub. Chop n spice, a mass spectrometric approach that allows identification of endogenous sumo conjugated peptides. *Mol Cell Proteomics*, 2009.

[135] R. Mahajan, L. Gerace, and F. Melchior. Molecular characterization of the sumo-1 modification of rangap1 and its role in nuclear envelope association. *J Cell Biol*, 140(2):259–270, Jan 1998.

[136] M. S. Rodriguez, J. M. Desterro, S. Lain, C. A. Midgley, D. P. Lane, and R. T. Hay. Sumo-1 modification activates the transcriptional response of p53. *EMBO J*, 18(22):6455–6461, Nov 1999.

[137] T. Sternsdorf, K. Jensen, B. Reich, and H. Will. The nuclear dot protein sp100, characterization of domains necessary for dimerization, subcellular localization, and modification by small ubiquitin-like modifiers. *J Biol Chem*, 274(18):12555–12566, Apr 1999.

[138] Francis Crick. *What Mad Pursuit: A Personal View of Scientific Discovery.* Basic Books, 1990.

[139] G. Seydoux and T. Schedl. The germline in c. elegans: origins, proliferation, and silencing. *Int Rev Cytol*, 203:139–185, 2001.

[140] A. Williamson and R. Lehmann. Germ cell development in drosophila. *Annu Rev Cell Dev Biol*, 12:365–391, 1996.

[141] C. Wylie. Germ cells. *Cell*, 96(2):165–174, Jan 1999.

[142] A. McLaren. Signaling for germ cells. *Genes Dev*, 13(4):373–376, Feb 1999.

[143] Shinya Aramaki, Fuminori Sato, Takako Kato, Tomoki Soh, Yukio Kato, and Masa-Aki Hattori. Molecular cloning and expression of dead end homologue in chicken primordial germ cells. *Cell Tissue Res*, 330(1):45–52, Oct 2007.

[144] Shinya Aramaki, Kaiyu Kubota, Tomoki Soh, Nobuhiko Yamauchi, and Masa-Aki Hattori. Chicken dead end homologue protein is a nucleoprotein of germ cells including primordial germ cells. *J Reprod Dev*, Feb 2009.

[145] Chitralekha Bhattacharya, Sita Aggarwal, Madhu Kumar, Amatul Ali, and Angabin Matin. Mouse apolipoprotein b editing complex 3 (apobec3) is expressed in germ cells and interacts with dead-end (dnd1). *PLoS ONE*, 3(5):e2315, 2008.

[146] Katja Horvay, Maike Claussen, Mathias Katzer, Jobst Landgrebe, and Tomas Pieler. Xenopus dead end mrna is a localized maternal determinant that serves a conserved function in germ cell development. *Dev Biol*, 291(1):1–11, Mar 2006.

[147] Krasimir Slanchev, Jürg Stebler, Guillermo de la Cueva-Méndez, and Erez Raz. Development without germ cells: the role of the germ line in zebrafish sex differentiation. *Proc Natl Acad Sci U S A*, 102(11):4074–4079, Mar 2005.

[148] Chitralekha Bhattacharya, Sita Aggarwal, Rui Zhu, Madhu Kumar, Ming Zhao, Marvin L Meistrich, and Angabin Matin. The mouse dead-end gene isoform alpha is necessary for germ cell and embryonic viability. *Biochem Biophys Res Commun*, 355(1):194–199, Mar 2007.

[149] Martijn Kedde, Markus J Strasser, Bijan Boldajipour, Joachim A F Oude Vrielink, Krasimir Slanchev, Carlos le Sage, Remco Nagel, P. Mathijs Voorhoeve, Josyanne van Duijse, Ulf Andersson Ørom, Anders H Lund, Anastassis Perrakis, Erez Raz, and Reuven Agami. Rna-binding protein dnd1 inhibits microrna access to target mrna. *Cell*, 131(7):1273–1286, Dec 2007.

[150] Ramesh S Pillai, Suvendra N Bhattacharyya, and Witold Filipowicz. Repression of protein synthesis by mirnas: how many mechanisms? *Trends Cell Biol*, 17(3):118–126, Mar 2007.

[151] Lee P Lim, Nelson C Lau, Philip Garrett-Engele, Andrew Grimson, Janell M Schelter, John Castle, David P Bartel, Peter S Linsley, and Jason M Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769–773, Feb 2005.

[152] Wigard P Kloosterman and Ronald H A Plasterk. The diverse functions of micrornas in animal development and disease. *Dev Cell*, 11(4):441–450, Oct 2006.

[153] Yoontae Lee, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Rådmark, Sunyoung Kim, and V. Narry Kim. The nuclear rnase iii drosha initiates microrna processing. *Nature*, 425(6956):415–419, Sep 2003.

[154] Richard I Gregory, Thimmaiah P Chendrimada, Neil Cooch, and Ramin Shiekhattar. Human risc couples microrna biogenesis and posttranscriptional gene silencing. *Cell*, 123(4):631–640, Nov 2005.

[155] Elisavet Maniataki and Zissimos Mourelatos. A human, atp-independent, risc assembly machine fueled by pre-mirna. *Genes Dev*, 19(24):2979–2990, Dec 2005.

[156] Ramesh S Pillai, Suvendra N Bhattacharyya, Caroline G Artus, Tabea Zoller, Nicolas Cougot, Eugenia Basyuk, Edouard Bertrand, and Witold Filipowicz. Inhibition of translational initiation by let-7 microrna in human cells. *Science*, 309(5740):1573–1576, Sep 2005.

[157] Shveta Bagga, John Bracht, Shaun Hunter, Katlin Massirer, Janette Holtz, Rachel Eachus, and Amy E Pasquinelli. Regulation by let-7 and lin-4 mirnas results in target mrna degradation. *Cell*, 122(4):553–563, Aug 2005.

[158] Cécile Esnault, Odile Heidmann, Frédéric Delebecque, Marie Dewannieux, David Ribet, Allan J Hance, Thierry Heidmann, and Olivier Schwartz. Apobec3g cytidine deaminase inhibits retrotransposition of endogenous retroviruses. *Nature*, 433(7024):430–433, Jan 2005.

[159] Cécile Esnault, Stéphane Priet, David Ribet, Odile Heidmann, and Thierry Heidmann. Restriction by apobec3 proteins of endogenous retroviruses with an extracellular life cycle: ex vivo effects and in vivo "traces" on the murine iape and human herv-k elements. *Retrovirology*, 5:75, 2008.

[160] F. William Studier. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif*, 41(1):207–234, May 2005.

[161] Joseph E Tropea, Scott Cherry, and David S Waugh. Expression and purification of soluble his(6)-tagged tev protease. *Methods Mol Biol*, 498:297–307, 2009.

[162] Elisabeth Gasteiger, Alexandre Gattiker, Christine Hoogland, Ivan Ivanyi, Ron D Appel, and Amos Bairoch. Expasy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*, 31(13):3784–3788, Jul 2003.

[163] Scott McGinnis and Thomas L Madden. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32(Web Server issue):W20–W25, Jul 2004.

[164] Jian Ye, Scott McGinnis, and Thomas L Madden. Blast: improvements for better sequence analysis. *Nucleic Acids Res*, 34(Web Server issue):W6–W9, Jul 2006.

[165] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. Ncbi blast: a better web interface. *Nucleic Acids Res*, 36(Web Server issue):W5–W9, Jul 2008.

[166] Edouard de Castro, Christian J A Sigrist, Alexandre Gattiker, Virginie Bulliard, Petra S Langendijk-Genevaux, Elisabeth Gasteiger, Amos Bairoch, and Nicolas Hulo. Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic Acids Res*, 34(Web Server issue):W362–W365, Jul 2006.

[167] Robert D Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R Eddy, Erik L L Sonnhammer, and Alex Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–D251, Jan 2006.

[168] Robert D Finn, John Tate, Jaina Mistry, Penny C Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R Eddy, Erik L L Sonnhammer, and Alex Bateman. The pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–D288, Jan 2008.

[169] Rune Linding, Lars Juhl Jensen, Francesca Diella, Peer Bork, Toby J Gibson, and Robert B Russell. Protein disorder prediction: implications for structural proteomics. *Structure*, 11(11):1453–1459, Nov 2003.

[170] Rune Linding, Robert B Russell, Victor Neduva, and Toby J Gibson. Globplot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, 31(13):3701–3708, Jul 2003.

[171] Christian Cole, Jonathan D Barber, and Geoffrey J Barton. The jpred 3 secondary structure prediction server. *Nucleic Acids Res*, 36(Web Server issue):W197–W201, Jul 2008.

[172] Burkhard Rost, Guy Yachdav, and Jinfeng Liu. The predictprotein server. *Nucleic Acids Res*, 32(Web Server issue):W321–W326, Jul 2004.

[173] L. J. McGuffin, K. Bryson, and D. T. Jones. The psipred protein structure prediction server. *Bioinformatics*, 16(4):404–405, Apr 2000.

[174] Kevin Bryson, Liam J McGuffin, Russell L Marsden, Jonathan J Ward, Jaspreet S Sodhi, and David T Jones. Protein structure prediction servers at university college london. *Nucleic Acids Res*, 33(Web Server issue):W36–W38, Jul 2005.

[175] C. Geourjon and G. Deléage. Sopma: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci*, 11(6):681–684, Dec 1995.

[176] L. A. Kelley, R. M. MacCallum, and M. J. Sternberg. Enhanced genome annotation using structural profiles in the program 3d-pssm. *J Mol Biol*, 299(2):499–520, Jun 2000.

[177] Christophe Combet, Martin Jambon, Gilbert Deléage, and Christophe Geourjon. Geno3d: automatic comparative molecular modelling of protein. *Bioinformatics*, 18(1):213–214, Jan 2002.

[178] Lawrence A Kelley and Michael J E Sternberg. Protein structure prediction on the web: a case study using the phyre server. *Nat Protoc*, 4(3):363–371, 2009.

[179] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The swiss-model workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, Jan 2006.

[180] Warren A Kibbe. Oligocalc: an online oligonucleotide properties calculator. *Nucleic Acids Res*, 35(Web Server issue):W43–W46, Jul 2007.

[181] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415, Jul 2003.

[182] S. Nakano, M. Fujimoto, H. Hara, and N. Sugimoto. Nucleic acid duplex stability: influence of base composition on cation effects. *Nucleic Acids Res*, 27(14):2957–2965, Jul 1999.

[183] P. M. Howley, M. A. Israel, M. F. Law, and M. A. Martin. A rapid method for detecting and mapping homology between heterologous dnas. evaluation of polyomavirus genomes. *J Biol Chem*, 254(11):4876–4883, Jun 1979.

[184] A. M. Maxam and W. Gilbert. A new method for sequencing dna. *Proc Natl Acad Sci U S A*, 74(2):560–564, Feb 1977.

[185] S. Chaudhuri, B. Jana, and T. Basu. Why does ethanol induce cellular heat-shock response? *Cell Biol Toxicol*, 22(1):29–37, Jan 2006.

[186] Manuel Ferrer, Tatyana N Chernikova, Michail M Yakimov, Peter N Golyshin, and Kenneth N Timmis. Chaperonins govern growth of escherichia coli at low temperatures. *Nat Biotechnol*, 21(11):1266–1267, Nov 2003.

[187] Perttu Permi. Aspects of coherence transfer in high molecular weight proteins. *Annu Rep NMR Spectr.*, 53:245–296, 2004.

[188] Catherine H. Schein. Production of soluble recombinant proteins in bacteria. *Bio/Technology*, 7:1141–1149, 1989.

[189] Andrea Vera, Nuria González-Montalbán, Anna Arís, and Antonio Villaverde. The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. *Biotechnol Bioeng*, 96(6):1101–1106, Apr 2007.

[190] Frank Hoffmann and Ursula Rinas. Roles of heat-shock chaperones in the production of recombinant proteins in escherichia coli. *Adv Biochem Eng Biotechnol*, 89:143–161, 2004.

[191] S. Kedzierska, M. Staniszewska, A. Wegrzyn, and A. Taylor. The role of dnak/dnaj and groel/groes systems in the removal of endogenous proteins aggregated by heat-shock from escherichia coli cells. *FEBS Lett*, 446(2-3):331–337, Mar 1999.

[192] François Baneyx and Joanne L Palumbo. Improving heterologous protein folding via molecular chaperone and foldase co-expression. *Methods Mol Biol*, 205:171–197, 2003.

[193] F. Arsène, T. Tomoyasu, and B. Bukau. The heat shock response of escherichia coli. *Int J Food Microbiol*, 55(1-3):3–9, Apr 2000.

[194] Lisa D Cabrita and Stephen P Bottomley. Protein expression and refolding–a practical guide to getting the most out of inclusion bodies. *Biotechnol Annu Rev*, 10:31–50, 2004.

[195] Nicola A Burgess-Brown, Sujata Sharma, Frank Sobott, Christoph Loenarz, Udo Oppermann, and Opher Gileadi. Codon optimization can improve expression of human genes in escherichia coli: A multi-gene study. *Protein Expr Purif*, 59(1):94–102, May 2008.

[196] C. I. Murphy and H. Piwnica-Worms. Overview of the baculovirus expression system. *Curr Protoc Neurosci*, Chapter 4:Unit 4.18, May 2001.

[197] Richard B Hitchman, Robert D Possee, and Linda A King. Baculovirus expression systems for recombinant protein production in insect cells. *Recent Pat Biotechnol*, 3(1):46–54, 2009.

[198] R. B. Kapust and D. S. Waugh. Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci*, 8(8):1668–1674, Aug 1999.

[199] Jeffrey D Fox and David S Waugh. Maltose-binding protein as a solubility enhancer. *Methods Mol Biol*, 205:99–117, 2003.

[200] Krasimir Ivanov Slanchev. *Molecular mechanisms governing germ line development in zebrafish and the role of this lineage in sexual differentiation.* PhD thesis, Georg August University Göttingen, 2005.

[201] Krasimir Slanchev, Juerg Stebler, Mehdi Goudarzi, Vlad Cojocaru, Gilbert Weidinger, and Erez Raz. Control of dead end localization and activity–implications for the function of the protein in antagonizing mirna function. *Mech Dev*, 126(3-4):270–277, 2009.

[202] A. A. Russo, P. D. Jeffrey, A. K. Patten, J. Massagué, and N. P. Pavletich. Crystal structure of the p27kip1 cyclin-dependent-kinase inhibitor bound to the cyclin a-cdk2 complex. *Nature*, 382(6589):325–331, Jul 1996.

[203] I. Kim, Y. Muto, M. Inoue, S. Watanabe, A. Kitamura, S. Yokoyama, K. Hosono, H. Takaku, A. Ono, M. Kainosho, H. Sakamoto, and Y. Shimura. Nmr analysis of the hydrogen bonding interactions of the rna-binding domains of the drosophila sex-lethal protein with target rna fragments with site-specific [3-15n]uridine substitutions. *Nucleic Acids Res*, 25(8):1565–1569, Apr 1997.

[204] M. R. Conte, T. Grüne, J. Ghuman, G. Kelly, A. Ladas, S. Matthews, and S. Curry. Structure of tandem rna recognition motifs from polypyrimidine tract binding protein reveals novel features of the rrm fold. *EMBO J*, 19(12):3132–3141, Jun 2000.

[205] M. Inoue, Y. Muto, H. Sakamoto, and S. Yokoyama. Nmr studies on functional structures of the au-rich element-binding domains of hu antigen c. *Nucleic Acids Res*, 28(8):1743–1750, Apr 2000.

[206] X. Wang and T. M. Tanaka Hall. Structural basis for recognition of au-rich element rna by the hud protein. *Nat Struct Biol*, 8(2):141–145, Feb 2001.

[207] Carina Johansson, L. David Finger, Lukas Trantirek, Thomas D Mueller, Soyoun Kim, Ite A Laird-Offringa, and Juli Feigon. Solution structure of the complex formed by the two n-terminal rna-binding domains of nucleolin and a pre-rrna target. *J Mol Biol*, 337(4):799–816, Apr 2004.

[208] R. C. Deo, J. B. Bonanno, N. Sonenberg, and S. K. Burley. Recognition of polyadenylate rna by the poly(a)-binding protein. *Cell*, 98(6):835–845, Sep 1999.

[209] M. Inoue, Y. Muto, H. Sakamoto, T. Kigawa, K. Takio, Y. Shimura, and S. Yokoyama. A characteristic arrangement of aromatic amino acid residues in the solution structure of the amino-terminal rna-binding domain of drosophila sex-lethal. *J Mol Biol*, 272(1):82–94, Sep 1997.

[210] M. Mar Carrió and Antonio Villaverde. Localization of chaperones dnak and groel in bacterial inclusion bodies. *J Bacteriol*, 187(10):3599–3601, May 2005.

[211] Margaret A. Speed, Daniel I. Wang, and Jonathan King. Specific aggregation of partially folded polypeptide chains: the molecular basis of inclusion body composition. *Nat Biotechnol*, 14(10):1283–1287, Oct 1996.

[212] Salvador Ventura and Antonio Villaverde. Protein quality in bacterial inclusion bodies. *Trends Biotechnol*, 24(4):179–185, Apr 2006.

[213] Mar Carrió, Nuria González-Montalbán, Andrea Vera, Antonio Villaverde, and Salvador Ventura. Amyloid-like properties of bacterial inclusion bodies. *J Mol Biol*, 347(5):1025–1037, Apr 2005.

[214] Lei Wang, Samir K Maji, Michael R Sawaya, David Eisenberg, and Roland Riek. Bacterial inclusion bodies contain amyloid-like structure. *PLoS Biol*, 6(8):e195, Aug 2008.

[215] Ursula Rinas, Frank Hoffmann, Eriola Betiku, David Estapé, and Sabine Marten. Inclusion body anatomy and functioning of chaperone-mediated in vivo inclusion body disassembly during high-level recombinant protein production in escherichia coli. *J Biotechnol*, 127(2):244–257, Jan 2007.

[216] Eliana De Bernardez Clark. Refolding of recombinant proteins. *Curr Opin Biotechnol*, 9(2):157–163, Apr 1998.

[217] Luis Vallejo and Ursula Rinas. Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microb Cell Fact*, 3(1):11, Sep 2004.

[218] Renaud Vincentelli, Stéphane Canaan, Valérie Campanacci, Christel Valencia, Damien Maurin, Frédéric Frassinetti, Loréna Scappucini-Calvo, Yves Bourne, Christian Cambillau, and Christophe Bignon. High-throughput automated refolding screening of inclusion bodies. *Protein Sci*, 13(10):2782–2792, Oct 2004.

[219] Melissa Swope Willis, James K Hogan, Prakash Prabhakar, Xun Liu, Kuenhi Tsai, Yunyi Wei, and Ted Fox. Investigation of protein refolding using a fractional factorial screen: a study of reagent effects and interactions. *Protein Sci*, 14(7):1818–1826, Jul 2005.

[220] Nick T. Spark. The fastest man on earth - history and background of murphy's law. *Annals of Improbable Research*, 9:4–26, 2003.

[221] Raymond C. Stevens. Design of high-throughput methods of protein production for structural biology. *Structure*, 8(9):R177–R185, Sep 2000.

[222] Renaud Vincentelli, Stéphane Canaan, Julien Offant, Christian Cambillau, and Christophe Bignon. Automated expression and solubility screening of his-tagged proteins in 96-well format. *Anal Biochem*, 346(1):77–84, Nov 2005.

[223] Sharon A Doyle. Screening for the expression of soluble recombinant protein in escherichia coli. *Methods Mol Biol*, 310:115–121, 2005.

[224] Cameron Neylon. Chemical and biochemical strategies for the randomization of protein encoding dna sequences: library construction methods for directed evolution. *Nucleic Acids Res*, 32(4):1448–1459, 2004.

[225] Amir Aharoni, Leonid Gaidukov, Shai Yagur, Lilly Toker, Israel Silman, and Dan S Tawfik. Directed evolution of mammalian paraoxonases pon1 and pon3 for bacterial expression and catalytic specialization. *Proc Natl Acad Sci U S A*, 101(2):482–487, Jan 2004.

[226] Darren J Hart and Franck Tarendeau. Combinatorial library approaches for improving soluble protein expression in escherichia coli. *Acta Crystallogr D Biol Crystallogr*, 62(Pt 1):19–26, Jan 2006.

[227] Franck Tarendeau, Julien Boudet, Delphine Guilligay, Philippe J Mas, Catherine M Bougault, Sébastien Boulo, Florence Baudin, Rob W H Ruigrok, Nathalie Daigle, Jan Ellenberg, Stephen Cusack, Jean-Pierre Simorre, and Darren J Hart. Structure and nuclear import function of the c-terminal domain of influenza virus polymerase pb2 subunit. *Nat Struct Mol Biol*, 14(3):229–233, Mar 2007.

[228] B. Keil. *Specificity of proteolysis.* Springer-Verlag Berlin-Heidelberg-New York, 1992.

Index

Curriculum Vitae

Benedikt T.C. Frank

born 31st December 1982 in Ulm, Germany

Obere Masch Str. 14, 37073 Göttingen
Phone +49 (0) 551 2829255
E-Mail benf@mpibpc.mpg.de

Education

2005 - present	Max Planck Institute for biophysical chemistry, Göttingen, Germany: Dept. NMR based structural biology. PhD thesis: *Investigation of Protein Structure and Dynamics*
2004 - 2005	Georg-August-University, Göttingen, Germany: International MSc/PhD program (direct admission to PhD)
2001 - 2004	Selwyn College, University of Cambridge, UK: Natural Sciences Tripos (MA *cantab.*).
1999 - 2001	Harrow School, London, UK: Sixth Form Scholar (5 A-levels).
1993 - 1999	Humboldt Gymnasium, Ulm, Germany (Mittlere Reife).

Scholarships & Prizes

2005 - 2008	Stipend Max-Planck Society
2004 - 2005	Stipend International Max-Planck Research School
2000 - 2002	Stipend Robert Bosch Foundation
1999 - 2001	Sixth Form Scholar, Harrow School, London
1999 - 2000	'UK Senior Maths Challenge': 1x Silver (1999), 1x Gold medal (2000)
1995 - 1998	Landeswettbewerb Mathematik, Baden-Württemberg: 1x 3rd Prize (1995), 1x 2nd Prize (1996), 2x 1st Prize (1997, 1998).