Evidence for multiple functions of Aprataxin in DNA damage repair

Dissertation

To Fulfill the
Requirements for the Degree of
"doctor rerum naturalium" (Dr. rer. nat.)

Submitted to the Council of the Faculty
of Biology and Pharmacy
of the Friedrich Schiller University Jena

by Diplom Biochemist Peter Bellstedt

born on 07.02.1985 **in** Eisenach (Thuringia)

Gutachter:

1. **Prof. Dr. Jens Wöhnert** (Frankfurt am Main)
2. **Prof. Dr. Frank Große** (Jena)
3. **PD Dr. Manuel E. Than** (Jena)

Datum der Verteidigung: **06.03.2014**

„Ein Experte ist ein Mann, der genau weiß, wie alles kommen wird, und der hinterher genau sagen kann, warum alles ganz anders gekommen ist.“

(Jack Lemmon)

Abreviations

APTX aprataxin

CD circular dichroism

CP cross polarization

ds double-stranded

HIT histidine triad

lsNMR liquid-state NMR spectroscopy

MAS magic angle spinning

MLS mitchondrial localization sequence

Mth *Methanobacterium thermoautotrophicum*

NLS nuclear localization sequence

NMR nuclear magnetic resonance

ORF open reading frame

PAA polyacrylamide

ss single-stranded

SSB single strand break

ssNMR solid-state NMR spectroscopy

wt wild-type

ZnF C_2H_2-type zinc finger

Table of contents

1 Introduction

The ability to store and propagate genetic information was probably the most crucial evolutionary achievement in the course of the development of life. The basis for our current molecular understanding of maintenance, replication and transmission of genetic material was provided by the identification of DNA as the genetic storage media [4], followed by the pioneering description of the double-stranded helical structure of DNA [105]. It took eight years, to answer the question of *how* exactly genetic information is encoded: The amino acid sequence of proteins is encoded by triplets of bases in a DNA strand [23, 66] and transient RNA species are responsible for shuttling this information from the storage location (DNA) to the ribosomes as the place of protein synthesis [69, 104]. The description of how DNA is packed (at least in eukaryotes) inside the nucleus of cells and the finding that the expression of genes might be controlled by dynamic association of DNA with chromatin proteins [32], increased the complexity of the topic and opened a wide array of research areas. The possibility to extract genetic information by DNA sequencing [67, 84] enabled scientists to track a large number of complex clinical syndromes such as premature aging, neurodegeneration, developmental defects and cancer to mutations in genes coding for proteins active in DNA repair [68, 79]. This dissertation addresses the structural and functional characterization of the DNA repair protein Aprataxin (APTX), which - if mutated or absent - is responsible for a distinct subtype of autosomal recessive cerebellar ataxias (ARCAs) called ataxia with oculomotor apraxia type 1 (AOA1). The clinical classification and phenotype of AOA1 is summarized in more detail in section 1.1. The classification of APTX with respect to protein family and its unique domain organization is outlined in section 1.2, followed by an review of APTX unique role during DNA single-strand break repair in section 1.3. As reported in section 1.4, the initial goal of this project was the characterization of the structure-function relationship of human APTX. Yet, the most interesting findings turned out to be novel and unexpected nucleolytic activities of this protein *in vitro*. Although these biochemical observations, which are presented in chapter 3, raise much more questions than this dissertation can answer, it provides a basis for comprehensive *in vivo* studies and may significantly widen the view on current models of DNA damage repair in the future.

1.1 Clinical classification of autosomal cerebellar ataxias including AOA1

ARCAs are a diverse group of rare neurodegenerative disorders phenotypically dominated by a progressive cerebellar degeneration, which manifests itself in the impairment of gait, balance and speech, disturbed upper limb coordination and eye movement abnormalities [43, 102]. Clinical features are observed before the age of 20, placing this subgroup of hereditary ataxias into early onset diseases [76]. AOA1 is after Friedreich's ataxia (FA), which accounts for 30% to 40% of patients diagnosed with ARCA, the most frequent recessive ataxia in many populations but with an increased incidence in Portugal, Japan, France and Tunisia [43, 60]. Interestingly FA is virtually not found in Japan, rendering AOA1 the most common ARCA within this population [54, 76]. Although in 1988 patients with cerebellar ataxia and oculomotor apraxia (AOA) lacking other typical features of ARCAs have been reported, AOA was recognized only in 2001 as a distinct subclass of ARCAs [6]. Up to now, AOA consists of at least 4 distinct neurodegenerative phenotypes: ataxia-telangiectasia (AT), ataxia-telangiectasia-like disorder (ATLD) as well as type 1 and 2 of ataxia with oculomotor apraxia (AOA1 / AOA2). All of these phenotypes are caused by mutations in single genes, namely *ATM*, *MRE11*, *APTX* and *SETX*, respectively, and all of the affected proteins act in DNA repair pathways. ATM and MRE11 both function at the initial phase of the response to DNA double-strand breaks [60]. The physiological function of SETX has recently been associated with the resolution of R-loops (DNA/RNA hybrids), which arise during transcriptional pausing and, if not resolved efficiently, lead to genomic instability [106]. APTX predominantly operates at DNA single-strand breaks [2, 71] and its unique catalytic activity in resolving abortive ligation intermediates will be addressed in more detail in section 1.2. The clinical term AOA refers to the characteristical oculomotor abnormalities. AOA1 patients which are *e.g.* asked to look at a lateral target, suffer from asynchronous movements, with the result, that the head reaches the lateral target before the eyes ("head-eye-dissociation") [60]. Strictly speaking, the neurological term oculomotor apraxia is defined by a failure of saccade *initiation*, but oculomotor apraxia in AOA1 is characterized by the loss of vestibulo-ocular reflex *cancellation*. Still, this terminology has been maintained since the first description of AOA1 although not perfectly matching the medical definition [60]. After the early onset of the disease with a mean age of 7 years, the neuropathy leads to rapid and severe disability. The patients become wheelchair-bound within 18 years on average [59]. Histological examination of a sural nerve[1] biopsy

[1]The sural nerve (also called saphenous nerve) is a nerve in the leg. Since it has a purely sensory function it is often used for biopsies.

revealed a severe loss of myelinated nerve fibres, whereas the unmyelinated nerve fibres were preserved [59]. This loss of myelinated nerve fibres may account for the impaired vibration sense, decreased general reflexes, chorea and neuropathy in AOA1. The ataxia and the oculomotor apraxia potentially can be attributed to the atrophic cerebellum [59] and the reduced number of Purkinje cells in the cerebellum [75, 96], which play an important role in the control and coordination of (ocular) movements [48, 50].

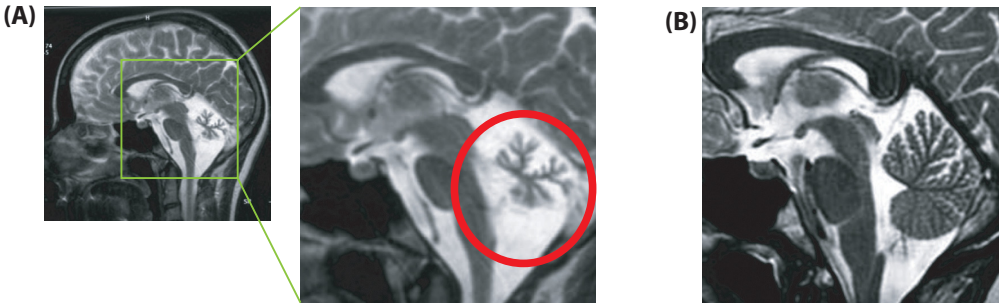


Figure 1.1: Brain imaging of AOA1 patient compared to a healthy person. Sagittal MRI section revealing severe cerebellar atrophy (A) as compared to a healthy individual (B). Due to the T2-weighted MRI method used here, cellular structures appear dark, regions with high water content bright. Figure adapted from Le Ber et al. [59] and Naidich et al. [72], respectively.

An extensive account of clinical symptoms associated with AOA1 including the non-neurological phenotypes as *e.g.* hypoalbuminaemia and hypercholesterolaemia is given in [59].

1.2 Molecular and structural properties of APTX as a representative of histidine triade proteins

APTX consists of an N-terminal forkhead associated (FHA) domain followed by a nuclear localization sequence (NLS), the enzymatically active histidine triade (HIT) domain and a C_2H_2-type zinc finger (ZnF) motif (Figure 1.2). APTX interacts *via* the FHA domain with XRCC1 [25, 39] and XRCC4 [16, 21], both of which are important scaffolding proteins involved in the repair of DNA single strand and double strand breaks, respectively. Furthermore the FHA domain interacts with the N-terminal acidic region of nucleolin [7], which is found in the nucleolus of cells active in rRNA synthesis but the functional role of this interaction has yet to be identified.

Interestingly, the FHA domain of APTX is replaced by other functional domains in non-vertebrates and is not found in lower eukaryotes [80]. However all APTX orthologs are characterized by a common core of at least one structurally highly conserved HIT-ZnF combina-

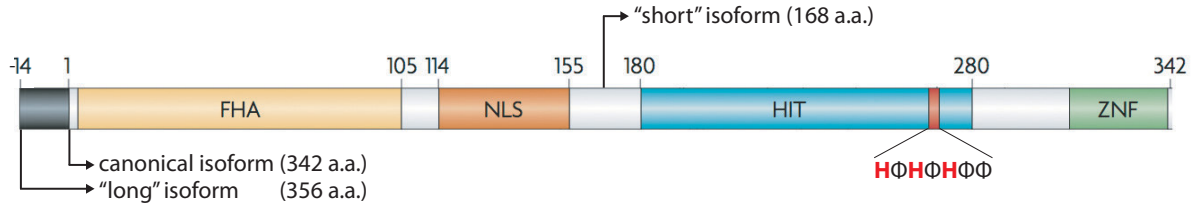


Figure 1.2: Domain organization of human Aprataxin including the major isoforms. FHA: Forkhead associated domain; NLS: Nuclear localization sequence; HIT: Histidine triade motif (highlighted in red) characterized by HΦHΦHΦΦ, whereas Φ is a hydrophobic residue; ZNF: C_2H_2-type zinc finger motive. Colored in Black: N-terminal extension with the putative mitochondrial targeting sequence. Figure modified from Caldecott [18].

tion. This suggests that the function of this unique combination is essential in eukaryotes but that additional functions carried out by "accessory domains" have evolved differently (Figure 1.3). More importantly, the Hit-ZnF combination is obviously the functional core unit of all APTX orthologs.

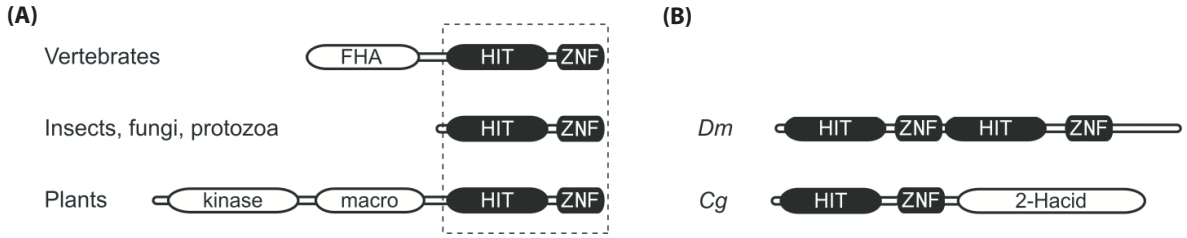


Figure 1.3: Domain organization of Aprataxin orthologs. **(A)** The combination of forkhead associated domain (FHA) with the combination of histidine tirade (HIT) and zinc finger (ZNF) is restricted to vertebrates. In plants the FHA domain is replaced by a macro and kinase domain. **(B)** Examples of atypical APTX orthologs: A dimeric head to tail fusion is found in *Drosophila melanogaster* (Dm) comprises a dimeric head to tail function. In the one-celled alge *C. globosum* (Cg), the HIT-ZNF combination is linked to a D-isomer-specific 2-hydroxyacid dehydrogenase (2-Hacid) domain. Interestingly, no APTX-like function has been identified in *S. cerevisiae* and *C. elegans* so far, even though HIT-family members are encoded in their genomes. Figure adapted from Rass et al. [80].

APTX is an atypical representative of histidine triade (HIT) proteins, which are a superfamily of nucleotide hydrolases and transferases acting on the α-phosphate of ribonucleotides [13] and in various cellular pathways. HIT proteins were first identified [88] solely on the basis of the common characteristic motif HΦHΦHΦΦ, whereas Φ is a hydrophobic residue (Figure 1.2). APTX forms a discrete branch within this HIT superfamily [51] with a unique and highly conserved HIT-ZnF domain combination. Based on sequence analysis and functional similarities there are four more branches comprising the histidine triad nucleotide-binding protein (HINT), fragile histidine triad (FHIT), galactose-1 phosphate uridylyltransferase (GALT), and finally DcpS, a scavenger pyrophosphatase [13]. Although the physiological nucleotide-containing substrate can differ, the biochemical (nucleotide transferase / hydrolase) reactions catalyzed by HIT proteins are mechanistically similar [12]. Furthermore, the residues involved

in substrate binding seem to be structurally conserved among the members of this superfamily [11]. Domains, adjacent to the HIT domain, seem to determine the distinct role of the respective HIT protein within their cellular context, yet only APTX (DNA repair) and DcpS (mRNA scavenger decapping) function in nucleic acid metabolism [13, 21, 62]. Typically, HIT proteins operate as homodimers with two independently active sites or comprise an imperfect internal polypeptide repeat retaining a single active site only [13, 63]. However, APTX only operates in a monomeric state [100] and in addition, no other protein has yet been identified which is able remove abortive ligation intermediates resulting from improper single strand break repair as catalyzed by APTX (see below, section 1.3).

Although several splice variants of APTX have been reported [24, 40, 44, 85], the protein encoded by the major isoform of the transcript comprises of 342 amino acids (a.a.) and is ubiquitously expressed, above all in the liver, the whole brain and there particularly in the cerebellum and the hypothalamus [97]. The most frequently reported minor isoforms are derived from alternative splicing of the *APTX* transcript and result in the expression of a longer (356 a.a.) or an N-terminally truncated, shorter isoform (168 a.a.), respectively (Figure 1.2). Nevertheless, the 342 a.a. APTX protein is considered to be the canonical form since it is expressed in high amounts in most tissues (including the brain) [24] and will herein be referred to - unless stated otherwise. Inside the cell APTX localizes to the nucleus with a significant fraction found in the nucleolus [7]. The biological role of the shorter isoform, which is predominantly found in the cytoplasm [39], is unclear. In contrast, there is some evidence that the additional 14 a.a. at the N-terminus, as found in the longer isoform, harbors a putative mitochondrial targeting sequence. This suggests a role in the repair of damaged *mitochondrial* DNA for this isoform [97].

1.3 Role of APTX in DNA repair of single-strand breaks

Since the seminal paper linking Xeroderma pigmentosum with DNA repair deficiency [20], a large number of complex clinical syndromes have been shown to be caused by mutations in genes coding for proteins involved in DNA damage repair [68, 79]. Even though such DNA repair genes have been identified, in many cases it remains to be elucidated, why mutations of a certain repair gene primarily manifest themselves in only a subset of tissues or distinct cell populations within a given tissue.

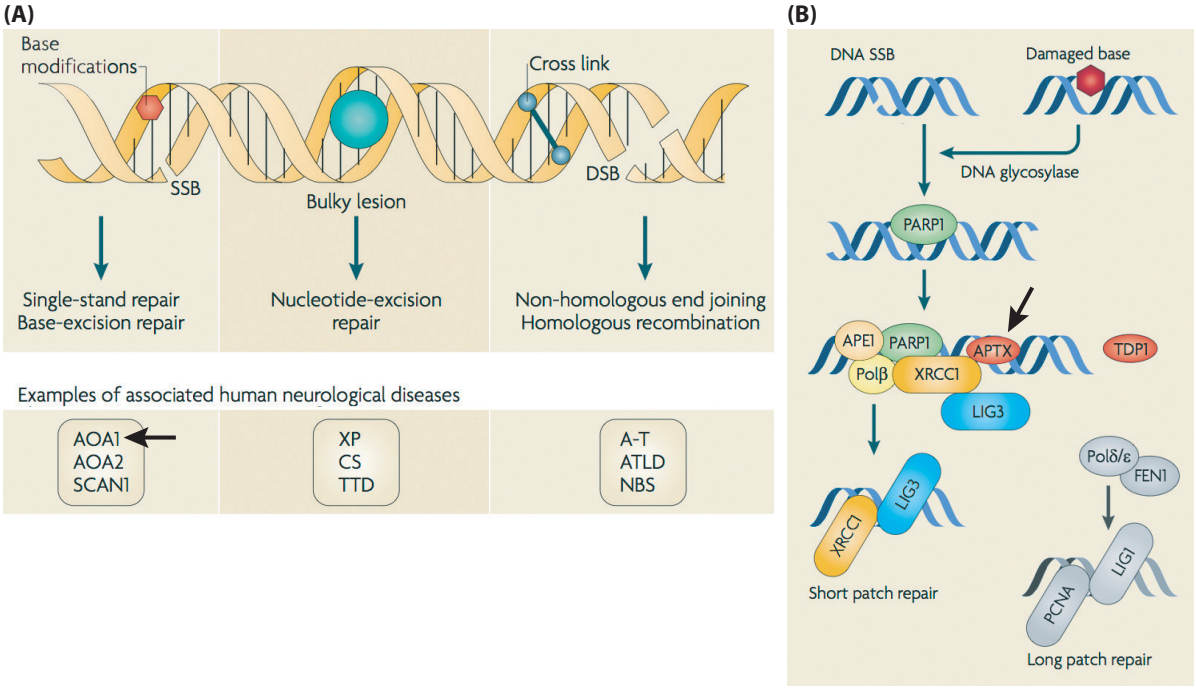


Figure 1.4: Types of DNA damage and repair (A) Defective repair of SSBs can lead to ataxia with oculomotor apraxia 1 (AOA1), AOA2 or spinocerebellar ataxia with axonal neuropathy 1 (SCAN1), whereas defects in nucleotide excision repair can result in xeroderma pigmentosum (XP), Cockayne Syndrome (CS) or trichothiodystrophy (TTD). Defective responses to DSBs can lead to ataxia telangiectasia (AT), AT-like disease (ATLD) or Nijmegen breakage syndrome (NBS). **(B)** Direct modification of the DNA backbone and base-excision repair cause single strand breaks, which lead to a local accumulation of poly(ADP-ribose)polymerase 1 (PARP1). PARP1 facilitates the recruitment of the scaffold protein XRCC1, which is essential for promoting the repair process by in turn recruiting further repair factors to the site of damage. The intention of the repair is restoring both ends of the break to allow for a rejoining mediated by Ligase 3. Depending on the nature of damage these repair factors *e.g.* resolve 5′-topoisomerase I DNA adducts (TDP1) or 3′-abortive ligation products (APTX). In non-proliferating cells, the short patch repair pathway involving the removal of single nucleotides is more operational than the long patch repair mediated by PCNA and DNA Ligase 1. APE1: apurinic/apyrimidinic endonuclease 1; APTX: Aprataxin; FEN1: flap structure-specific endonuclease 1; LIG3: DNA Ligase III; Pol: DNA polymerase; TDP1: tyrosyl DNA phosphodiesterase 1. Figure adapted from McKinnon [68].

DNA damage can be classified into four categories requiring single strand break (SSB) repair, base excision repair (BER), nucleotide excision repair (NER) and double strand break repair (Figure 1.4 A). Each day more than 10.000 DNA SSBs arise in our cells due to direct chemical modification by reactive oxygen species (ROS), irradiation or chemicals [18]. In addition, damage of DNA bases trigger the BER involving the AP endocuclease or DNA glycosylases resulting in the introduction of “indirect” breaks [27]. Unrepaired direct or indirect SSBs cause blockage or collapse of the replication fork and produce double strand breaks in mitotically active cells and/or lead to abortive transcription or stalling of RNA polymerases. The repair of SSBs may occur as “short patch” repair (Figure 1.4 B) carried out by a well known set of proteins, which detect SSBs (PARP1), process DNA termini (PNPK, APE1,

DNA Polβ) and ligate the processed 3′-OH and 5′-phosphate termini (DNA ligase 3, Lig3). XRCC1 functions as scaffolding protein in this process by interacting with PNPK, PARP1, Lig3, DNA Polβ and APTX [16, 17]. Furthermore, in cycling cells the flap endonuclease FEN1, PCNA and replicative DNA polymerases expedite the "long patch" repair. Importantly, postmitotic cells such as terminally differentiated neurons probably rely mainly on the short patch mechanism, as activities required for long patch are available to a lesser extend [17, 18].

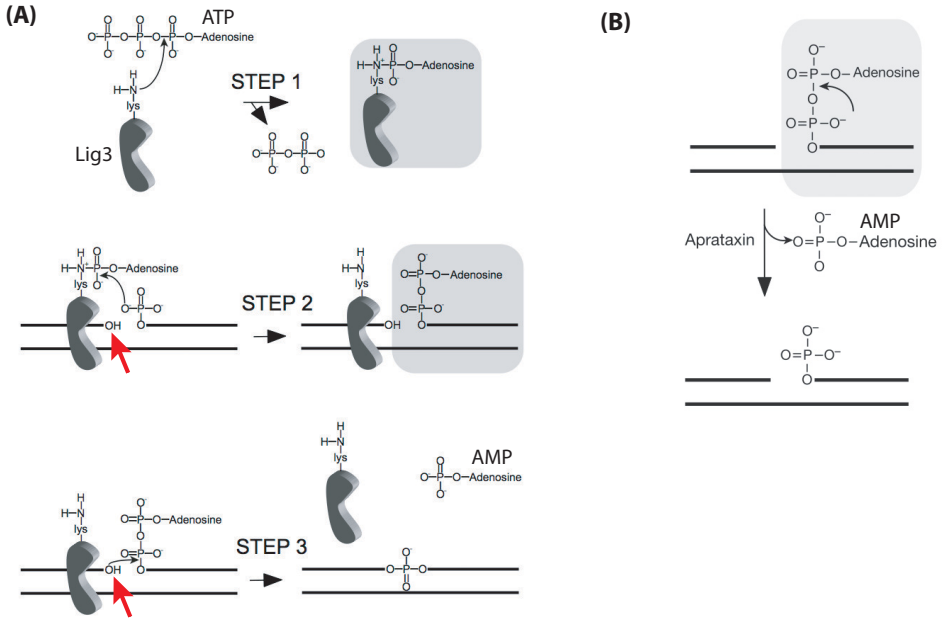


Figure 1.5: Ligase 3 mediated SSB rejoining and deadenylation catalyzed by Aprataxin (A) Reaction mechanism of DNA Ligase 3 (Lig3): Step 1 - Activation of Lig3 by hydrolysis of ATP; Step 2 - Activation of 5′-phosphate at SSB leading to an rAMP-p-DNA intermediate; Step 3 - Nucleophilic attack of 3′-hydroxy group resulting in the rejoining of the break. Red arrows indicate, that the hydroxy group is essential for the rejoining step, but not for the activation of the phosphate group. Attributable to the latter fact, the absence of a 3′-hydroxy group results in an abortive ligation with a characteristical adenylate structure at the 5′ end of the break, preventing further ligation events even if the 3′-hydroxy group is being restored. **(B)** Aprataxin resolves abortive ligation intermediates by removing the adenylate from the 5′ end of the DNA break, enabling a new round of ligation after the restoring of the 3′-OH by other repair factors. Figure modified from Ahel et al. [2]

During the repair of direct or indirect SSBs so-called "dirty" DNA breaks such as 3′-phosphates, oxidation products and adducts of the sugar moiety at the 5′ or 3′ end of the break [18] may arise. Such chemical modifications prevent either gap filling by Polβ or the final ligation step by Lig3, which requires 3′-OH and 5′-phosphate termini (Figure 1.5 A). Lig3 activates the 5′-phosphate to be ligated by attaching a rAMP to it, thereby creating a 5′-5′ rAMP-p-DNA intermediate called adenylate. Subsequently, the rejoining of DNA termini via the 3′-OH and the 5′-phosphate at the break by transesterification releases rAMP. If this last

ligation step cannot be executed due to “dirty breaks”, the adenylate intermediate remains on the 5-end of the break. This abortive ligation product cannot be ligated by Lig3 but requires APTX to remove the adenylate (Figure 1.5 B) prior to a new ligation attempt [2, 80]. Hence, APTX serves as a *unique* proofreading or recovery enzyme in base excision and single strand break repair. In addition, there is some evidence, that APTX potentially also operates during the repair of double strand breaks in the same manner by removing adenylate structures arising from abortive Lig4 ligation events there [21]. APTX is up to now the only known enzyme which is capable of removing adenylates from DNA. Consistent with this, mutations in the *APTX* gene, which cause loss or functional deficiencies of APTX [2, 42, 45, 51, 79, 86] manifest themselves in reduced SSB repair *in vivo* [42, 46]. Since alternative repair pathways such as the “long patch” repair are underrepresented in postmitotic cells, *e.g.* in neurons, the loss of Purkinje cells and of myelinated nerve fibres, which are clinical hallmarks of AOA1 [59, 75], may well be the pathophysiological consequences of unrepaired SSB damage. Interestingly, AOA1 patients lack predisposition for cancer in contrast to *e.g.* Ataxia telangiectasia [58, 98], which is caused by defects in double strand break repair. Furthermore, lower levels of APTX are a positive prognostic marker in the context of chemotherapy in the therapy of colon cancer and acute leukemia with agents promoting DNA damage [28, 49]. This observation can be taken as evidence for APTX function in DNA repair pathways in human beings and renders APTX a potential “auxiliary drug target” in chemotherapy. However, contrary to the situation in humans, the absence of APTX alone appears as not sufficient to elicit ouvert defects in mice [31], indicating an added layer of complexity for the understanding of APTX’s function(s).

1.4 Research outline

The initial goal of the project was the determination of the three dimensional (3D) structure of the HIT-ZnF part of human APTX in order to provide mechanistic insight into its interaction with the adenylate substrate and a structural rational for the deadenylation reaction mechanism proposed on the basis of biochemical data [80]. The prerequisite for structure determination *via* liquid-sate NMR, besides labeling with stable isotopes, is that the purified protein in question has to be stable at relatively high concentration ($\approx$1 mM, corresponding to 20 mg/ml for a 20 kDa protein) for at least several days at room temperature. However, all APTX constructs tested, tended to precipitate at already moderate concentrations, essentially excluding the route *via* standard NMR procedures. This intrinsic behavior turned out to be the major *technical* challenge of the project. Alternative approaches were chosen to extract structural information *via* different and complementing techniques utilizing circular dichroism spectroscopy (CD), as well as solid-state and liquid-state NMR spectroscopy (Figure 1.6, left side). CD is a fast and well established label-free method for structural characterization requiring only a small amount of purified protein. However, the information obtained is limited to the secondary structure, in particular the overall content of α-helix, β-sheet and random coil. To analyze CD data employing the largest publicly available reference data set of proteins with known 3D structure, a novel web server based tool was developed (***Publication 1***), providing a reliable estimate of secondary structure content. The second approach to deal with the precipitation propensity of APTX was the use of solid-state NMR. In contrast to liquid-state NMR, solid-state NMR depends on the availability of a microcrystalline[2] sample. Once (micro-)crystallization conditions are identified and a suitable NMR sample is obtained, neither the propensity to precipitate nor the size of the protein is in principle a restricting factor of this method. However, the experimental lifetime of these microcrystals is often limited due to technical issues (*e.g.* sample heating) [30, 95]. One approach to minimize the exposure time of the sample to these method-related "stress factors" constitutes the *simultaneous* acquisition of multiple NMR data sets, that are required for structural studies of proteins. Several crystallization conditions[3] were tested to obtain HIT-ZnF microcrystals, but none of them led to a satisfactory sample quality preventing further investigations *via* solid-state NMR[4]. Although

[2] In contrast to *e.g* X-ray crystallography, where a regular and sufficiently large *single* crystal is needed for diffraction, ssNMR a great number of small (micro-)crystals ($\approx$5-10 mg) are utilized.

[3] Crystallization conditions included (1) 1.7 M ammonium sulfate [37], (2) 50% (v/v) 2-methyl-2,4-pentanediol and 25% (v/v) 2-propanol [34], (3) 0.2 M sodium-potassium tartrate and 20% (w/v) polyethylene glycol 3350 [100], (4) 30% (v/v) polyethylene glycol 8000 [65].

[4] In contrast to the situation in lsNMR, there is no generally applicable "standard procedure" for structure determination via ssNMR, for which rather different approaches have to be tested [55].

not applicable to APTX, the possibility to collect different NMR data sets at the same time were demonstrated using the protein domains GB1 and SH3 as two well-characterized model systems (***Publication 2***), which were used to develop and implement this method. Facing the apparent limitation that APTX precipitates in solution at moderate concentration, finally we focused on highly sensitive liquid-state NMR experiments only. The assignment of resonances is the essential first step and the prerequisite for any following NMR work. This assignment is based on information, which is normally obtained with complex but relatively insensitive and longer-lasting NMR experiments. As these types of experiments were not practicable (due to the relatively fast precipitation of APTX), a novel strategy was used, based upon specific individual isotopic *unlabeling* of all 20 standard amino acids, to achieve the resonance assignment. Following this approach, 78% of the observable protein backbone resonances of the HIT-ZnF domain could be identified and assigned successfully[5] (***Publication 3***). Although in the common route of NMR structure determination, this assignment of resonances would only be the first step, this information *per se* reports on structural characteristics and was directly used to derive a structural model for the HIT-ZnF utilizing the CS-Rosetta approach [56]. Equally important, by knowing which resonance belongs to which atom (nucleus) in which residue, one can now design NMR experiments for site-specific detection and characterization of the interaction with chemical compounds, substrates or other proteins.

The biochemical part (unpublished) addresses the evaluation of APTX's capability to cleave nucleotide-containing substrates. Members of the HIT superfamily are acting on various substrates and, therefore, are linked to a variety of metabolic and cellular pathways [11, 13]. Even though, the particular role of the respective HIT protein within their cellular context is apparently linked to domains adjacent to the HIT domain, the combination of a HIT and a zinc finger motif is found in APTX only. Importantly, only this Hit-ZnF domain is highly conserved among all APTX orthologs and the adjacent FHA domain, found in vertebrates only [80], is dispensable for the catalytic activity of APTX [51, own work]. This prompted us to focus on the HIT-ZnF only, instead of characterizing the full-length protein (for which an even greater tendency to aggregate and precipitate was found). Besides APTX's capability to hydrolyze abortive ligation intermediates arising during SSB repair (see above, section 1.3), it was reported that APTX also interacts with the RNA-binding protein nucleolin [7] and that APTX binds dsRNA and dsDNA with equal affinity [51]. This somewhat promiscuous binding behavior motivated me to analyze APTX's capability to bind and to act upon alternative nucleotide-containing substrates that differ from "canonical" adenylated DNA (Figure

[5]Missing assignments result from non-observability of at least one of the resonances for a given amino acid required for unambiguous data analysis.

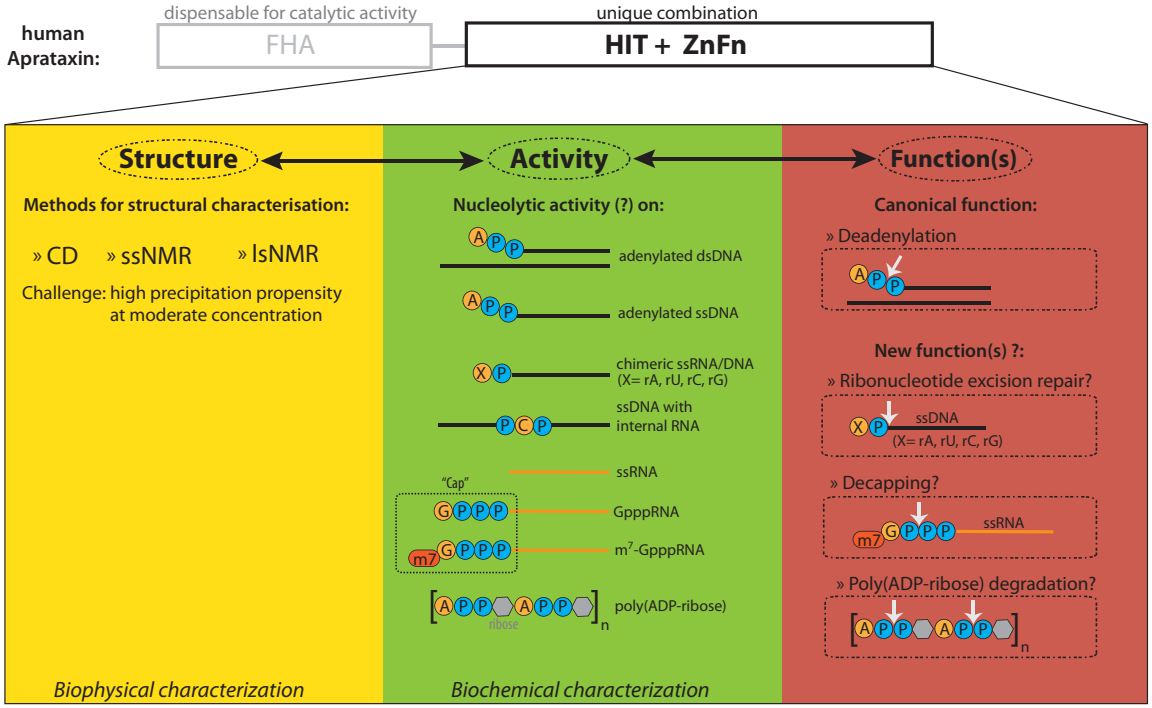


Figure 1.6: Research outline for structural and functional characterization of the HIT-ZnF domain. For the structural characterization a parallel approach utilizing circular dichroism and NMR spectroscopy in solid- and liquid-state was followed. Biochemical characterization of substrate specificity comprised of the design, production and evaluation of adenylated dsDNA (canonical substrate) as well as novel substrates including chimeric RNA/DNA-oligonucleotides, modified RNA and poly(ADP-ribose). CD: circular dichroism spectroscopy; ssNMR: solid-state NMR; lsNMR: liquid-state NMR

1.6, right side). Since most of these substrates were not commercially available (or only at exceedingly high costs), several enzymes including the *Mth* RNA ligase and hPARP1 were recombinantly expressed, purified and the respective reaction for the production of the substrates were established. These efforts finally led to the most interesting discovery of my work (described in more detail in chapter 3): There is now strong experimental evidence, that APTX exhibits unanticipated additional hydrolytic *in vitro* activities, of which the degradation of poly(ADP-ribose) (PAR) appears by far the most interesting one. PAR-ylation of target proteins mediated by PARP1 is an essential signal for many cellular processes [15, 36] including the "marking" of DNA damage sites by PARP1 to attract repair proteins. These findings provide a basis to address the potentially more global role of APTX in DNA damage signaling and DNA repair pathways in the future and might eventually also deliver more insight into the pathological consequences of APTX deficiency.

2 Publications

2.1 Publication 1

Title:	CAPITO - A web server based analysis and plotting tool for circular dichroism data
Authors:	C.Wiedemann, **P.Bellstedt**, M.Görlach
Contributions:	CW conceived of the concept. PB and CW programmed the scripts and wrote the manuscript. PB programmed the scripts for a web service. CW did the experiments. MG supervised the work and corrected the manuscript.
Status:	Accepted for publication in *Bioinformatics*: 13.05.2013 doi: 10.1093/bioinformatics/btt278
Summary:	This manuscript describes an novel web server-based tool for the analysis of circular dichroism (CD) data, in particular with respect to the prediction of secondary structure content. We utilized the largest publicly available reference data set comprising of CD data of proteins with known three dimensional structure to implement different methods for the evaluation of a given CD spectrum, thereby providing an reliable estimate of protein secondary structure content.

Bioinformatics Advance Access published June 7, 2013

BIOINFORMATICS ORIGINAL PAPER

2013, pages 1–8
doi:10.1093/bioinformatics/btt278

Structural bioinformatics

Advance Access publication May 15, 2013

CAPITO—a web server-based analysis and plotting tool for circular dichroism data

Christoph Wiedemann*, Peter Bellstedt and Matthias Görlach*

Biomolecular NMR Spectroscopy, Leibniz Institute for Age Research—Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Circular dichroism (CD) spectroscopy is one of the most versatile tools to study protein folding and to validate the proper fold of purified proteins. Here, we aim to provide a readily accessible, user-friendly and platform-independent tool capable of analysing multiple CD datasets of virtually any format and returning results as high-quality graphical output to the user.

Results: CAPITO (CD Anaylsis and Plotting Tool) is a novel web server-based tool for analysing and plotting CD data. It allows reliable estimation of secondary structure content utilizing different approaches. CAPITO accepts multiple CD datasets and, hence, is well suited for a wide application range such as the analysis of temperature or pH-dependent (un)folding and the comparison of mutants.

Availability: http://capito.nmr.fli-leibniz.de.

Contact: cwiede@fli-leibniz.de or mago@fli-leibniz.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 20, 2012; revised on May 10, 2013; accepted on May 13, 2013

1 INTRODUCTION

The past 20 years have witnessed a dramatic growth of the number of high-resolution protein structures deposited in the protein data bank (PDB: Berman *et al.*, 2000). The progress in structural biology has been driven by developments in recombinant protein expression technology, as well as by advances in methodology, data analysis and bioinformatics. *Escherichia coli* is, so far, the most widely used host for structural studies, which in turn require significant amounts of recombinant protein. Before resource-intensive detailed structural and functional studies, it is extremely helpful, if not essential, to validate the proper fold of purified recombinant proteins and one of the most versatile tools to study protein fold(ing) constitutes circular dichroism (CD) spectroscopy. Compared with X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, the structural information obtained from CD is limited. However, CD spectroscopy carries a number of advantages: it is a well-established label-free technique requiring comparably small amounts of material, and a short time is necessary for assessing structural parameters of proteins like secondary structure, conformational changes, (un)folding and interactions (Whitmore *et al.*, 2010).

*To whom correspondence should be addressed.

A broad range of mathematical methods have been devised to extract structural information from CD spectra to provide for an estimate of the secondary structure composition of proteins via multilinear regression (Greenfield and Fasman, 1969), singular value decomposition (Hennessey and Johnson, 1981), ridge regression (Provencher and Glöckner, 1981), principal component factor analysis (Pribič, 1994), convex constraint analysis (Perczel *et al.*, 1991), neural network-based analysis (Andrade *et al.*, 1993; Böhm *et al.*, 1992) and the self consistent method (Sreerama and Woody, 1993), respectively. All these methods are based on the assumption that the CD spectrum of a given protein represents a linear combination of basis spectra. Different secondary structural elements give rise to bands characteristic in wavelength and intensity (Raussens *et al.*, 2003).

$$[\Theta]_\lambda = \sum f_n S_{\lambda n} + noise \quad (1)$$

The CD spectrum of a given protein can be represented by the molar ellipticity $[\Theta]_\lambda$ as a function of wavelength λ, where f_n is the fraction of each secondary structure n, and $S_{\lambda n}$ is the ellipticity at each wavelength of each *nth* secondary structural element (Greenfield, 2006). The sum of all fractional weights $\sum f_n$ is equal to 1 in constrained fits.

The quality of the output of the aforementioned methods relies on the availability of a reference database of CD spectra of proteins whose 3D structure is known (Lees *et al.*, 2006). With the advent of the Protein Circular Dichroism Data Bank (PCDDB), a public repository for far ultraviolet (far-UV) and synchrotron radiation CD spectral data and their associated experimental metadata, the number of publicly available CD spectra increased enormously (Lees *et al.*, 2006; Wallace *et al.*, 2006; Whitmore *et al.*, 2011). In the PCDDB, each entry contains sequence and experimental information for the respective protein and includes the PDB code for proteins of which 3D structures are available.

During the recent past, different web services (Louis Jeune *et al.*, 2012; Raussens *et al.*, 2003; Whitmore and Wallace, 2004, 2008) or programmes (Böhm *et al.*, 1992; Johnson, 1999) for analysing of CD data and estimating secondary structure content became available. Recently, Janes and co-workers (Klose *et al.*, 2012) launched the tool DichroMatch for matching spectra against reference data. Here, we describe a novel web server-based tool combining different methods for estimating secondary structure content and analysing far-UV CD data based on a selected set of far-UV CD data as available from the PCDDB.

2 METHODS

2.1 CD Data collection and processing

Lysozyme (chicken), cytochrome C (horse), β-amylase (sweet potato) and carbonic anhydrase II (bovine) were purchased from Sigma Aldrich at the highest purity available. Recombinant ubiquitin (human) was provided by Thomas Seiboth (FLI Jena) and the β1 immunoglobulin-binding domain of protein G (GB1) was expressed and purified as described (Bellstedt *et al.*, 2012). All proteins were dissolved in 50 mM borate (cytochrome C, β-amylase and carbonic anhydrase II), pH 7.5, or exchanged into pure water (lysozyme, ubiquitin and GB1) using NAP-5 columns (GE Healthcare). The protein concentrations were in the 10 µM range and verified spectrophotometrically at 280 nm with extinction coefficients calculated using ProtParam (http://web.expasy.org/protparam/). CD spectra were collected on a JASCO J-710 CD spectropolarimeter at 4°C in a 1 mm quartz cuvette. The instrument was calibrated with D-10-camphorsulphonic acid. Each CD spectrum represents the average of 10 accumulated scans at 100 nm/min with a 1 nm slit width and a time constant of 1 s for a nominal resolution of 1.7 nm. Data were collected between 185 and 260 nm with the appropriate buffer and solvent background subtraction. No further zeroing was applied (after background subtraction) because none of the six proteins we used for the experimental part exhibited, for our chosen 10 µM concentration range, a CD between 260 and 320 nm—as tested in preliminary experiments. We tested a whole range of scan rates and time constants and did not notice significant changes in CD and the outcome of the CAPITO analysis.

2.2 Reference datasets

For this study, we used the PCDDB dataset of October 2012 as a well-calibrated, wide wavelength range reference dataset containing a large number of proteins, which effectively cover a large combination of secondary structures and fold space (Lees *et al.*, 2006). That database does not include structures of oligopeptides. From this dataset, only entries linked to an existing PDB code were selected. For multiple entries referring to the same PDB code, only the spectral data recorded at the lowest temperature were used. Our selected dataset contains 107 entries (Supplementary Table S1). Note that for each PCDDB entry, the values for α, 3_{10} and π-helix are summarized as helical (h), β-strand (b) also includes β-bridge and bonded turn, bend, loop and irregular are combined as irregular (i), respectively.

In addition, as reference for significantly unfolded and pre-molten globule states, 95 datasets containing the CD values for $\lambda = 200$ and 222 nm were used as published (Uversky, 2002).

2.3 CAPITO input

Spectral data in millidegrees or mean residue CD extinction coefficient ($\Delta\epsilon$) or mean residue ellipticity ([Θ]), respectively, can be submitted in different data formats as text (txt) file: AVIV 60DS, Aviv, Aviv CDS, BP (Wallace and Teeters, 1987; Whitmore and Wallace, 2004) and Jasco. Example files are available through the CAPITO web page. The user also has the possibility to manually enter or copy/paste spectral data, where wavelength and CD data are separated by a blank or a tab stop with one wavelength per row. In addition, it is possible to upload CD data collected with either smaller or larger step size than 1 nm. The default input data dimension is in millidegrees. Following input of additional experimental parameters such as protein concentration, cuvette pathlength and the number of amino acids, millidegrees are converted to either mean residue CD extinction coefficient ($\Delta\epsilon$ in M^{-1} cm^{-1}) or mean residue ellipticity ([Θ] in deg cm^2 $dmol^{-1}$). Optionally, the amino acid sequence can be submitted as one-letter code for prediction of secondary structure using an implemented Chou-Fasman-algorithm (Chou and Fasman, 1978).

2.4 CAPITO output

CAPITO provides for the spectral data converted into either $\Delta\epsilon$ or [Θ] as a graph (for review see Greenfield, 2006; Kelly *et al.*, 2005; Sreerama and Woody, 2004). In addition, the spectral values at 200 versus 222 nm are plotted for an estimate of the folding state of the protein in question. The prediction of the secondary structure elements is realised via extraction of information from a calculated set of basis spectra and a matching-based approach as described later in the text. Of all CD spectra in our reference dataset, the three curves best matching the submitted query are plotted as well. All graphs can be downloaded as high-quality portable network graphic (png) files.

3 RESULTS AND DISCUSSION

One of the most widely used applications of CD is the estimation of protein secondary structure content from far-UV CD spectra. Not only the relative proportion of secondary structure (e.g. helical, β-strand and others) provides for a characteristic contribution to the far-UV CD spectrum of a protein but also aromatic and sulphur-containing side chains, the length of α-helices and the twist in β-sheets (Johnson, 1999). A large reference dataset is necessary to cover all these features for analysis. In principle, the number of CD spectra in a reference dataset defines the number of structural features that can be determined. Based on considerations of Hennessey and Johnson (1981) and Johnson (1992, 1999), the number of different secondary structural elements significantly depends on the shortest wavelength used in a CD spectrum. For example, a lower spectral limit set to 190 nm reduces information content so that three to four different structural elements can be safely predicted. As using lower wavelengths might be impractical, in particular for biochemists, we restrict the evaluation of the CD data entered into CAPITO to three structural elements: the combination of α, 3_{10} and π-helix as helical content (h), β-strand (b) also includes β-bridge and bonded turn, bend and loop are included in the structural feature irregular (i), respectively.

3.1 Reference dataset derived basis spectra

The optical activity of individual secondary structural elements is assumed to be additive and can be expressed as given in Equation (2). At any particular wavelength λ, the sum of $f's$ is equal to 1 and all $f's \geq 0$ for a constrained approach.

$$[\Theta]_\lambda = f_h[\Theta]_{h,\lambda} + f_b[\Theta]_{b,\lambda} + f_i[\Theta]_{i,\lambda} \quad (2)$$

If the relative proportion for the secondary structural elements is known (from an X-ray or NMR spectroscopy-based protein structure) and the corresponding CD spectrum is at hand, it is possible to calculate the ellipticity for any given wavelength within the range of the CD spectrum. The least-square method was used for solving the $f's$ from a system of equations for our reference dataset. Solving the matrix [Equation (3)] by least-square fitting for each selected protein j in the reference dataset, a calculated $[\Theta]_{h,b,i}$ for each secondary structure element, is returned over the wavelength range of 180–240 nm (Supplementary Table S2). [Θ] for each secondary structure element is plotted in Figure 1 against the wavelength.

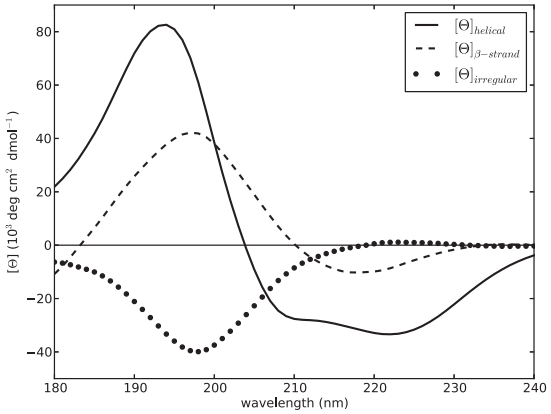


Fig. 1. Calculated CD spectra for a content of 100% helix ($[\Theta]_{helical}$), β-strand ($[\Theta]_{\beta-strand}$) and irregular ($[\Theta]_{irregular}$), respectively, based on the solution of the matrix [Equation (3)] and using all 107 proteins in the reference dataset (Supplementary Table S1)

$$\begin{pmatrix} [\Theta]_{180} \\ [\Theta]_{181} \\ [\Theta]_{182} \\ . \\ . \\ . \\ [\Theta]_{240} \end{pmatrix}_{[j,n]} = \begin{pmatrix} [\Theta]_h \\ [\Theta]_b \\ [\Theta]_i \end{pmatrix}_{\lambda} \times \begin{pmatrix} f_{h,j} & f_{b,j} & f_{i,j} \\ f_{h,j+1} & f_{b,j+1} & f_{i,j+1} \\ f_{h,j+2} & f_{b,j+2} & f_{i,j+2} \\ & . & \\ & . & \\ & . & \\ f_{h,n} & f_{b,n} & f_{i,n} \end{pmatrix} \quad (3)$$

With this approach, we derived three basis CD spectra (helical, β-strand and irregular) using our reference dataset. For the helical basis spectrum, the presence of one positive band at 194 nm and two major negative bands at 210 and 222 nm are most evident. This is consistent with basis CD spectra derived by others (Brahms and Brahms, 1980; Chen and Yang, 1971; Chen *et al*., 1974, 1972; Hennessey and Johnson, 1981; Perczel *et al*., 1992b; Reed and Reed, 1997; Sreerama and Woody, 1993; Toumadje *et al*., 1992). The β-strand basis spectrum is characterized by a less intensive positive band at 197 nm and a negative band at 217 nm. A negative band at 197 nm and a weak positive band at 223 nm are features of the irregular basis spectrum. The irregular structure closely resembles that of polypeptides in extended poly-L-proline II-like structures (Woody, 1992). Also the irregular structure fits well the CD curve for poly(Pro-Lys-Leu-Lys-Leu)$_n$ in salt-free solution as model compound for unordered conformation (Brahms and Brahms, 1980). A plot overlay of different basis spectra in comparison with those calculated in this work are given in the Supplementary Data (Supplementary Figs S1–S3).

In a simple model, the ellipticity of one secondary structure content is related to certain wavelengths. For identifying wavelengths best representing secondary structure content, a constrained approach was performed. With the three basis spectra ($[\Theta]_{h,b,i}$) as derived from solving the aforementioned matrix [Equation (3)], every possible combination of $f's$ [Equation (4)] with the boundary condition [Equation (5)] was calculated using Equation (2).

$$f_{h,b,i} = 0.00, 0.01, 0.02, \ldots, 1.00 \quad (4)$$

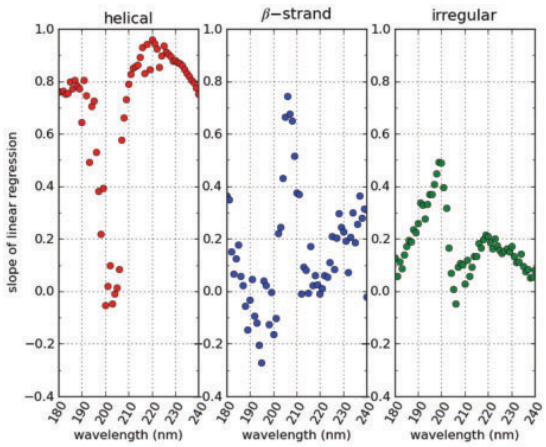


Fig. 2. For all of the proteins in the reference dataset, the secondary structure content was calculated, and the resulting slope of the linear regression of the auto-correlation—as a measure of accuracy of the prediction—is plotted against the respective wavelength

$$f_h + f_b + f_i = 1 \quad (5)$$

Subsequently, the minimum of the square difference between the calculated $[\Theta]_{cal}$ and the observed molar ellipticity $[\Theta]_{obs}$ in the reference dataset was determined. This was carried out for all proteins of the reference dataset for each wavelength in the range of 180–240 nm. For each wavelength, the best matching combination of $f's$ was plotted against the experimental f values to assess the correlation between calculated and experimental data points, and a linear regression analysis was performed (Figs 2 and 3).

For a perfect correlation, the actual slope of the linear regression of the auto-correlation would be one, and the closer the slope is approaching unity, the higher the quality of the prediction of secondary structure elements for a given wavelength.

The results of the linear regression analysis are plotted against the wavelength. Maxima for the prediction of secondary structure contents are found for helical at 220 nm, for β-strand at 206 nm and for irregular at 199 nm, respectively (Fig. 2). To extract information about the secondary structure content from a CD curve, the three basis spectra ($[\Theta]_{h,b,i}$) as derived from matrix [Equation (3)] were used in Equation (2) at defined wavelengths as follows. For prediction of the helical content, the calculated molar ellipticity of a pure helix $[\Theta]_h$ at 220 nm was used. By varying the f_h, f_b and f_i [Equation (4)], with the constraint that the sum of the $f's$ is equal to one [Equation (5)], the difference between the calculated and the observed $[\Theta]_{220nm}$ of a query protein was minimized. The f_h resulting from this solution represents the helical content of the respective query protein. The whole procedure was then independently repeated at 206 nm, and the resulting f_b at this wavelength is considered as the predicted β-strand content. The irregular content (f_i) was calculated at 199 nm in the same manner. In summary, for the prediction of the content of each of the three secondary structure elements, Equation (2) is used. Based on the maxima found in the linear regression analysis aforementioned (Fig. 2), as a measure for the accuracy of prediction, the programme extracts helical content at

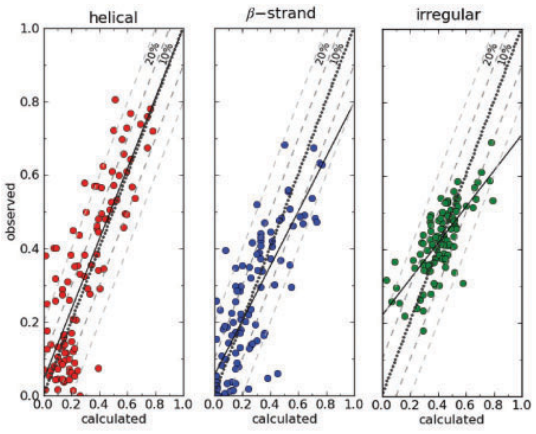


Fig. 3. Cross-validation of the basis spectra method: the secondary structure content for each of the proteins in the reference dataset was calculated and plotted against the secondary structure content available in the PCDDB. Helical content was obtained at 220 nm, β-strand at 206 nm and irregular at 199 nm. The hatched line indicates 10 and 20% deviation from the ideal auto-correlation (dotted line). Solid line represents the linear regression of the calculated auto-correlation

220 nm, β-strand at 206 nm and irregular at 199 nm only (neglecting the respective contributions of the other two elements at those wavelengths). Hence, the sum for the three secondary structural elements can differ from 100%. Based on the calculated secondary structural content, optionally a theoretical CD curve can be back calculated and plotted against the query curve, and the goodness of fit between the two curves is presented as normalized root-mean-square deviation (NRMSD) (Supplementary Equation S1) to the user, where an ideal fit would approach an NRMSD of zero (Mao *et al.*, 1982). To evaluate the quality of the model, a cross-validation was carried out. For each reference protein in the dataset, the secondary structure content was calculated as described earlier in the text and compared with deposited data within the PCDDB. This prediction was repeated for all spectra within our reference dataset. The calculated secondary structure contents are plotted against the actual values (Fig. 3). From this, it becomes evident that this procedure delivers a reliable estimate of helical, β-strand and irregular secondary structure content for a given protein.

As illustrated in Figure 3, the cross-validation provides the most accurate results for helical secondary structure. Only 7 of 107 validated proteins are beyond the margin of 20% deviation. The prediction of β-strand secondary structure is somewhat less accurate than for helical content and shows that eight proteins of our reference dataset are not within the range of 20% deviation. Irregular structures (neither helix nor β-strand) adopt a wide range of backbone angles exhibiting CD spectra of heterogeneous character. This hampers the accurate estimation of such secondary structure elements. With our approach, only four proteins of our reference dataset are not in the range of 20% deviation for the prediction of irregular structures. The irregular structure content in our reference data is not spread over the range observed for helical or β-strand content. In such a situation, outliers have a more significant influence on the linear regression, resulting in a slope further away from approaching unity.

This approach appears somewhat less accurate in comparison with other methods [SELCON3 (Sreerama and Woody, 1993), CDSSTR (Johnson, 1999) or CONTIN (Provencher and Glöckner, 1981), Table 1], but the results and their validation strongly suggest that the input of measured values at three wavelengths (220 nm for helical, 206 nm for β-strand and 199 nm for irregular) are sufficient to describe the information contained within a given CD spectrum.

3.2 Matching-based prediction: nearest-neighbour and area difference method

Under the general assumption that proteins with similar secondary structure content give rise to comparable CD curves, we tested alternative prediction methods. If the reference dataset effectively covers a large combination of secondary structures and fold space, matched reference proteins should comprise the actual secondary structure content of a query protein. To test this hypothesis, we used two methods to evaluate a query CD spectrum against our reference dataset. For either method, the query CD spectrum is compared with each CD spectrum of the reference dataset. As standard pattern recognition method, we used a k-nearest-neighbour algorithm (Cover, 1968; Kowalski and Bender, 1972) in the following two approaches.

For the nearest-neighbour approach, for each wavelength within the range of 180–240 nm, the 25 best matching reference curves defined by closest proximity were determined. Here, proximity is defined as the 1D distance between the query and the reference CD spectra at each wavelength. Subsequently, the frequency (N) of a given reference protein among the 25 nearest neighbours to the query protein is used for ranking.

To assess the lowest area difference (AD), the best matching reference curves for the range of 180–240 nm were selected. AD was defined as shown in Equation (6).

$$\mathrm{AD}_i = \sum_{\lambda} \sqrt{\left(([\Theta]_{Q,\lambda} - [\Theta]_{Ref,\lambda,i}) \times 1\,nm\right)^2} \qquad (6)$$

AD (in deg cm^2 dmol^{-1} nm) represents the CD curve area difference between query (Q) and the reference (Ref) protein, where $[\Theta]_{Q,\lambda}$ is the molar ellipticity of the query protein at wavelength λ, and $[\Theta]_{Ref,\lambda}$ is the molar ellipticity of the protein i in the reference dataset at the same wavelength λ, respectively. In extension of the rotational strength approach presented by Klose *et al.* (2012), the AD approach here evaluates in 1 nm steps over the wavelength range λ the difference in area between a reference and a query curve rather than the area of a given CD curve per se. It, hence, also includes an (although indirect) evaluation of the shape of the query CD curve. The AD output provides the best ranked CD reference curves with the smallest area difference. The NRMSD (Supplementary Equation S1) between the best matching reference CD spectra and the query CD spectra is calculated and shown for comparison. AD and NRMSD are always calculated over the same wavelength range as provided by the query dataset. However, the NRMSD is neither used for matching nor for ranking of identified hits.

To validate our approach for extracting structural information by matching query and reference CD spectra an auto-correlation

Table 1. Estimation of structural contents in comparison

Protein	SELCON3	CDSSTR	CONTINLL	CDNN2.1[a]	Raussens[b]	K2D3[c]	CAPITO[d]	PDB[e]
Ubiquitin								
h	15,5	17,6	26,3	14,3	29,7	10,1	11 (4–25)	25
b	27,6	28,1	20	42,9	16,3	28,6	30 (34–45)	34
i	52,8	53,7	53,7	48,2	47,6	61,3	52 (41–54)	41
Lysozyme								
h	41,4	41,8	41,9	34	30,9	32,6	25 (31–50)	40
b	9,1	10,4	8,9	12,9	16,3	17,5	14 (4–21)	10
i	49,4	47,4	49,2	55,8	46,8	49,3	51 (46–49)	50
Cytochrome C								
h	41,3	43,5	44	31,6	24,7	26,3	31 (13–48)	40
b	5,2	10,1	3,1	12,7	19,2	23,4	25 (2–34)	1
i	50,6	47,8	52,9	58,2	46,3	50,3	46 (49–53)	59
β-Amylase								
h	36,3	40,9	36,6	31,8	25	29,4	28 (31–50)	38
b	14,1	12,3	11,3	12,1	22,7	17,9	15 (4–21)	13
i	51,8	47,2	52	62,3	47,5	52,7	53 (46–48)	49
CA-II								
h	9,6	9,6	7,6	10,3	0,9	2,5	0 (12–16)	15
b	35,8	41,8	32,3	32,6	34,6	37,9	37 (30–37)	30
i	55,9	52,4	60	54,9	51,5	59,6	52 (48–54)	55
GB1								
h	42,9	45,4	42,5	40,3	40	34,1	39 (29–44)	25
b	13,6	13,5	15,2	15,4	13,4	19,8	13 (12–30)	42
i	43,1	41,3	42,3	45	42,7	46,1	37 (39–54)	33

Note: The CD spectra of the indicated proteins were recorded and processed. The resulting data are analysed with different programmes. Helical content is represented as h, β-strand content as b and irregular as i. SELCON3 (Sreerama and Woody, 1993), CDSSTR (Johnson, 1999) and CONTINLL (variant of CONTIN (Provencher and Glöckner, 1981) are provided in the CDPro software package (Sreerama and Woody, 2000).
[a]Böhm *et al.* (1992), [b]Raussens *et al.* (2003), [c]Louis Jeune *et al.* (2012). [d]For CAPITO, the result on the basis of spectra-based method is given. In parenthesis, the range of the three best hits is provided based on the area difference method. [e]Secondary structure content obtained through the PDB website using the PDB ID 1UBI (ubiquitin), 193L (lysozyme), 1HRC (cytochrome C), 1FA2 (β-amylase), 1V9E (carbonic angydrsae II, CA-II) and 2LGI (β1 immunoglobulin-binding domain of protein G, GB1), respectively. For K2D3 and PDB, the helical and β-strand contents were subtracted from 100 to calculate the irregular content. All values are given as percentage. *Of note*: SELCON, CDSSTR, CONTINLL and K2D2 are used by DichroWeb (Whitmore and Wallace, 2004). Here, we have used the most recent versions of these packages for comparing their results with the results returned by CAPITO.

was performed. For this purpose, single-reference spectra were removed in turn from the reference dataset, and matching was performed with the remaining spectra against the structure content as derived from the removed spectrum. The structural contents of the three best matching reference proteins are combined to give a range for f_h, f_b and f_i. In this approach, the secondary structure content of a protein in question is not defined by single calculated value but by a range of secondary structural content derived from the three best matching reference proteins. This prediction was repeated for all the spectra within the reference dataset.

Figure 4 depicts the cross-validation for the structural content predicted by the nearest-neighbour method (a) and by the area difference method (b), respectively. Both methods deliver a reliable and comparable estimate of helical, β-strand and irregular content for a given protein. Only a few outliers in the prediction of helical and β-strand content are not within the margin of 20% deviation. Surprisingly, for irregular content, all validated proteins fall within the boundary of 20% deviation. The range of secondary structural content present in the three best matches covers in most cases the actual content of the query protein. A consequence of this is that an increasing number of protein structures in the reference database will further restrict the range covered by the three best matches and, hence, lead to an increased accuracy of the secondary structure content for a given protein as predicted by CAPITO.

3.3 Validation by protein identification and comparison with other programmes

Whitmore *et al.* (2011) mentioned the idea of identifying proteins based on their CD spectral characteristics, i.e. if a protein is deposited in a database, it should be possible to identify this protein based on its CD curve. Here, we recorded CD spectra of freshly prepared solutions of six different proteins (see Section 2) and processed the recorded CD data with the CAPITO web server. Our reference dataset contains five (β-amylase, carbonic anhydrase II, cytochrome C, lysozyme and ubiquitin) of six of the selected proteins used here. As seen in Figure 5a (and also Supplementary Fig. S4), our matching-based prediction allows for the identification of proteins present in a reference dataset—even under buffer conditions slightly different from the ones in the reference dataset. All tested proteins were identified as indicated by the best score (Fig. 5a and Supplementary Fig. S4).

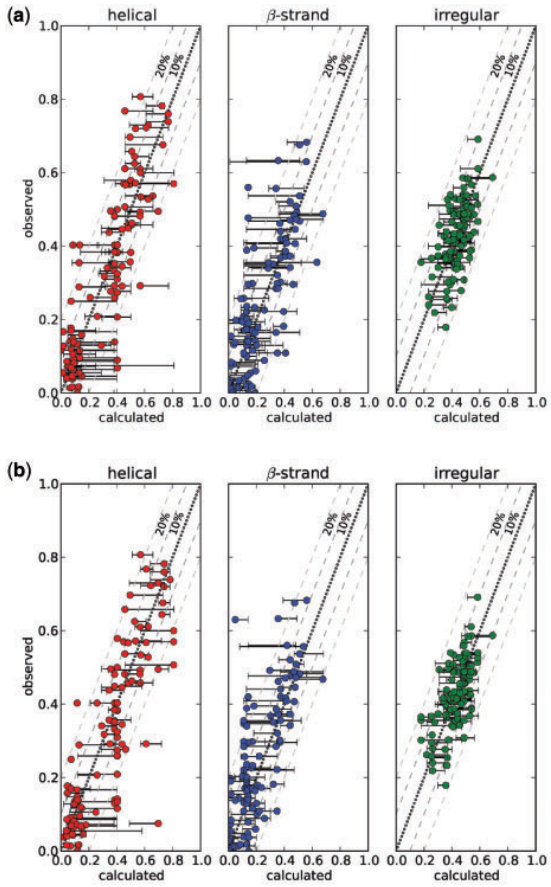


Fig. 4. Cross-validation of the nearest-neighbour method (**a**) and area difference method (**b**). The secondary structure content for each of the proteins in the reference dataset was calculated and plotted against its actual secondary structure content. The top-ranked hit is plotted as dot, error bars connect second- and third-ranked hit. The hatched line indicates 10 and 20% deviation from the ideal auto-correlation (dotted line)

To compare our method of secondary structure prediction, we used our recorded and processed spectra to a selection of other available programmes and web services (Böhm *et al*., 1992; Louis Jeune *et al*., 2012; Raussens *et al*., 2003; Sreerama and Woody, 2000). The results (Table 1) were compared with secondary structure assignments deposited in the Protein Data Bank, which in turn are based on the DSSP programme (Kabsch and Sander, 1983). A standard set of reference CD data was tested to compare the accuracy of CAPITO with other available programmes. The standard set consisted of 16 proteins and poly-L-glutamic acid (Sreerama and Woody, 1993). CAPITO returns a good correlation coefficient for the helical and β-strand secondary structure elements found in the X-ray structure of the 16 test proteins (Table 2). As most programmes that rely on a protein dataset, CAPITO is not suitable for evaluating helical and β-strand conformation content for long homopolymeric peptides such as poly-L-lysine, as our reference dataset does not include such homopolymers.

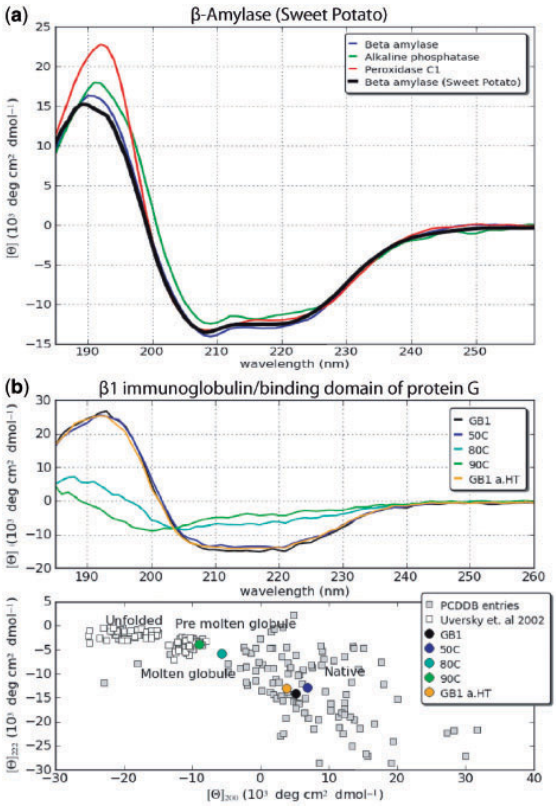


Fig. 5. Experimental validation: (**a**) a CD spectrum of β-amylase (sweet potato) was recorded and analysed with CAPITO (black curve). As an example, the graphical output for the area difference method is shown. The best-matching CD spectra of the reference dataset are β-amylase (blue), the second best alkaline phosphatase (green) and the third best hit peroxidase C1 (red), respectively. (**b**) Recorded CD spectra for GB1 at different temperatures. Upper panel: CD curve for the starting temperature of 4°C is depicted in black, other temperatures as indicated in the inset. CD spectrum of GB1 after heat treatment and cooled down to 4°C (GB1 a.HT) is shown in orange. Lower panel: CD values at $\lambda = 200$ nm plotted versus the values at $\lambda = 222$ nm to deduce the folding state of GB1 at the respective temperatures

3.4 Monitoring protein folding under different conditions

Different folding states of polypeptide chains are characterized by specific shapes of their far-UV CD spectrum. For example, unfolded polypeptides or proteins containing mainly irregular structural elements show a spectral minimum in the vicinity of 200 nm and an ellipticity close to zero in the vicinity of 222 nm. Helical proteins show a characteristic double minimum at 222 and 208 nm and an intensive maximum near 195 nm (Fig. 1). The 'double wavelength' plot ($[\Theta]_{222}$ versus $[\Theta]_{200}$) as described (Uversky, 2002) allows the direct visualization of the folding state of a protein for different conditions (e.g. temperature, pH, buffer type and ionic strength). Here, we use this plotting routine within CAPITO to carry out an assessment of the GB1 folding state as a function of temperature. As shown in Figure 5b (lower panel) from the double wavelength plot, it can be concluded that GB1 shows a well folded state at lower temperatures,

Table 2. Comparisons of methods of analysing protein secondary structure content from CD data

Program	Standards	Wavelength	Helical		β-Strand		Irregular	
			P	σ	P	σ	P	σ
Linear regression—unconstrained								
MLR[a]	4 peptides	178–240	0.91	0.13	0.43	0.21	0.07	0.16
MLR	4 peptides	200–240	0.92	0.14	0.74	0.16	0.23	0.16
Linear regression—constrained								
G&F[b]	poly-L-lysine	208–240	0.92	0.13	0.61	0.18	ND	ND
LINCOMB[c]	4 peptides	178–240	0.93	0.11	0.58	0.15	0.61	0.11
LINCOMB	17 proteins	178–240	0.94	0.09	0.62	0.14	0.21	0.13
Singular value decomposition								
SVD[d]	17 proteins	178–240	0.98	0.05	0.68	0.12	0.22	0.10
Convex constraint algorithm								
CCA[e]	17 proteins	178–260	0.96	0.10	0.62	0.18	0.39	0.18
Ridge regression								
CONTIN[f]	17 proteins	178–260	0.93	0.11	0.56	0.15	0.58	0.08
Variable selection								
VARSLC[g]	17 proteins	178–260	0.97	0.07	0.81	0.10	0.60	0.07
Variable selection—self-consistent method								
SELCON[h]	17 proteins	178–260	0.95	0.09	0.84	0.08	0.77	0.05
SELCON	33 proteins	178–260	0.93	0.09	0.91	0.07	0.53	0.09
Neural network analysis								
CDNN2.1[i]	17 proteins	178–260	0.93	0.10	0.73	0.11	0.82	0.05
K2D2[j]	19 proteins	200–240	0.95	0.09	0.77	0.10	ND	ND
This work								
CAPITO	107 proteins	178–260	0.96	0.11	0.80	0.13	ND	ND

Note: Table 2 lists the correlation coefficient (P) and the mean-square errors (σ) between the calculated and the observed contents of each secondary structure. The table was abstracted from Greenfield (1996).
[a]Perczel *et al.* (1992a), [b]Greenfield and Fasman (1969), [c]Perczel *et al.* (1992a), [d]Hennessey and Johnson (1981), [e]Perczel *et al.* (1991), [f]Provencher and Glöckner (1981), [g]Manavalan and Johnson (1987), [h]Sreerama and Woody (1993), [i]Böhm *et al.*, (1992) and [j]Andrade *et al.* (1993).

which is stable even at higher temperatures. At 80°C, a transition from the native folded state to the molten globule is observed. The temperature shift from 80°C to 90°C changes the GB1 fold from a molten globule towards a pre-molten globule state. This observation is consistent with the previous determined midpoint of denaturation at ~75°C (Alexander *et al.*, 1992; Minor and Kim, 1994). Although, this does not replace a detailed analysis (e.g. melting curve), it allows for a quick and coarse estimation of a transition point enabling analysis of unstable proteins, which may not withstand a time-consuming detailed analysis.

4 CONCLUSION

We have developed CAPITO, a novel web server-based analysis tool for interpreting CD spectra. It allows the simultaneous evaluation of multiple datasets. Hence, it is suitable for the investigation of a protein in question under different conditions (temperature, pH, buffer solvent and mutations). Our approaches (basis spectra and matching-based method) to extract secondary structure information from a CD spectrum take advantage of a recent significant increase in the availability of well-calibrated far-UV CD spectra linked to available tertiary structures. The accuracy of our methods in predicting α-helical, β-strand or irregular content is reliable compared with other frequently used programmes or web services. In summary, we here provide a freely accessible, user-friendly and robust tool for the analysis of CD spectra.

ACKNOWLEDGEMENT

The authors are grateful to G. Peiter and F. Haubensak for help with the IT infrastructure.

Funding: The FLI is a member of Science Association 'Gottfried Wilhelm Leibniz (WGL)' and is financially supported by the Federal Government of Germany and the State of Thuringia. P.B. and C.W. are supported by the Leibniz Graduate School on Ageing and Age-Related Diseases (LGSA).

Conflict of Interest: none declared.

REFERENCES

Alexander,P. *et al.* (1992) Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: why small proteins tend to have high denaturation temperatures. *Biochemistry*, **31**, 3597–3603.

Andrade,M.A. *et al.* (1993) Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Protein Eng.*, **6**, 383–390.

Bellstedt,P. *et al.* (2012) Solid state NMR of proteins at high MAS frequencies: symmetry-based mixing and simultaneous acquisition of chemical shift correlation spectra. *J. Biomol. NMR*, **54**, 325–335.
Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
Böhm,G. *et al.* (1992) Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng.*, **5**, 191–195.
Brahms,S. and Brahms,J. (1980) Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J. Mol. Biol.*, **138**, 149–178.
Chen,Y.H. and Yang,J.T. (1971) A new approach to the calculation of secondary structures of globular proteins by optical rotatory dispersion and circular dichroism. *Biochem. Biophys. Res. Commun.*, **44**, 1285–1291.
Chen,Y.H. *et al.* (1972) Determination of the secondary structures of proteins by circular dichroism and optical rotatory dispersion. *Biochemistry*, **11**, 4120–4131.
Chen,Y.H. *et al.* (1974) Determination of the helix and beta form of proteins in aqueous solution by circular dichroism. *Biochemistry*, **13**, 3350–3359.
Chou,P.Y. and Fasman,G.D. (1978) Empirical predictions of protein conformation. *Annu. Rev. Biochem.*, **47**, 251–276.
Cover,T.M. (1968) Estimation by the nearest neighbor rule. *IEEE Trans. Information Theory*, **14**, 50–55.
Greenfield,N. and Fasman,G.D. (1969) Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, **8**, 4108–4116.
Greenfield,N.J. (1996) Methods to estimate the conformation of proteins and polypeptides from circular dichroism data. *Anal. Biochem.*, **235**, 1–10.
Greenfield,N.J. (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.*, **1**, 2876–2890.
Hennessey,J. Jr and Johnson,W. Jr (1981) Information content in the circular dichroism of proteins. *Biochemistry*, **20**, 1085–1094.
Johnson,W. Jr (1992) Analysis of circular dichroism spectra. *Methods Enzymol.*, **210**, 426–447.
Johnson,W.C. (1999) Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins*, **35**, 307–312.
Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
Kelly,S.M. *et al.* (2005) How to study proteins by circular dichroism. *Biochim. Biophys. Acta*, **1751**, 119–139.
Klose,D.P. *et al.* (2012) DichroMatch: a website for similarity searching of circular dichroism spectra. *Nucleic Acids Res.*, **40**, W547–W552.
Kowalski,B.R. and Bender,C.F. (1972) The K-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal. Chem.*, **44**, 14051411.
Lees,J.G. *et al.* (2006) A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics*, **22**, 1955–1962.
Louis Jeune,C. *et al.* (2012) Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. *Proteins*, **80**, 374–381.
Manavalan,P. and Johnson,W. Jr (1987) Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Anal. Biochem.*, **167**, 76–85.
Mao,D. *et al.* (1982) Folding of the mitochondrial proton adenosinetriphosphatase proteolipid channel in phospholipid vesicles. *Biochemistry*, **21**, 4960–4968.
Minor,D. Jr and Kim,P.S. (1994) Measurement of the beta-sheet-forming propensities of amino acids. *Nature*, **367**, 660–663.
Perczel,A. *et al.* (1991) Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein Eng.*, **4**, 669–679.
Perczel,A. *et al.* (1992a) Analysis of the circular dichroism spectrum of proteins using the convex constraint algorithm: a practical guide. *Anal. Biochem.*, **203**, 83–93.
Perczel,A. *et al.* (1992b) Deconvolution of the circular dichroism spectra of proteins: the circular dichroism spectra of the antiparallel beta-sheet in proteins. *Proteins*, **13**, 57–69.
Pribicc̀,R. (1994) Principal component analysis of Fourier transform infrared and/or circular dichroism spectra of proteins applied in a calibration of protein secondary structure. *Anal. Biochem.*, **223**, 26–34.
Provencher,S.W. and Glöckner,J. (1981) Estimation of globular protein secondary structure from circular dichroism. *Biochemistry*, **20**, 33–37.
Raussens,V. *et al.* (2003) Protein concentration is not an absolute prerequisite for the determination of secondary structure from circular dichroism spectra: a new scaling method. *Anal. Biochem.*, **319**, 114–121.
Reed,J. and Reed,T.A. (1997) A set of constructed type spectra for the practical estimation of peptide secondary structure from circular dichroism. *Anal. Biochem.*, **254**, 36–40.
Sreerama,N. and Woody,R.W. (1993) A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal. Biochem.*, **209**, 32–44.
Sreerama,N. and Woody,R.W. (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal. Biochem.*, **287**, 252–260.
Sreerama,N. and Woody,R.W. (2004) Computation and analysis of protein circular dichroism spectra. *Methods Enzymol.*, **383**, 318–351.
Toumadje,A. *et al.* (1992) Extending CD spectra of proteins to 168 nm improves the analysis for secondary structures. *Anal. Biochem.*, **200**, 321–331.
Uversky,V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739–756.
Wallace,B.A. and Teeters,C.L. (1987) Differential absorption flattening optical effects are significant in the circular dichroism spectra of large membrane fragments. *Biochemistry*, **26**, 65–70.
Wallace,B.A. *et al.* (2006) The protein circular dichroism data bank (PCDDB): a bioinformatics and spectroscopic resource. *Proteins*, **62**, 1–3.
Whitmore,L. and Wallace,B.A. (2004) DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res.*, **32**, W668–W673.
Whitmore,L. and Wallace,B.A. (2008) Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers*, **89**, 392–400.
Whitmore,L. *et al.* (2010) The protein circular dichroism data bank, a Web-based site for access to circular dichroism spectroscopic data. *Structure*, **18**, 1267–1269.
Whitmore,L. *et al.* (2011) PCDDB: the protein circular dichroism data bank, a repository for circular dichroism spectral and metadata. *Nucleic Acids Res.*, **39**, D480–D486.
Woody,R.W. (1992) Circular dichroism and conformation of unordered poly-peptides. *Adv. Biophys. Chem.*, **2**, 31–79.

Supplementary data
CAPITO - A web server based analysis and plotting tool for circular dichroism data

Christoph Wiedemann, Peter Bellstedt, and Matthias Görlach

Biomolecular NMR Spectroscopy
Leibniz Institute for Age Research - Fritz Lipmann Institute
Beutenbergstr. 11, 07745 Jena, Germany

May 6, 2013

PCDDB_id	name	pdb_id	temperature	helical	beta	irregular
CD0000028000	3-dehydroquinate dehydratase	1qfe	4.0	0.4570	0.2030	0.3410
CD0000029000	3-dehydroquinate dehydratase	2dhq	4.0	0.4380	0.1710	0.3900
CD0000100000	Acriflavine resistance protein B	2gif	20.0	0.5070	0.1900	0.3020
CD0000001000	Aldolase	1ado	4.0	0.4580	0.1440	0.3980
CD0000002000	Alkaline phosphatase	1ed9	4.0	0.3100	0.2090	0.4810
CD0000003000	Alpha amylase	1vjs	4.0	0.2770	0.2260	0.4970
CD0000004000	Alpha bungarotoxin	1hc9	4.0	0.0460	0.3680	0.5860
CD0000005000	Alpha chymotrypsin	5cha	4.0	0.1160	0.3470	0.5370
CD0000006000	Alpha chymotrypsinogen	2cga	4.0	0.1340	0.3510	0.5140
CD0000099000	Ammonia channel	2nop	20.0	0.6250	0.0140	0.3610
CD0000007000	Aprotinin	5pti	4.0	0.2070	0.2580	0.5340
CD0000008000	Avidin	1rav	4.0	0.0710	0.4920	0.4370
CD0000101000	Bacteriorhodopsin	1qhj	4.0	0.6970	0.0480	0.2540
CD0000009000	Beta amylase	1fa2	4.0	0.3830	0.1320	0.4840
CD0000010000	Beta galactosidase	1bgl	4.0	0.1390	0.3890	0.4720
CD0000011000	Beta lactoglobulin	1b8e	4.0	0.1420	0.4200	0.4380
CD0000022000	Beta-B2 crystallin	2bb2	4.0	0.0730	0.4410	0.4860
CD0003671000	Beta-crystallin B1	1oki	4.0	0.0860	0.4090	0.5040
CD0003672000	Beta-crystallin B2	1bd7	4.0	0.1000	0.4310	0.4570
CD0003670000	Beta-crystallin B2	1ytq	4.0	0.0690	0.3730	0.5590
CD0003669000	Beta-crystallin S	1a7h	4.0	0.1390	0.4070	0.3380
CD0000012000	c-Phycocyanin	1ha7	4.0	0.7600	0.0000	0.2400
CD0000013000	Calmodulin	1lin	4.0	0.5680	0.0540	0.3780
CD0000014000	Carbonic anhydrase I	1hcb	4.0	0.1770	0.2960	0.5260
CD0000015000	Carbonic anhydrase II	1v9e	4.0	0.1580	0.3020	0.5400
CD0000016000	Carboxypeptidase A1	5cpa	4.0	0.3820	0.1700	0.4500
CD0000017000	Catalase	1dgf	4.0	0.3320	0.1870	0.4820
CD0000018000	Ceruloplasmin	1kcw	4.0	0.1170	0.3560	0.5270
CD0000019000	Citrate synthase	2cts	4.0	0.6110	0.0250	0.3640
CD0000104000	ClC-ec1	1kpk	20.0	0.6440	0.0040	0.3510
CD0000020000	Concanavalin A	1nls	4.0	0.0380	0.4680	0.4940
CD0000105000	cytochrome bc1	1be3	10.0	0.5370	0.0370	0.4270
CD0000021000	Cytochrome C	1hrc	4.0	0.4810	0.0190	0.5000
CD0000106000	cytochrome C oxidase	2dyr	4.0	0.5700	0.0600	0.3700
CD0000030000	Deoxyribonuclease-1	3dni	4.0	0.2890	0.2850	0.4270
CD0000031000	Elastase	3est	4.0	0.0960	0.3210	0.5840
CD0000032000	Ferredoxin	2fdn	4.0	0.1280	0.1820	0.6910

PCDDB_id	name	pdb_id	temperature	helical	beta	irregular
CD0000108000	Ferrichrome-iron receptor	1fcp	4.0	0.0460	0.5370	0.4180
CD0000107000	Ferrienterobactin receptor	1fep	4.0	0.0550	0.4880	0.4570
CD0000023000	Gamma-B crystallin	4gcr	4.0	0.0920	0.4830	0.4250
CD0000024000	Gamma-D crystallin	1elp	4.0	0.0900	0.4770	0.4340
CD0000027000	Gamma-D crystallin	1hk0	4.0	0.0810	0.5090	0.4110
CD0000025000	Gamma-E-crystallin	1m8u	4.0	0.0640	0.4460	0.4920
CD0000026000	Gamma-s-crystallin C terminus	1ha4	4.0	0.1200	0.4710	0.4080
CD0000033000	Glucose oxidase	1cf3	4.0	0.3410	0.2330	0.4250
CD0000034000	Glutamate dehydrogenase I	1hwx	4.0	0.4950	0.1380	0.3670
CD0000035000	Glycogen phosphorylase-b	1gpb	4.0	0.4930	0.1550	0.3530
CD0000036000	Haloalkane dehalogenase	1bn6	4.0	0.4490	0.1940	0.3560
CD0000037000	Hemoglobin	1hda	4.0	0.7680	0.0000	0.2330
CD0000038000	Human serum albumin	1n5u	4.0	0.7200	0.0000	0.2790
CD0000039000	Immunoglobulin G	1igt	4.0	0.0720	0.4620	0.4670
CD0000040000	Insulin	1trz	4.0	0.5790	0.0790	0.3430
CD0000111000	inwardly rectifying k+ channel	1xl4	4.0	0.2740	0.2960	0.4300
CD0000041000	Jacalin	1ku8	4.0	0.0000	0.6330	0.3680
CD0000042000	Lactoferrin	1blf	4.0	0.3180	0.1790	0.5030
CD0000112000	Lactose permease	2cfq	4.0	0.6760	0.0070	0.3160
CD0000115000	Large-conductance mechanosensitive channel	2oar	10.0	0.5280	0.0340	0.4390
CD0000043000	Lectin (lentil)	1les	4.0	0.0430	0.4830	0.4740
CD0000053000	Lectin (pea)	1ofs	4.0	0.0410	0.4730	0.4870
CD0000044000	Leptin	1ax8	4.0	0.6100	0.0140	0.3770
CD0000114000	Light harvesting protein	1nkz	4.0	0.7230	0.0000	0.2770
CD0000045000	Lysozyme	193l	4.0	0.4030	0.1090	0.4880
CD0000046000	Monellin	1mol	4.0	0.1700	0.5210	0.3080
CD0000047000	Myoglobin	1ymb	4.0	0.7390	0.0000	0.2620
CD0000048000	Myoglobin	1a6m	4.0	0.7820	0.0000	0.2190
CD0000113000	Na(+):neurotransmitter symporter (Snf (nss) family)	2a65	20.0	0.7300	0.0120	0.2580
CD0000116000	NalP	1uyn	4.0	0.0750	0.6300	0.2950
CD0000049000	Nitrogen metabolite repression regulator	1k6j	4.0	0.3550	0.1650	0.4800
CD0000128000	outer membrane lipoprotein Wza	2j58	20.0	0.2600	0.2850	0.4560
CD0000118000	Outer membrane protein G	2iwv	20.0	0.0140	0.6830	0.3030
CD0000119000	Outer membrane protein OPCA	2vdf	4.0	0.0160	0.6760	0.3090
CD0000050000	Ovalbumin	1ova	4.0	0.2910	0.2980	0.4110
CD0000051000	Ovotransferrin	1dot	4.0	0.2920	0.1730	0.5340
CD0000052000	Papain	1ppn	4.0	0.2590	0.2210	0.5190

PCDDB_id	name	pdb_id	temperature	helical	beta	irregular
CD0000054000	Pectate lyase C	1air	4.0	0.1300	0.3400	0.5290
CD0000055000	Pepsinogen	2psg	4.0	0.1480	0.3730	0.4780
CD0000056000	Peroxidase C1	7atj	4.0	0.5000	0.0380	0.4610
CD0000060000	Phenylethanolamine N-methyltransferase	1hnn	4.0	0.3520	0.2360	0.4130
CD0000057000	Phosphoglucomutase 1	3pmg	4.0	0.3570	0.2570	0.3860
CD0000058000	Phosphoglycerate kinase	3pgk	4.0	0.3450	0.1160	0.5400
CD0000059000	Phospholipase A2	1une	4.0	0.4960	0.0890	0.4150
CD0000122000	Photosynthetic reaction centre	2wjn	25.0	0.4820	0.0840	0.4340
CD0000124000	Preprotein translocase subunit secY	1rh5	20.0	0.5720	0.0280	0.4000
CD0000061000	Pyruvate kinase	1a49	4.0	0.3850	0.2010	0.4140
CD0000121000	Reaction centre protein	1pcr	20.0	0.4820	0.1000	0.4170
CD0000062000	Rhodanese	1rhs	4.0	0.3250	0.1490	0.5270
CD0000123000	Rhodopsin (dark)	1hzx	20.0	0.6000	0.0340	0.3650
CD0000109000	Rhomboid protease glpG	2nr9	10.0	0.6580	0.0000	0.3420
CD0000063000	Ribonuclease, pancreatic	3rn3	4.0	0.2090	0.3550	0.4350
CD0000064000	Rubredoxin	1r0i	4.0	0.1670	0.2220	0.6110
CD0000125000	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1	1t5s	15.0	0.4630	0.1640	0.3740
CD0000117000	Sensory rhodopsin-2	1h2s	20.0	0.8070	0.0140	0.1790
CD0001166000	Sodium/potassium-transporting ATPase	3kdp	20.0	0.3840	0.1300	0.4860
CD0001180000	Sodium/potassium-transporting ATPase	2zxe	20.0	0.4030	0.1410	0.4560
CD0000066000	Streptavidin	1stp	4.0	0.0630	0.4210	0.5160
CD0000067000	Subtilisin Carlsberg	1scd	4.0	0.2920	0.1980	0.5110
CD0000126000	Succinate dehydrogenase	1nek	4.0	0.5670	0.0180	0.4150
CD0000127000	Sucrose porin	1a0s	4.0	0.0680	0.5600	0.3740
CD0000068000	Superoxide dismutase [Cu-Zn]	1cbj	4.0	0.0430	0.4170	0.5390
CD0000069000	Thaumatin I	1thw	4.0	0.1060	0.3860	0.5080
CD0000120000	TraF protein	3jqo	25.0	0.1340	0.2980	0.5680
CD0000070000	Triose phosphate isomerase	7tim	4.0	0.4460	0.1700	0.3850
CD0000065000	Trypsin inhibitor A	1ba7	4.0	0.0170	0.3950	0.5870
CD0000071000	Ubiquitin	1ubi	4.0	0.2500	0.3420	0.4080
CD0000103000	Vitamin B12 import system permease protein BtuC	1l7v	20.0	0.5700	0.0730	0.3560
CD0000102000	Vitamin B12 transporter BtuB	1nqh	4.0	0.0550	0.5570	0.3870
CD0000110000	Voltage-gated potassium channel	1j95	20.0	0.5340	0.0000	0.4660

Table S 1: Reference data set: Note that for each PCDDB entry the values for α, 3_{10} and π-helix are summarized as helical (h), β-strand (b) also includes β-bridge, and bonded turn, bend, loop or irregular are combined as irregular (i), respectively.

wavelength (nm)	$[\Theta]_h$	$[\Theta]_b$	$[\Theta]_i$
180.0	21897.2324233	-10786.1648387	-6341.5980588
181.0	24922.7984252	-7790.61322816	-6858.04829562
182.0	28446.593157	-4256.95870651	-7650.54285156
183.0	32496.5644661	-465.825788649	-8673.62505425
184.0	37147.8083036	3354.05757256	-9788.88027785
185.0	42160.9227785	7223.53050364	-11054.8862436
186.0	47762.120681	11272.4343689	-12779.9107409
187.0	54006.7830234	15522.4822828	-14948.8628902
188.0	60495.6155691	19796.9637669	-17560.8883413
189.0	66847.7529971	23781.2774896	-20375.0006403
190.0	72773.8319741	27471.0789071	-23383.8947459
191.0	77554.1230454	30851.911299	-26443.6143346
192.0	81017.2409935	33926.6678852	-29448.6971261
193.0	83004.850573	36743.4059435	-32406.8252748
194.0	83309.2518658	39224.9105674	-35193.8004486
195.0	81507.2927616	41058.5454296	-37466.6045022
196.0	77386.8625786	42332.1637733	-39319.5609908
197.0	70518.8975416	42713.2648671	-40287.1057364
198.0	61132.6821629	42136.6031229	-40261.9682276
199.0 (irregular)	**50011.7649952**	**40461.5388483**	**-39152.0887305**
200.0	38359.9790986	37736.6674905	-37128.3587338
201.0	27029.0390822	34290.8005439	-34472.5135507
202.0	16563.1938101	30492.7277057	-31580.8614786
203.0	7044.2431275	26471.258034	-28453.9142965
204.0	-1670.68884035	22356.1876909	-25160.7020103
205.0	-9478.26342647	18244.5298804	-21765.5404446
206.0 (β-strand)	**-16120.1031708**	**14030.6953031**	**-18330.8201628**
207.0	-21390.3052576	9939.69127498	-15020.9774813
208.0	-25044.6320684	6023.36653704	-11941.6121579
209.0	-27124.255373	2543.52298135	-9381.43822642
210.0	-28160.3945008	-622.44387509	-7106.41229829
211.0	-28510.2452539	-3295.45410171	-5288.9939695
212.0	-28670.9595085	-5592.82049287	-3800.15446302
213.0	-28923.0103588	-7361.73846951	-2670.76376673
214.0	-29372.1559214	-8886.7141394	-1587.34570764
215.0	-29968.407717	-10129.8445386	-677.213121794
216.0	-30689.9260991	-10987.2959381	122.143108235
217.0	-31426.7599341	-11431.1415768	744.241837405
218.0	-32173.4533802	-11514.0723119	1248.02863836
219.0	-32902.2780758	-11338.8260718	1679.96760881
220.0 (helical)	**-33535.6207757**	**-11125.4780732**	**2203.36930204**
221.0	-33938.5595645	-10652.6537734	2568.44163159
222.0	-34019.2147652	-9877.04402691	2725.97052961
223.0	-33806.7793123	-8902.67470572	2731.60123618
224.0	-33193.0027133	-7780.00176343	2594.36007758
225.0	-32215.2719389	-6740.53880018	2458.40085882
226.0	-30851.3094111	-5749.91839891	2285.68730693
227.0	-29158.703389	-4794.92705078	2079.67012263
228.0	-27176.1051116	-3899.14738261	1829.25897716
229.0	-24934.1116119	-3049.74710591	1501.903277
230.0	-22529.073774	-2362.85258982	1211.77553217
231.0	-20054.3823061	-1751.08199578	961.849223596
232.0	-17616.3772258	-1158.05524337	690.502215213
233.0	-15293.0754101	-745.855507674	510.454819138
234.0	-13090.6484294	-355.915648024	296.509810137
235.0	-11075.599311	-130.347203996	119.224861287
236.0	-9270.43428981	50.4071004754	-24.5644635126
237.0	-7619.61457117	66.5195971873	-90.5248068546

wavelength (nm)	$[\Theta]_h$	$[\Theta]_b$	$[\Theta]_i$
238.0	-6211.46211225	-8.227459105	-59.1221479046
239.0	-4956.74341616	-4.15154282102	-110.134346627
240.0	-3893.94232353	-77.1535381017	-75.2577348608

Table S 2: Calculated values ($[\Theta]$ (grad cm^2 dmol^{-1})) for the CD basis spectra. Coloured rows: final values used for the calculation of secondary structure content (red: helical, blue: β-strand, green: irregular).

Goodness of fit

A goodness of fit between a reference curve ($[\Theta_r]$) and a query curve ($[\Theta_q]$) in the range of λ is presented as normalized root mean square deviation (NRMSD). An ideal fit would approach an NRMSD of zero (Mao et al., 1982).

$$\mathrm{NRMSD} = \sqrt{\frac{\sum\limits_{\lambda}([\Theta]_r - [\Theta]_q)^2}{\sum\limits_{\lambda}[\Theta_r]^2}} \tag{1}$$

Comparison of basis spectra

An insightful summary of commonly used methods and reference data sets is provided in the supplementary information of Greenfield (2006), and there in the software collection (README.TXT in the subfolder LINEAR).
Brahms and Brahms (1980) define reference spectra for α-helix, β- sheet, β-turn, and random coil. The α-helix spectrum was derived from sperm whale myoglobin in 0.1 M NaF at pH 7. The data were corrected for the contributions of turns and random coil and normalized to 1. Poly(Lys$^+$- Leu)$_n$ in 0.5 M NaF at pH 7 was utilized to obtain an β-sheet spectrum. Poly(Ala$_2$,Gly)$_n$ was measured in water to obtain β-turn and Poly(Pro-Lys-Leu-Lys-Leu)$_n$ in salt free solution was used for random coil.
Reed and Reed (1997) calculated their reference spectra for α-helix, β- sheet, β-turn type 1, β-turn type 2 and random coil as average values from various model peptides at different conformation.
Chen et al. (1974) contains the spectra for α- helix, β- sheet, β-turn and random coil as extracted from 15 proteins by multilinear regression.
Sreerama and Woody (1993) used 16 proteins plus poly-L-glutamate as standard. The values for α-helix, β-pleated sheet, β-turn and random coil were extracted by multilinear regression. The structural contents of the standards were calculated according to the method of Kabsch and Sander (1983).
From a combination of 32 proteins plus poly-L-glutamate, the contribution of α-helix, antiparallel β-pleated sheet, (3) parallel β-pleated sheet, β-turn and remainder was extracted by multilinear regression as described in Toumadje et al. (1992).
Perczel et al. (1992) provides standard curves for α-helix, β-turn and/or parallel β-pleated sheet, aromatic and disulfide content, unordered content and antiparallel pleated sheet. The basis spectra were deconvoluted by convex constraint analysis of 25 proteins.

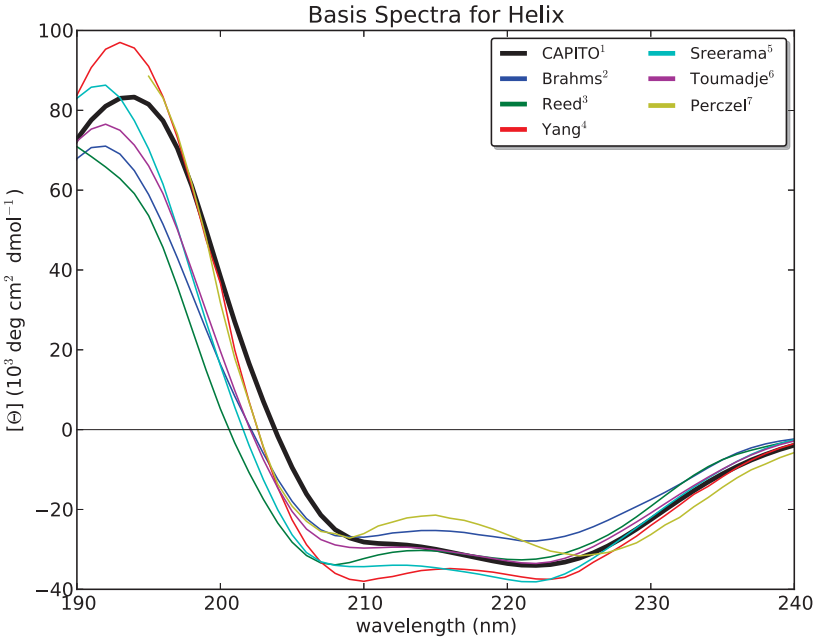
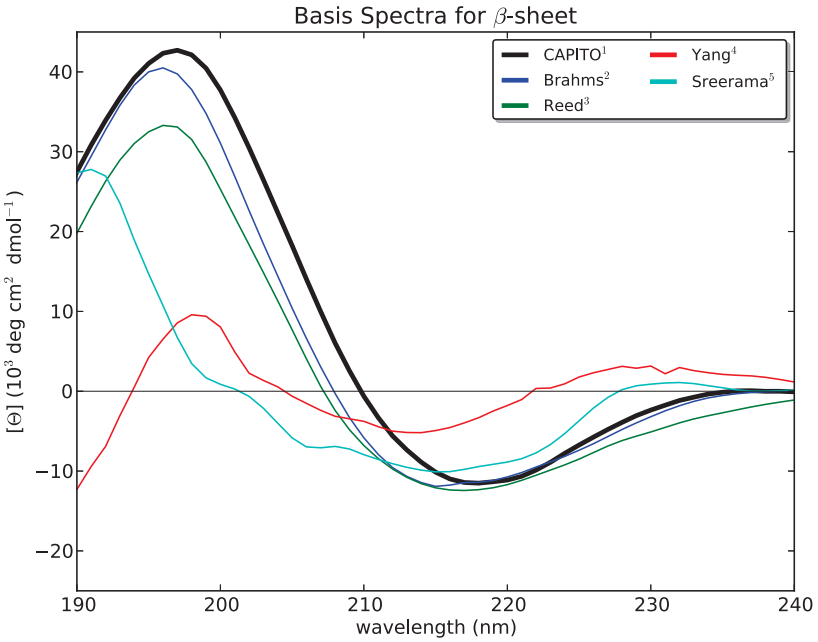
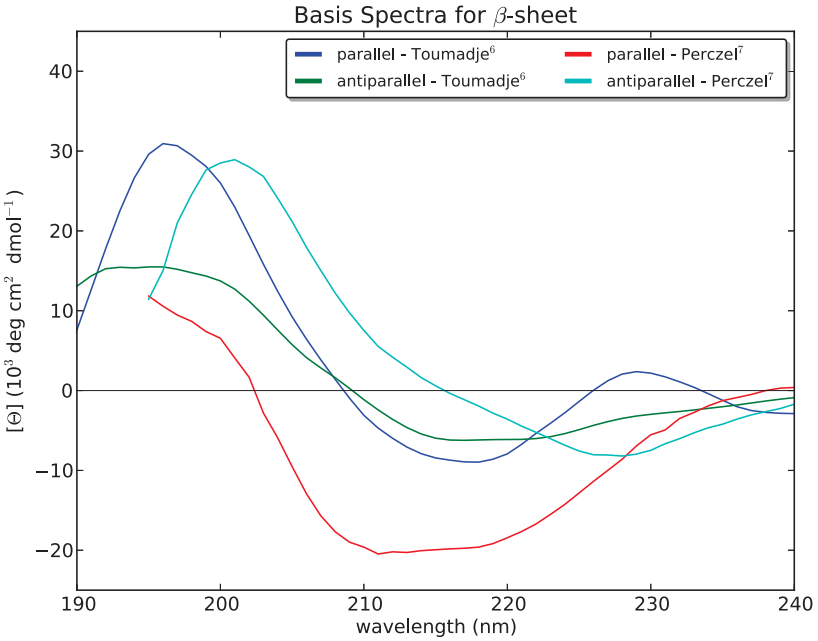


Fig. S 1: Graphical representation of helical basis spectra derived from different reference data sets and methods. References: [1]this work, [2] Brahms and Brahms (1980), [3] Reed and Reed (1997), [4] Chen et al. (1974), [5] Sreerama and Woody (1993), [6] Toumadje et al. (1992), [7] Perczel et al. (1992)



(a)



(b)

Fig. S 2: Graphical representation of β-sheet basis spectra derived from different reference data sets and methods. (a) [1]this work, [2] Brahms and Brahms (1980), [3] Reed and Reed (1997), [4] Chen et al. (1974), [5] Sreerama and Woody (1993). (b) [6] Toumadje et al. (1992) discriminate between parallel (blue) and antiparallel (green) β-pleated sheets. [7] Perczel et al. (1992) discriminate between β-turn and/or parallel (red) and antiparallel (cyan) β-pleated sheets.

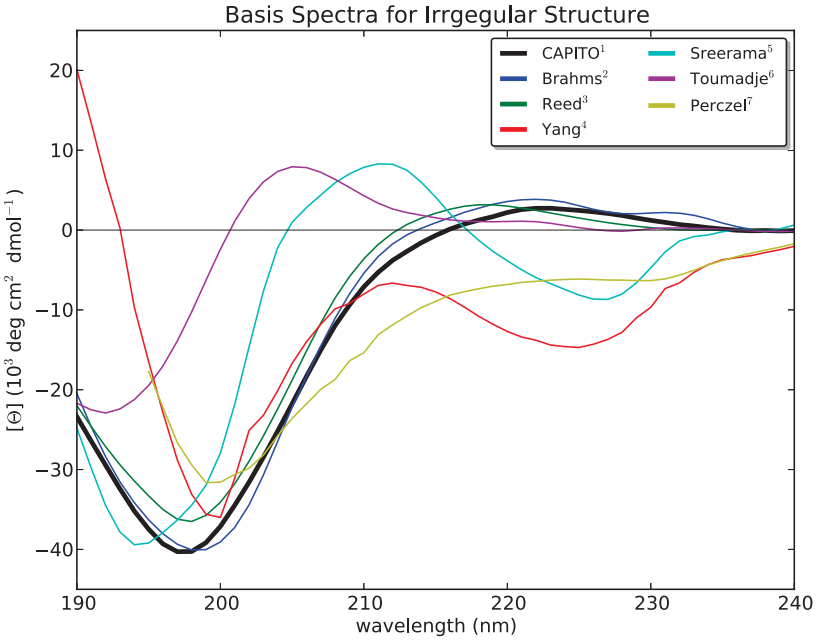


Fig. S 3: Graphical representation of irregular structure basis spectra derived from different reference data sets and methods. References: [1]this work, [2] Brahms and Brahms (1980), [3] Reed and Reed (1997), [4] Chen et al. (1974), [5] Sreerama and Woody (1993), [6] Toumadje et al. (1992), [7] Perczel et al. (1992)

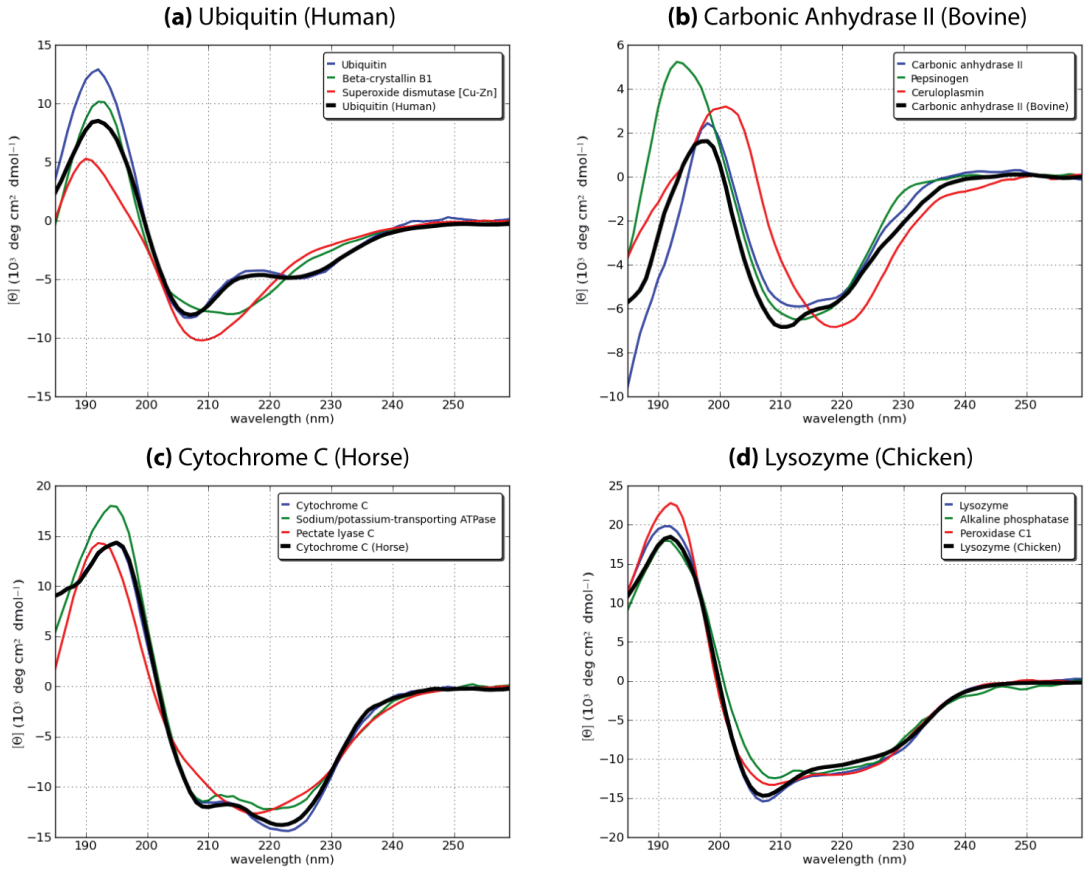


Fig. S 4: Experimental validation of CD spectra matching: CD data of (a) Ubiquitin (Human), (b) Carbonic Anhydrase II (Bovine), (c) Cytochrome C (Horse), and (d) Lysozyme (Chicken) were collected and analysed with CAPITO (black curves). As an example, the graphical output for the area difference method is shown. The best matching CD spectra from the reference data set are coloured in blue, the second best in green, and the third best hit in red, respectively.

References Supplementary

S. Brahms and J. Brahms. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J Mol Biol*, 138(2):149–178, Apr 1980.

Y. H. Chen, J. T. Yang, and K. H. Chau. Determination of the helix and beta form of proteins in aqueous solution by circular dichroism. *Biochemistry*, 13(16):3350–3359, Jul 1974.

N. J. Greenfield. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc*, 1(6): 2876–2890, 2006. doi: 10.1038/nprot.2006.202. URL `http://dx.doi.org/10.1038/nprot.2006.202`.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983. doi: 10.1002/bip.360221211. URL `http://dx.doi.org/10.1002/bip.360221211`.

D. Mao, E. Wachter, and B. A. Wallace. Folding of the mitochondrial proton adenosinetriphosphatase proteolipid channel in phospholipid vesicles. *Biochemistry*, 21(20):4960–4968, Sep 1982.

A. Perczel, K. Park, and G. D. Fasman. Deconvolution of the circular dichroism spectra of proteins: the circular dichroism spectra of the antiparallel beta-sheet in proteins. *Proteins*, 13(1):57–69, May 1992. doi: 10.1002/prot.340130106. URL `http://dx.doi.org/10.1002/prot.340130106`.

J. Reed and T. A. Reed. A set of constructed type spectra for the practical estimation of peptide secondary structure from circular dichroism. *Anal Biochem*, 254(1):36–40, Dec 1997. doi: 10.1006/abio.1997.2355. URL `http://dx.doi.org/10.1006/abio.1997.2355`.

N. Sreerama and R. W. Woody. A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal Biochem*, 209(1):32–44, Feb 1993. doi: 10.1006/abio.1993.1079. URL `http://dx.doi.org/10.1006/abio.1993.1079`.

A. Toumadje, S. W. Alcorn, and W. Johnson, Jr. Extending CD spectra of proteins to 168 nm improves the analysis for secondary structures. *Anal Biochem*, 200(2):321–331, Feb 1992.

2.2 Publication 2

Title: Solid state NMR of proteins at high MAS frequencies: symmetry-based mixing and simultaneous acquisition of chemical shift correlation spectra

Authors: **P.Bellstedt**, C.Herbst, S.Häfner, J.Leppert, M.Görlach, R. Ramachandran

Contributions: RR conceived of the NMR experiments. PB and SH expressed, PB purified and crystallized the proteins. CH designed the mixing sequences. JL technically supported NMR experiments, PB analyzed the data. RR and PB performed the experiments and wrote the manuscript. MG supervised the work and corrected the manuscript.

Status: Published in *Journal of Biomolecular NMR* 2012; 54(4): 325-35.
doi: 10.1007/s10858-012-9680-z.

Summary: Structural characterization of proteins *via* solid-state NMR depends upon the availability of a suitable microcrystalline sample. Once obtained, the lifetime of such a sample is often limited by technical stress factors (*e.g.* rotation at high frequencies or sample heating). This manuscript describes an approach for the *simultaneous* acquisition of NMR data sets to minimize the data acquisition time, thereby potentially maximizing the sample lifetime. Although finally not applicable to APTX, the method described here is demonstrated using the two model systems GB1 and SH3, of which both were essential to initially establish basic ssNMR methods.

J Biomol NMR
DOI 10.1007/s10858-012-9680-z

COMMUNICATION

Solid state NMR of proteins at high MAS frequencies: symmetry-based mixing and simultaneous acquisition of chemical shift correlation spectra

Peter Bellstedt · Christian Herbst · Sabine Häfner · Jörg Leppert · Matthias Görlach · Ramadurai Ramachandran

Received: 18 September 2012 / Accepted: 29 October 2012

Abstract We have carried out chemical shift correlation experiments with symmetry-based mixing sequences at high MAS frequencies and examined different strategies to simultaneously acquire 3D correlation spectra that are commonly required in the structural studies of proteins. The potential of numerically optimised symmetry-based mixing sequences and the simultaneous recording of chemical shift correlation spectra such as: 3D NCAC and 3D NHH with dual receivers, 3D NC′C and 3D C′NCA with sequential ^{13}C acquisitions, 3D NHH and 3D NC′H with sequential ^{1}H acquisitions and 3D CANH and 3D C'NH with broadband ^{13}C–^{15}N mixing are demonstrated using microcrystalline samples of the β1 immunoglobulin binding domain of protein G (GB1) and the chicken α-spectrin SH3 domain.

Keywords Solid state NMR · Magic angle spinning · Symmetry-based mixing · Dual receivers · Chemical shift correlation

Introduction

Multi-dimensional chemical shift correlation experiments with mixing periods leading to ^{15}N–^{13}C and ^{13}C–^{13}C dipolar and scalar coupling mediated magnetisation transfers, both band-selective and broadband, are performed in MAS solid state NMR studies of proteins (Kehlet et al. 2007; Franks et al. 2007; Chen et al. 2007; Nielsen et al. 2009; Loening et al. 2012). Although weak dipolar couplings between low γ nuclei are typically averaged out under magic angle spinning (MAS) conditions, the spatial averaging of weak dipolar interactions can be inhibited via the application of suitable RF pulse sequences. The CN_n^ν and RN_n^ν symmetry-based approach (Levitt 2002) provides a general framework for the design of different homo- and heteronuclear mixing sequences and has found a variety of applications in biomolecular MAS solid state NMR studies. The CN_n^ν class of RF pulse schemes involves the application of a basic element "*C*", corresponding to an RF cycle with unity propagator $U_{RF}(t_c) = 1$, *N* times over *n* rotor periods τ_r with successive *C* elements incremented in phase by $\nu 2\pi/N$. RF pulse sequences belonging to RN_n^ν symmetry involve the application of the pulse sandwich $\{R\varphi R\text{-}\varphi\}$, where $\varphi = \pi\nu/N$ and *R* is a 180° pulse, *N*/2 times over *n* rotor periods so as to form an RF cycle with unity propagator $U_{RF}(t_c) = 1$. *N*, *n* and ν are all integers and appropriate values for these are chosen, via the selection rule for CN_n^ν and RN_n^ν symmetry, to generate the desired average Hamiltonian. One of the difficulties with the symmetry-based RF pulse schemes based on conventional composite RF pulses, however, is that the RF field strength requirements are related to the spinning speed. This may lead to situations where the RF field strength needed becomes too large, thus limiting its applicability at high MAS frequencies that are typically employed to minimise the effects of CSA at high Zeeman field strengths and in experiments involving direct proton detection. As a result, most of the biomolecular NMR studies involving symmetry-based mixing sequences reported till date have been

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9680-z) contains supplementary material, which is available to authorized users.

P. Bellstedt · S. Häfner · J. Leppert · M. Görlach · R. Ramachandran (✉)
Biomolecular NMR spectroscopy, Leibniz Institute for Age Research, Fritz Lipmann Institute, 07745 Jena, Germany
e-mail: raman@fli-leibniz.de

C. Herbst
Department of Physics, Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani 34190, Thailand

Published online: 23 November 2012

carried out essentially at moderate MAS frequencies. To overcome this problem, we have developed a numerical approach for the design of efficient symmetry-based mixing schemes (Herbst et al. 2009a, b, c, 2010, 2011) taking into account experimental requirements and constraints, e.g. MAS frequency and RF field strength. This exploits the fact that for the recoupling and decoupling of different nuclear spin interactions the symmetry-based approach provides a large number of inequivalent symmetries involving the application of basic "*R*"/"*C*" elements of different durations. Our studies have demonstrated that the design of mixing sequences via the symmetry-based approach is neither restricted to broadband mixing nor to moderate MAS frequencies. By selecting appropriate symmetries and optimising the RF field modulation profile of the basic elements, it is equally possible to implement both high-power broadband and low-power band-selective mixing sequences for any desired MAS frequency. Here, we have carried out, homo- and heteronuclear chemical shift correlation experiments at different MAS frequencies and Zeeman field strengths using microcrystalline samples of the β1 immunoglobulin binding domain of protein G (GB1) (Zhou et al. 2007; Franks et al. 2007) and the chicken α-spectrin SH3 domain (Castellani et al. 2003; Akbey et al. 2010; Linser et al. 2011), in their protonated as well as perdeuterated form. Different strategies to achieve simultaneous acquisition of chemical shift correlation spectra that are commonly required in protein MAS NMR investigations have been examined. We show that it is possible to effectively employ numerically optimised symmetry-based mixing sequences in protein solid state NMR at high MAS frequencies and to achieve *simultaneous* acquisition of chemical shift correlation spectra 3D NCAC and 3D NHH using dual receivers, 3D NC′C and 3D C′NCA employing *sequential* acquisitions with a single receiver, 3D NHH and 3D NC′H with *sequential* ^{1}H acquisitions and 3D CANH and 3D C′NH with broadband ^{13}C–^{15}N mixing sequences.

Materials and methods

Sample preparation

Uniformly (^{13}C–^{15}N) labelled microcrystalline samples of GB1 and the SH3 domain were used in these studies. The T3Q mutant of GB1 was expressed, purified and crystallized based on a published protocol (Franks et al. 2005). Protein expression in M9 media containing 1 g/l $^{15}NH_4Cl$ and 2 g/l ^{13}C-glucose or 2 g/l [^{13}C, ^{2}H]-glucose for perdeuterated sample was induced with 0.3 mM isopropyl ß-D-thiogalactoside for 3 h at 310 K. A 5 mg/ml solution of purified GB1 in 50 mM KH_2PO_4/K_2HPO_4 was used for precipitation, microcrystals were collected by centrifugation after 1 day at room temperature and packed into a 2.5 mm rotor, which was sealed with silicone disks to minimize sample dehydration during the experiments. In case of the perdeuterated GB1 sample partial back-exchange of labile protons was achieved by dissolving lyophylized GB1 in a D_20/H_20 mixture (ratio 90:10) and incubation at 323 K for 6 h prior to crystallization. Traces of NaCl through incomplete dialysis resulted in an alternative microcrystalline form of GB1 (Frericks Schmidt et al. 2007) and is referred to as crystal form B here. Perdeuterated SH3 was expressed identically to GB1 except that the M9 media contained 1.5 g/l [^{13}C, ^{2}H]-glucose. Cells were disrupted using a french press, debris was removed by centrifugation (16,000×*g*, 277 K, 30 min) and supernatant was subjected to an anion exchange chromatography (DEAE Sepharose). The flow-through was collected and a subsequent ultracentrifugation step (50,000×*g*, 277 K, 30 min) was used to remove unwanted high molecular weight proteins/aggregates prior to size exclusion chromatography (Sephadex 75). SH3 containing fractions were combined and dialysed 2 times against 4 l 10 mM sodium citrate pH 3.5, concentrated to 10 mg/ml, lyophylized, redisolved in D_2O/H_20 mixture (ratio 70:30) and incubated at 298 K overnight to partially back-exchange the amide protons. Crystallisation was initialized by changing the pH from 3.5 to ∼7.5 by stepwise addition of NH_4OH solution. Microcrystals were collected after at least 3 days of incubation at 293 K by centrifugation and packed into a 2.5 mm rotor, which was sealed with silicone discs. Approximately 10–15 mg of protein was used to pack the rotors. Although the two microcrystalline forms of GB1 lead to slightly different spectral patterns, correlation spectra of high quality with good ^{15}N and ^{13}C spectral resolution could be generated with both samples.

NMR experiments

Multi-dimensional chemical shift correlation experiments were carried out at different MAS frequencies with either a Bruker 500 MHz wide-bore or a 750 MHz narrow-bore Avance III solid state NMR spectrometer equipped with 2.5 mm triple resonance probes and with the cooling air kept at a temperature such that the sample temperature corresponded to ∼288 K. The numerically optimised symmetry-based mixing schemes reported recently (Herbst et al. 2009a, b, c, 2010, 2011) were used with appropriate scaling of the RF field strength and duration of the basic elements, where required. Unless indicated otherwise all spectra were collected with simultaneous ^{1}H decoupling during mixing. Phase sensitive chemical shift correlation spectra were generated by the States procedure (States et al. 1982). Standard phase cycling procedures were employed to select signals arising from desired coherence transfer pathways.

Results and discussion

One of the critical steps in multidimensional protein MAS solid state NMR experiments such as the 3D NCACX, NCOCX, CONCA and CONH is the magnetisation transfer between the backbone amide nitrogens and the ^{13}CA or ^{13}CO nuclei, respectively. Considering a resonance offset range of ±2 kHz and ^{13}C and ^{15}N RF field strengths of 10 kHz, optimised sequences such as $R16_{49}^{-5,4}$ and $R16_{63}^{-4,5}$ leading to ^{15}N–^{13}C γ-encoded heteronuclear double-quantum dipolar recoupling with suppression of chemical shift anisotropies and homonuclear dipolar coupling terms have been reported recently at MAS frequencies of 15 and 20 kHz, respectively (Herbst et al. 2011). Although designed for a specific MAS frequency, one of the advantages with our numerically optimised mixing sequences, in general, is the possibility to apply them to other MAS frequencies via appropriate scaling of the RF field strength and duration of the basic elements involved. The scalability and the efficacy of the optimised symmetry $R16_{49}^{-5,4}$ for band-selective mixing is demonstrated by 2D ^{15}N–^{13}CA and ^{15}N–^{13}CO correlation experiments using GB1 (Fig. 1a, b) collected at 750 MHz at a spinning speed of 27.5 kHz. ^{15}N → ^{13}C longitudinal magnetisation exchange was achieved using the $R16_{49}^{-5,4}$ symmetry with the corresponding scaled basic "*R*" elements of 111.36 μs duration and ^{13}C and ^{15}N RF field strengths of ∼18.3 kHz. As the ^{1}H decoupling field strength of ∼125 kHz used was still significantly larger than the recoupling RF field strengths, correlation spectra with good cross-peak intensities were observed due to minimal RF field interference effects during mixing. The crosspeak patterns are consistent with that reported for GB1 (Franks et al. 2007). The $R16_{49}^{-5,4}$ symmetry also yields satisfactory spectra at MAS frequencies of 20 and 33.333 kHz (supplementary Figs. 1, 2). Small variations in the ^{13}C and ^{15}N RF field strengths were not found to have any significant effect on the observed ^{15}N → ^{13}C magnetisation transfer characteristics (supplementary Fig. 3); i.e. the performance of the numerically optimised symmetries are expected to be unaffected by minor RF field inhomogeneities (±5 %). In addition to heteronuclear mixing, the performance of different optimised band-selective homonuclear mixing sequences were also experimentally evaluated. Optimised symmetry-based RF pulse schemes such as the $C7_{30}^{1}$ and $C9_{69}^{1}$ (Herbst et al. 2011) also work effectively at high MAS frequencies for obtaining dipolar and scalar coupling mediated ^{13}C–^{13}C correlation spectra, respectively, in the aliphatic region of proteins (supplementary Fig. 4).

Where sensitivity and resolution are favourable, the use of broadband mixing sequences leading to simultaneous magnetisation transfers from ^{15}N to both the ^{13}CA and ^{13}CO nuclei in a 3D NCC experiment may facilitate sequential resonance assignments using a single

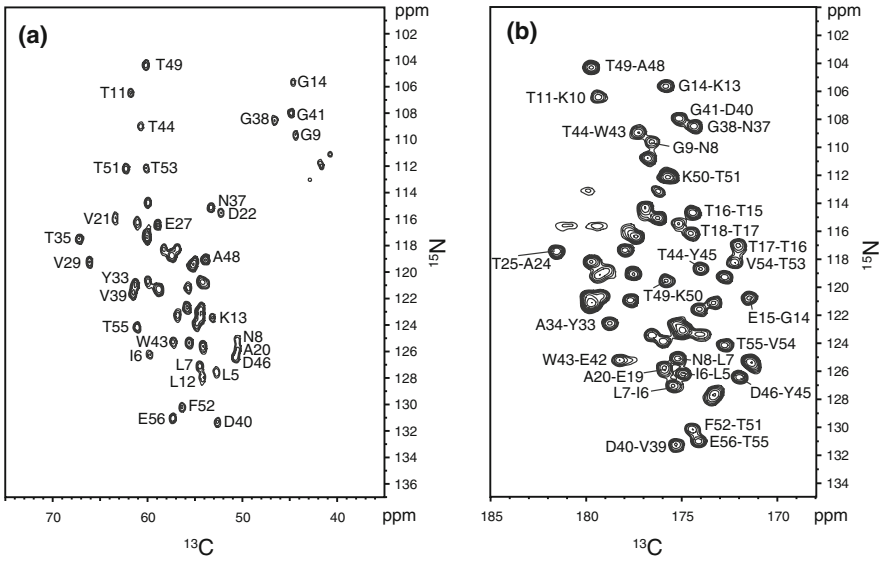


Fig. 1 (**a**) 2D ^{15}N–^{13}CA and (**b**)^{15}N–^{13}C' band-selective chemical shift correlation spectra of GB1 (crystal form A) recorded at 750 MHz and a spinning speed of 27.5 kHz. The $R16_{49}^{-5,4}$ symmetry with the corresponding numerically optimised (scaled) *R* elements reported earlier (Herbst et al. 2011), a CP contact time of 0.5 ms, τ_{mix} of 3.56 ms, ^{13}C and ^{15}N mixing RF field strengths of ∼18 kHz, 16 transients per t_1 increment, 80 t_1 increments, spectral width in the indirect dimension of 3,000 Hz and a recycle time of 2.0 s were used, keeping the ^{13}C RF carrier at 55 ppm (**a**) and 175 ppm (**b**) and the ^{15}N RF carrier at 120 ppm. Assignments are taken from the literature (Franks et al. 2007)

experiment by providing information about both the intra and inter-residue carbon chemical shifts. However, due to the large isotropic chemical shift separation between the ^{13}CA and ^{13}CO nuclei, large ^{13}C RF field strengths would typically be required to achieve efficient broadband mixing. This could lead to substantial signal losses in situations where the ^{15}N–^{13}C mixing is carried out in the presence of ^{1}H decoupling. However, in the study of perdeuterated samples at high MAS frequencies, efficient homo- and heteronuclear mixing can be carried out in the absence of ^{1}H decoupling (Huang et al. 2011; Knight et al. 2011). To achieve broadband mixing we have used a symmetry-based heteronuclear sequence (Brinkmann and Levitt 2001) designed for a simple ^{15}N–^{13}C two spin system. Figure. 2 shows the 2D ^{15}N–^{13}C broadband correlation spectrum of perdeuterated GB1 (NH:ND 10:90) acquired at 500 MHz and a spinning speed of 20 kHz using ^{15}N and ^{13}C recoupling RF field strengths of 44 kHz. The optimised symmetry $R24_{22}^{-5,7}$, designed (Herbst et al. 2009c) considering a resonance offset range of ±8 and ±2 kHz for the ^{13}C and ^{15}N nuclei, respectively, was employed. The ^{15}N–^{13}CA and ^{15}N–^{13}CO correlation peak patterns observed matches with that obtained via band-selective mixing sequences. It is also worth pointing that efficient broadband ^{15}N–^{13}C mixing can be realised with only moderate ^{15}N and ^{13}C RF field strengths. In addition to ^{15}N–^{13}C mixing, mixing periods leading to ^{13}C–^{13}C magnetisation transfers constitute another critical building block in many of the multi-dimensional chemical shift correlation experiments that are performed in the study of proteins. ^{13}C–^{13}C broadband mixing can be conveniently implemented via different proton driven spin diffusion based mixing schemes (Hou et al. 2010). However, in a variety of situations active dipolar recoupling schemes would be required; e.g. in double-quantum spectroscopy. In this context, the performance of different optimised broadband ^{13}C–^{13}C double-quantum dipolar recoupling sequences, such as $R16_{14}^{-7}$ and $C8_{10}^{3}$ reported recently (Herbst et al. 2009b, c), were also experimentally evaluated at high MAS frequencies and found to lead to correlation spectra of high quality (supplementary Fig. 5). Although symmetry-based mixing sequences have not been widely applied for protein solid state NMR studies at high MAS frequencies, the data presented here demonstrate the potential of our numerically optimised symmetry-based mixing sequences in such investigations.

Although the RF pulse schemes reported here can lead to efficient mixing, these symmetry-based mixing schemes do not lead to *complete* magnetisation transfers (Levitt 2002). Hence, sensitivity permitting, it is possible to effectively use also the *residual* magnetisation and achieve simultaneous acquisition of signals arising via different magnetisation transfer pathways of interest. For example,

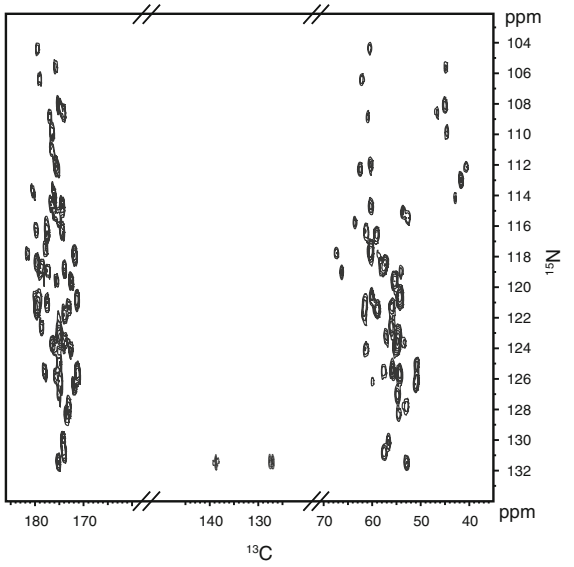


Fig. 2 2D ^{15}N–^{13}C broadband chemical shift correlation spectrum of perdeuterated GB1 (crystal form A) recorded at 500 MHz and a spinning speed of 20 kHz without ^{1}H decoupling during mixing. The $R24_{22}^{-5,7}$ symmetry with the corresponding numerically optimised *R* element reported earlier (Herbst et al. 2010, Fig. S2), a CP contact time of 2 ms, τ_{mix} of 2.2 ms, ^{13}C and ^{15}N mixing RF field strengths of 44 kHz, 512 transients per t_1 increment, 48 t_1 increments, spectral width in the indirect dimension of 2,027 Hz and a recycle time of 2.0 s were used, keeping the ^{13}C and ^{15}N RF carriers, respectively, at 115 and 120 ppm

with dual receivers, the RF pulse scheme given in Fig. 3a permits the one shot acquisition of 3D NCAC (Castellani et al. 2003; Franks et al. 2007; Schuetz et al. 2010; Sperling et al. 2010; Shi et al. 2011; Habenstein et al. 2011) and 3D NHH (Paulson et al. 2003; Zhou et al. 2007; Linser et al. 2011) correlation spectra. In this scheme, the initial transverse ^{15}N magnetisation generated by the first cross-polarisation (CP) step is allowed to evolve during the t_1/t_1' period. The magnetisation at the end of the evolution period is flipped to the *z* axis and then subjected to a period of ^{15}N → ^{13}CA longitudinal magnetisation exchange via the application of a band-selective mixing sequence. The ^{13}C polarisation generated at the end of the heteronuclear mixing period is brought to the transverse plane and allowed to evolve during the t_2 period. The ^{13}C magnetisation at the end of t_2 is then flipped to the *z* axis and subjected to longitudinal homonuclear mixing during the period τ_{mix}^{CC}. The magnetisation at the end of τ_{mix}^{CC} is rotated back to the transverse plane and detected in t_3 to generate the 3D NCAC spectrum. After the completion of the acquisition of the ^{13}C FID, proton saturation pulses with alternating *x* and *y* phases were first applied for 200 ms at a power level of ~25 kHz to achieve water suppression (Zhou and Rienstra 2008) and then the *residual* ^{15}N

longitudinal magnetisation remaining after the first $^{15}N \rightarrow {}^{13}CA$ transfer step is brought to the transverse plane and subjected to a CP step for transferring the t_1' modulated ^{15}N magnetisation to the directly attached proton. The proton magnetisation is allowed to evolve during the t_2' period and then flipped to the z axis. Longitudinal magnetisation exchange mediated by proton–proton dipolar couplings is allowed to take place during the mixing period τ_{mix}^{HH}. The proton magnetisation at the end of τ_{mix}^{HH} is rotated back to the transverse plane for direct detection in t_3' to generate the 3D NHH spectrum that provides information about proton–proton spatial proximities. With this approach using the optimised $R16_{49}^{-5,4}$ symmetry-based band-selective heteronuclear mixing scheme, we have succesfully acquired the 3D NCAC and NHH spectra of a perdeuterated sample of GB1 (Fig. 4) using RFDR (Bennett et al. 1998; Brinkmann et al. 2002; Leppert et al. 2003) during τ_{mix}^{HH} and τ_{mix}^{CC}.

While exploiting dual receiver capabilities is one way of achieving simultaneous acquisition of 3D correlation spectra of proteins in the solid state, an alternative approach to simultaneous data collection with a single receiver involves the use of sequential acquisition procedure reported recently (Gopinath and Veglia 2012a, b). For example, the RF pulse scheme shown in Fig. 3b, involving dual sequential proton detection, permits the simultaneous acquisition of 3D NCOH and NHH chemical shift correlation spectra. These spectra resulting from the magnetisation transfer pathways $^1H_N \rightarrow {}^{15}N \rightarrow {}^{13}CO \rightarrow {}^1H_N$ and $^1H_N \rightarrow {}^{15}N \rightarrow {}^1H_N \leftrightarrow {}^1H_N$, respectively, provide information about $CO \leftrightarrow H_N$ (Agarwal et al. 2010; Linser 2012) and $H_N \leftrightarrow H_N$ (Paulson et al. 2003; Zhou et al. 2007; Linser et al. 2011) spatial proximities. In this scheme, the initial transverse ^{15}N magnetisation generated by the first cross-polarisation (CP) step is allowed to evolve during the t_1/t_1' period. The magnetisation at the end of the evolution period is flipped to the z axis and then subjected to a period of $^{15}N \rightarrow {}^{13}CO$ longitudinal magnetisation exchange via the application of a band-selective mixing sequence. The ^{13}C polarisation generated at the end of the heteronuclear mixing period is brought to the transverse plane and allowed to evolve during the t_2' period. Homonuclear decoupling during ^{13}CO evolution was achieved via the application of a sandwich of pulses and delays $\{t2'/2 - (180)^{CA} - t2'/2 - (180) - \Delta\}$ with the delay Δ set to the duration of the band-selective $(180)^{CA}$ pulse. The ^{13}CO magnetisation at the end of t_2' is then flipped to the z axis and proton saturation pulses with alternating x and y phases were first applied for 200 ms at a power level of $\sim$25 kHz to achieve water suppression (Zhou and Rienstra 2008). The *residual* ^{15}N longitudinal magnetisation remaining after the first $^{15}N \rightarrow {}^{13}CO$ transfer step is brought to the transverse plane and subjected to a CP step for transferring

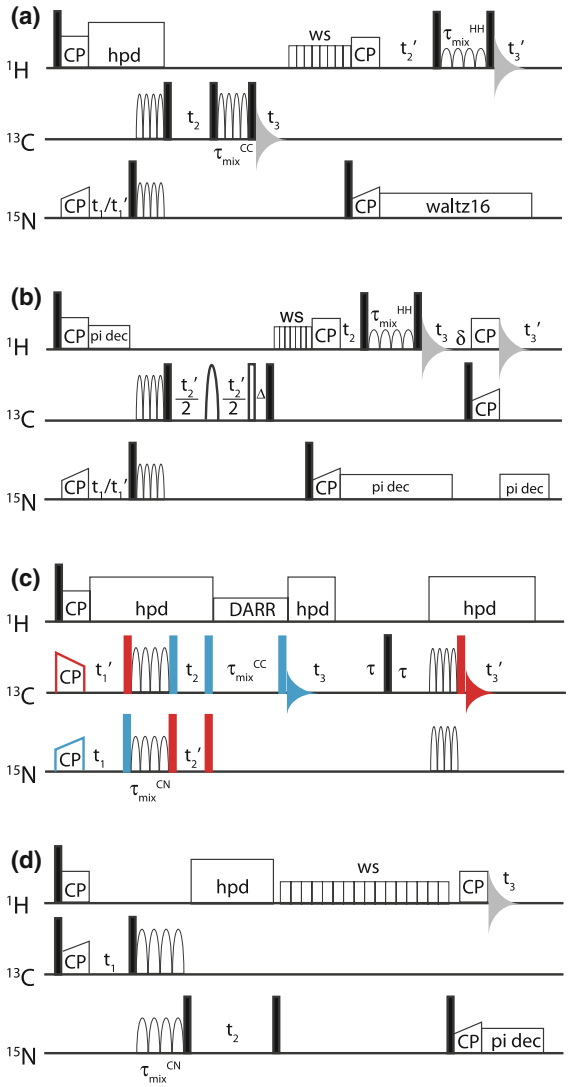


Fig. 3 RF pulse schemes for the simultaneous acquisition of (**a**) 3D NCAC and 3D NHH (**b**) 3D NCOH and 3D NHH **c** 3D NC'C (in *blue*) and 3D C'NC (in *red*) and **d** 3D CANH and 3D C'NH chemical shift correlation spectra with dual receivers (**a**), dual sequential acquisition in the direct dimension (**b**,**c**) and broadband ^{13}C–^{15}N mixing (**d**). Open and filled *rectangles* represent 180° and 90° pulses, respectively

the t_1 modulated ^{15}N magnetisation to the directly attached proton. The proton magnetisation is allowed to evolve during the t_2 period and then flipped to the z axis. Longitudinal magnetisation exchange mediated by proton–proton dipolar couplings is allowed to take place during the mixing period τ_{mix}^{HH}. The proton magnetisation at the end of τ_{mix}^{HH} is rotated back to the transverse plane for direct detection in t_3 to generate the 3D NHH spectrum. After the completion of the first ^1H acquisition and after a short delay, the ^{13}CO magnetisation is brought to the transverse

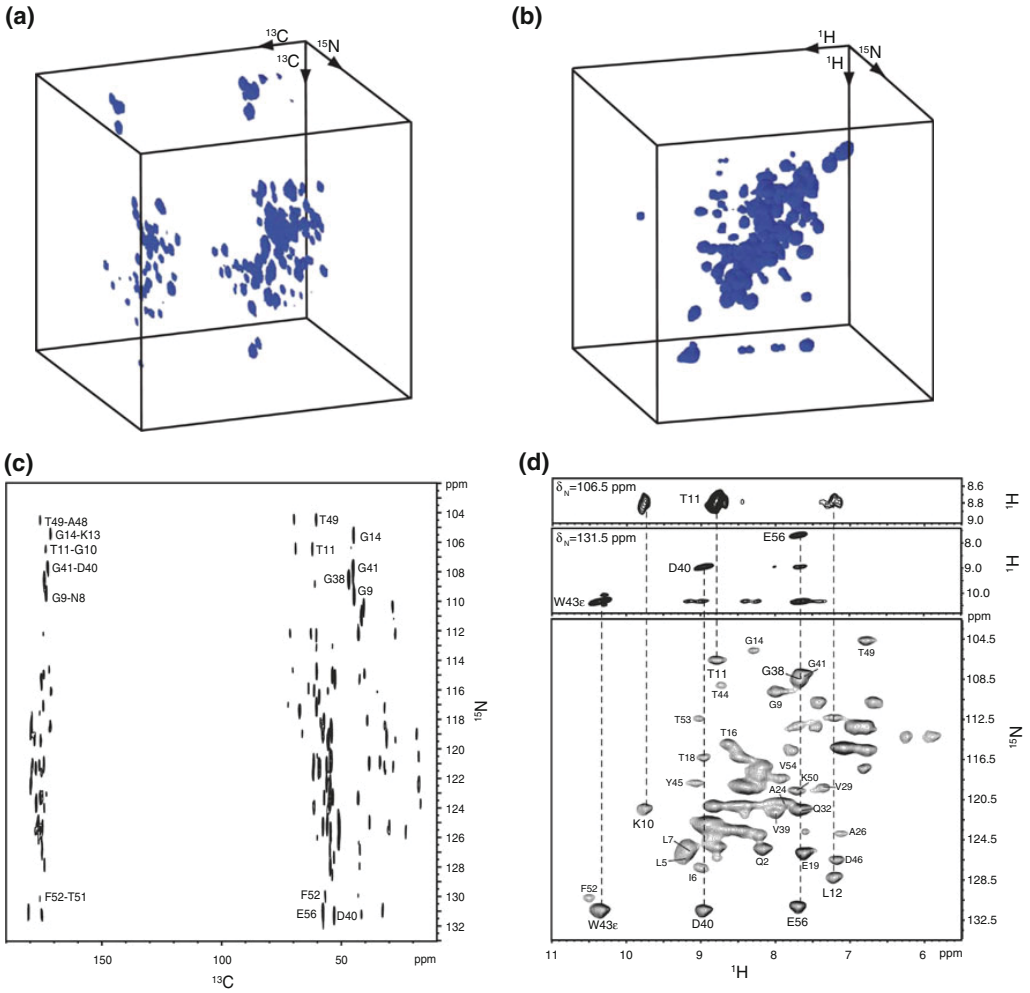


Fig. 4 Simultaneously acquired 3D NCAC (**a**) and 3D NHH (**b**) spectra of perdeuterated GB1 (NH:ND 10:90 (crystal form A)) recorded at 750 MHz and a spinning speed of 27.5 kHz. The $R16_{49}^{-5,4}$ symmetry with the corresponding numerically optimised *R* element (Fig. 1), ^{1}H–^{15}N CP contact time of 2 ms, ^{15}N–^{13}C mixing time of 3.56 ms, ^{13}C and ^{15}N RF field strength of 18 kHz during mixing, 32 transients per t_1 increment, 24 t_1 increments, 56 t_2 increments, spectral widths in the indirect dimensions of 3,000 Hz (^{15}N) and 6,036 (^{13}CA/^{1}H) and a recycle time of 2.0 s. The ^{1}H, ^{13}C and ^{15}N RF carriers were set to 2.5, 55 and 120 ppm, respectively. The ^{13}C RF carrier was switched to 100 ppm during ^{13}C–^{13}C mixing carried out in the absence of ^{1}H decoupling via RFDR using tanh/tan adiabatic pulses of 36.36 µs duration with $R6_6^2R6_6^{-2}$ supercycling (Brinkmann et al. 2002) for a period of ∼1.7 ms and employing an RF field strength of ∼70 kHz. The ^{1}H RF carrier was kept at 9.5 ppm during ^{1}H evolution in t_2 and during ^{1}H–^{1}H mixing via RFDR using tanh/tan adiabatic pulses of 20.36 ms duration with xy-16 supercycling for a period of ∼4.7 ms and employing an RF field strength of ∼125 kHz. ^{15}N decoupling during ^{1}H acquisition was carried out via waltz16 decoupling scheme at a power level of ∼7 kHz. (**c**) ^{15}N^{13}C spectrum from a 2D version of the 3D experiment acquired with 128 transients per t_1 increment, 64 t_1 increments and with other parameters as mentioned above. (**d**) ^{1}H–^{1}H strips taken from the 3D NHH spectrum at the ^{15}N chemical shifts indicated. The ^{15}N–^{1}H HSQC spectrum of perdeuterated GB1 (NH:ND 10:90) collected at a spinning speed of 27.5 kHz is shown for reference. Assignments in (d) were taken from the literature (Zhou et al. 2007)

plane and subjected to a reverse CP step for transferring the magnetisaton to different spatially proximal protons for direct detection in t_3'. This approach including the optimised $R16_{49}^{-5,4}$ symmetry-based band-selective heteronuclear mixing scheme, yields satisfactory 3D NC'H and NHH spectra (Fig. 5) of a perdeuterated sample of the SH3 domain (NH:ND 30:70). It is worth emphasising that no significant signal losses were observed by keeping the ^{13}CO magnetisation in the *z* state even for a period of ∼200 ms in the perdeuterated sample studied here.

The dual sequential acquisition procedure is also applicable for the collection of 3D data with ^{13}C direct

detection. For example, the RF pulse scheme (Fig. 3c) permits the simultaneous acquisition of 3D NCOC (Castellani et al. 2003; Franks et al. 2007; Schuetz et al. 2010; Sperling et al. 2010; Shi et al. 2011) and 3D CONCA (Astrof et al. 2001; Schuetz et al. 2010; Shi et al. 2011) correlation spectra for a fully protonated protein sample and resulting from the magnetisation transfer pathways $^{1}H \rightarrow {}^{15}N \rightarrow {}^{13}CO \rightarrow {}^{13}C$ and $^{1}H \rightarrow {}^{13}CO \rightarrow {}^{15}N \rightarrow {}^{13}CA$, respectively. Unlike the RF pulse scheme given in Fig. 3a, the first step in the scheme Fig. 3c involves the simultaneous generation of transverse $^{15}N/^{13}C$ magnetisations by the cross-polarisation (CP) procedure (see Herbst et al. 2008, 2010). These transverse magnetisations are allowed to evolve during the t_1/t_1' period, flipped to the

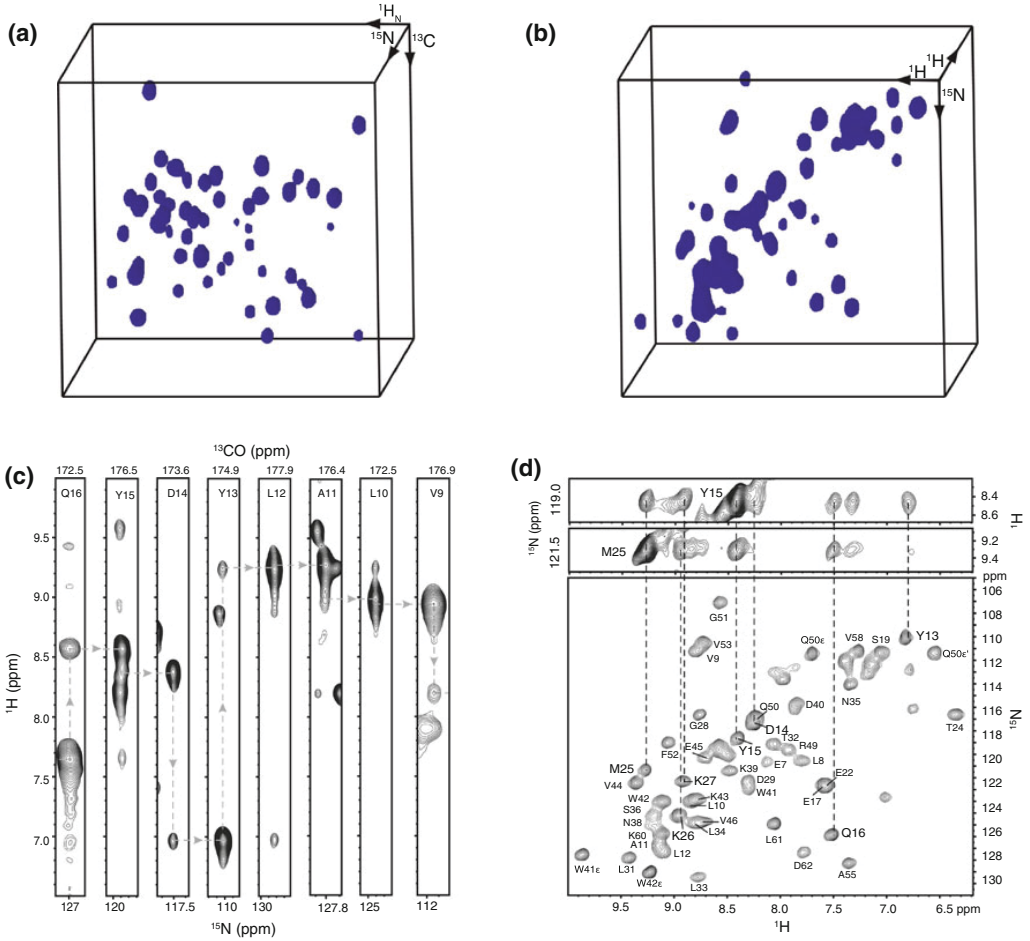


Fig. 5 Simultaneously acquired 3D NCOH (**a**) and 3D NHH (**b**) spectra of perdeuterated SH3 domain (NH:ND 30:70) recorded at 500 MHz and a spinning speed of 33.333 kHz. The $R16_{49}^{-5,4}$ symmetry with the corresponding numerically optimised *R* element, $^{1}H–^{15}N$ CP contact time of 0.5 ms, $^{1}H–^{13}C$ CP contact time of 4 ms, $^{15}N–^{13}C$ mixing time of 2.94 ms, ^{13}C and ^{15}N RF field strength of ~22 kHz during mixing, 64 transients per t_1 increment, 24 t_1 increments, 48 t_2 increments, spectral widths in the indirect dimensions of 1,500 Hz (^{15}N) and 4,000 Hz ($^{13}CO/^{1}H$) and a recycle time of 2.0 s were used, keeping the ^{1}H, ^{13}C and ^{15}N RF carriers at 9.0, 174 and 120 ppm, respectively. RFDR using tanh/tan adiabatic pulses of 20.00 μs duration with xy-16 supercycling for a period of ~7.68 ms and employing an RF field strength of ~125 kHz was employed for $^{1}H–^{1}H$ mixing. ^{15}N decoupling was carried out via the repetitive application of ^{15}N π pulses (Zhou et al. 2007) of 64 μs duration with an inter-pulse delay of 30 μs. ^{1}H decoupling during ^{15}N evolution was achieved via the repetitive application of ^{1}H π pulses of 8 μs duration with an inter-pulse delay of 30 μs. A numerically designed (Herbst et al., unpublished) phase-modulated band-selective refocussing 180° pulse of 200 μs duration was employed during the t_2' period to achieve homonuclear *J*-decoupling. (c) $^{15}N–^{1}H$ strips from the 3D NCOH spectrum taken at the ^{13}CO chemical shifts indicated. These spectra show the sequential walk along the backbone residues spanning the region Val9-Glu16. In addition to correlations arising from $CO^{i-1} \rightarrow H_N^i$ and $CO^{i-1} \rightarrow H_N^{i-1}$ magnetisation transfers, weak non-sequential correlations, e.g. due to hydrogen bonds across β-strands, are also observed. (d) $^{1}H–^{1}H$ strips taken from the 3D NHH spectrum at the ^{15}N chemical shifts indicated. The $^{15}N–^{1}H$ HSQC spectrum of perdeuterated SH3 domain (NH:ND 30:70) collected at a spinning speed of 33.333 kHz is given for reference. Assignments in (d) were taken from the literature (Lewandowski et al. 2011)

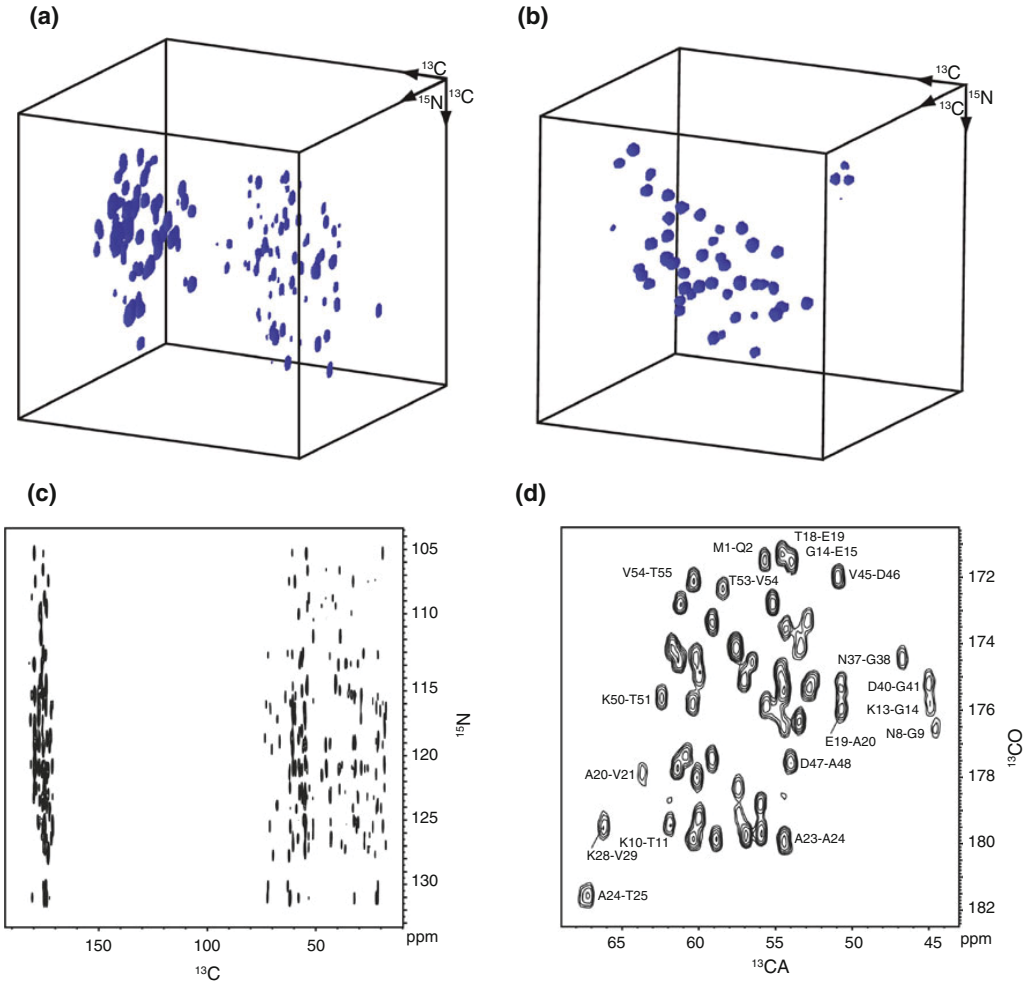


Fig. 6 Simultaneously acquired 3D NC'C (**a**) and 3D C'NCA (**b**) spectra of GB1 (crystal form A) recorded at 750 MHz and a spinning speed of 27.5 kHz. The $R16_{49}^{5,4}$ symmetry with the corresponding numerically optimised *R* element as in Fig. 1a, b, ^{1}H–^{13}C CP contact time of 1 ms, ^{1}H–^{15}N CP contact time of 1 ms, ^{15}N–^{13}C mixing time of 3.56 ms, ^{13}C and ^{15}N RF field strength of 18 kHz during mixing, 32 transients per t_1 increment, 32 t_1 increments, 48 t_2 increments, ^{15}N and ^{13}C spectral widths of 3,000 Hz in the indirect dimension and a recycle time of 2.0 s were used, keeping the ^{13}C and ^{15}N RF carriers, respectively, 175 and 120 ppm. ^{13}C–^{13}C mixing in (a) was achieved via the DARR procedure with a mixing time of 100 ms. The ^{13}C RF carrier was switched to 55 ppm at the beginning of the second ^{15}N → ^{13}C heteronuclear mixing period and switched back to 175 ppm during detection. (**c**) ^{15}N^{13}C and (d) ^{13}CO–^{13}CA spectra from a 2D version of the 3D experiment acquired with 256 transients per t_1 increment, 64 t_1 increments and with other parameters as mentioned above

z axis at the end of the evolution period and then subjected to a period of ^{15}N ↔ ^{13}CO longitudinal magnetisation exchange via the application of a band-selective symmetry-based mixing sequence. The ^{15}N/^{13}CO polarisations at the end of the heteronuclear mixing period are brought to the transverse plane and allowed to evolve during the t_2/t_2' period. The ^{15}N/^{13}CO magnetisations at the end of the evolution period are then flipped to the *z* axis. The ^{13}CO polarisation is first subjected to longitudinal homonuclear mixing during the period τ_{mix}^{CC}; e.g. via the DARR scheme (Takegoshi et al. 2001). The carbon magnetisation at the end of the mixing period is rotated back to the transverse plane and detected in t_3 to generate the 3D NCOC spectrum. After the completion of the acquisition of the first ^{13}C FID, the ^{15}N longitudinal magnetisation is then subjected to a band-selective ^{15}N → ^{13}CA longitudinal magnetisation exchange by switching the ^{13}C RF carrier frequency to the CA region. The ^{13}CA magnetisation at the end of the second heteronuclear mixing period is rotated back to the transverse plane for direct detection in t_3' to generate the 3D CONCA spectrum that provides connectivity information about two adjacent amino acid residues. After the

acquisition of the first ^{13}C FID, to minimise the effects of the residual transverse and longitudinal ^{13}C magnetisations on the ^{13}C FID collected subsequently, a sandwich of $\{\tau-(\pi/2)^{C}-\tau\}$ with $\tau = 4$ ms was applied in the absence of ^{1}H decoupling (Gopinath and Veglia 2012a, b). Eliminating signals arising from unwanted magnetisation transfer pathways via appropriate phase cycling of the RF pulses involved, this procedure allows to simultaneously acquire 3D NCOC and 3D CONCA correlation spectra (Fig. 6). The simultaneous collection of 3D NCACX and CANCO correlation spectra via dual sequential ^{13}C acquisition has also been demonstrated recently (Gopinath and Veglia 2012a).

While the 3D spectra shown in Figs. 4, 5, and 6 were generated with band-selective ^{15}N–^{13}C mixing schemes, correlation spectra resulting from ^{13}CO → ^{15}N and ^{13}CA → ^{15}N magnetisation transfer pathways can be simultaneously collected by making use of broadband mixing schemes, e.g. as shown by the RF pulse scheme (Fig. 3d) for the broadband acquisition of a 3D CNH spectrum of a protein. This RF pulse scheme has been successfully employed with a perdeuterated protein sample for the "one-shot" collection of 3D CANH (Knight et al. 2011; Ward et al. 2011) and 3D CONH (Knight et al. 2011; Ward et al. 2011) correlation spectra resulting from the magnetisation transfer pathway ^{1}H → ^{13}CA/^{13}CO → ^{15}N → ^{1}H. The initial transverse ^{13}C magnetisation is first generated by a combination of direct excitation and a cross-polarisation step (Linser 2011). It is then allowed to evolve in the absence of ^{1}H decoupling during the t_1 period. The magnetisation at the end of t_1 is flipped to the z axis and subjected to a period of ^{13}C → ^{15}N longitudinal magnetisation exchange via the application of a broadband mixing sequence. The ^{15}N polarisation generated at the end of the heteronuclear mixing period is brought to the transverse plane and allowed to evolve during the t_2 period. The magnetisation at the end of the t_2 is then flipped to the z axis. Proton saturation pulses with alternating x and y phases were then applied for 200 ms at a power level of ~30 kHz to achieve water suppression. Subsequently the ^{15}N longitudinal magnetisation is brought to the transverse plane and subjected to a CP step for transferring the magnetisaton to the directly attached proton and for direct detection in t_3. The ^{13}C RF carrier was kept near the CA region during the t_1 period and the spectral width was adjusted so as to fold the ^{13}CO resonances into the ^{13}CA spectral window. This approach including the optimised $R24_{22}^{-5,7}$ broadband heteronuclear mixing scheme (Herbst et al. 2010), yields a satisfactory 3D broadband CNH spectrum (Fig. 7) of the perdeuterated sample of the SH3 domain.

Although the acquisitions of only a few representative spectra have been demonstrated here, the approach outlined may be extended to acquire other correlation spectra. For example, with simultaneous excitation of transverse ^{15}N/^{13}C magnetisations by the first cross-polarisation step and employing dual sequential proton detection, it may be possible to achieve the simultaneous acquisition of 3D NCOH and 3D CON(H)H (Zhou et al. 2007) chemical shift correlation spectra, resulting from the magnetisation transfer pathways $^{1}H_N$ → ^{15}N → ^{13}CO → $^{1}H_N$ and $^{1}H_N$ → ^{13}CO → ^{15}N → $^{1}H_N$ ↔ $^{1}H_N$, respectively, and extract structural constraints. Dual receivers can also be employed to acquire simultaneously 3D NCC and 3D CNH correlation spectra of proteins. Broadband heteronuclear mixing sequences leading to simultaneous ^{13}CA → ^{15}N and ^{13}CO → ^{15}N magnetisation transfers can be exploited

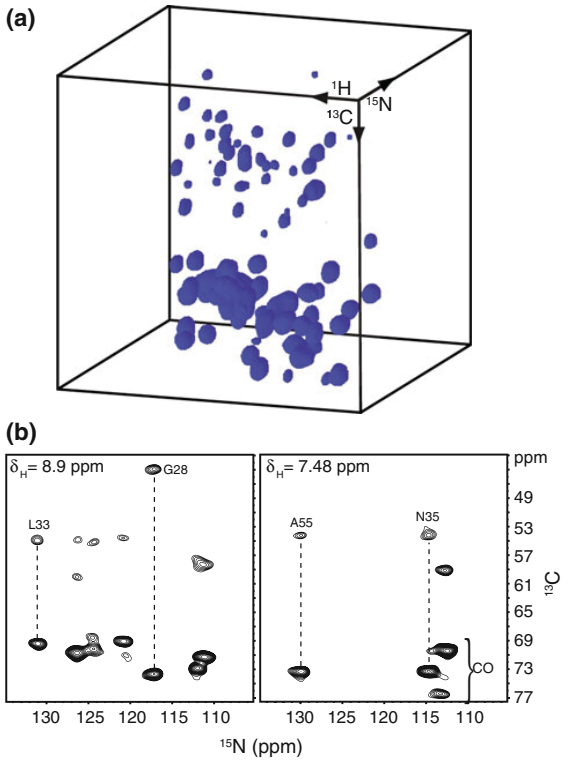


Fig. 7 3D CNH broadband chemical shift correlation spectrum (**a**) of perdeuterated SH3 domain of chicken α-spectrin (NH:ND 30:70) recorded at 500 MHz and a spinning speed of 20 kHz resulting from the magnetisation transfer pathway ^{1}H → ^{13}C → ^{15}N → ^{1}H. The $R24_{22}^{-5,7}$ symmetry with the corresponding numerically optimised R element, ^{1}H–^{13}C CP contact time of 3 ms, ^{1}H–^{15}N CP contact time of 1.5 ms, ^{15}N–^{13}C mixing time of 3.3 ms, ^{13}C and ^{15}N RF field strength of 44 kHz during mixing, 96 transients per t_1 increment, 20 t_1 increments, 32 t_2 increments, spectral widths in the indirect dimensions of 1,500 Hz (^{15}N) and 4,375 Hz (^{13}C) and a recycle time of 2.0 s. The ^{1}H, ^{13}C and ^{15}N RF carriers were set at 1, 58 and 120 ppm, respectively. The ^{13}C RF carrier was switched to 115 ppm during ^{15}N–^{13}C mixing in the absence of ^{1}H decoupling. ^{15}N decoupling in the direct dimension was carried out via the repetitive application of ^{15}N π pulses (Zhou et al. 2007) of 64 μs duration with an inter-pulse delay of 50 μs. (**b**) ^{13}C–^{15}N strips taken from the 3D CNH spectrum at the ^{1}H chemical shifts indicated

for the 3D DQ(CACO)NH-type of experiments reported recently (Ward et al. 2011). The approach outlined here for the simultaneous acquisition of multidimensional 3D data sets may also be combined with other approaches; e.g. sparse sampling in the indirect dimension, for further reducing the data acquisition times. In summary, the results presented demonstrate that it is possible to effectively use, even at high MAS frequencies, numerically designed high-power broadband and low-power band-selective homo- and heteronuclear symmetry-based mixing sequences and to achieve, with dual receivers, sequential acquisitions and broadband ^{15}N–^{13}C mixing schemes, the simultaneous acquisition of 3D correlation spectra of high quality and of interest in protein structural studies.

Acknowledgments The FLI is a member of the Science Association'Gottfried Wilhelm Leibniz' (WGL) and is financially supported by the Federal Government of Germany and the State of Thuringia. Also thanks to the Leibniz Graduate School on Aging and Age-Related Diseases (LGSA) for funding and support.

References

Agarwal V, Linser R, Fink U, Faelber K, Reif B (2010) Identification of hydroxyl protons, determination of their exchange dynamics and characterisation of hydrogen bonding in a microcrystallin protein. J Am Chem Soc 132:3187–3195

Akbey U, Lange S, Franks WT, Linser R, Rehbein K, Diehl A, Rossum BJ, Reif B, Oschkinat O (2010) Optimum levels of exchangeable protons in perdeuterated proteins for proton detection in MAS NMR spectroscopy. J Biomol NMR 46:67–73

Astrof NS, Lyon CE, Griffin RG (2001) Triple resonance solid state NMR experiments with reduced dimensionality evolution periods. J Magn Reson 152:303–307

Bennett AE, Rienstra CM, Griffiths JM, Zhen W, Lansbury PT, Griffin RG (1998) Homonuclear radio frequency-driven recoupling in rotating solids. J Chem Phys 108:9463–9479

Brinkmann A, Levitt M (2001) Symmetry principles in the nuclear magnetic resonance of spinning solids: heteronuclear recoupling by generalized Hartmann-Hahn sequences. J Chem Phys 115: 357–384

Brinkmann A, Schmedt auf der Günne J, Levitt MH (2002) Homonuclear zero-quantum recoupling in fast magic-angle spinning nuclear magnetic resonance. J Magn Reson 156:79–96

Castellani F, van Rossum BJ, Diehl A, Rehbein K, Oschkinat H (2003) Determination of solid-state NMR structures of proteins by means of three-dimensional 15 N–13C–13C dipolar correlation spectroscopy and chemical shift analysis. Biochemistry 42:11476–11483

Chen L, Kaiser JM, Lai J, Polenova T, Yang J, Rienstra CM, Mueller LJ (2007) *J*-based 2D homonuclear and heteronuclear correlation in solid-state proteins. Magn Reson Chem 45:S84–S92

Franks WT, Zhou DH, Wylie BJ, Money BG, Graesser DT, Frericks HL, Sahota G, Rienstra CM (2005) Magic-angle spinning solid-state NMR spectroscopy of the β1 immunoglobulin binding domain of protein G (GB1): ^{15}N and ^{13}C chemical shift assignments and conformational Analysis. J Am Chem Soc 127:12291–12305

Franks WT, Kloepper KD, Wylie BJ, Rienstra CM (2007) Four-dimensional heteronuclear correlation experiments for chemical shift assignments of solid proteins. J Biomol NMR 39: 107–131

Frericks Schmidt HL, Sperling LJ, Gao YG, Wylie BJ, Boettcher JM, Wilson SR, Rienstra CM (2007) Crystal polymorphism of protein GB1 examined by solid-state NMR spectroscopy and X-ray diffraction. J Phys Chem B 111:14362–14369

Gopinath T, Veglia G (2012a) Dual acquisition magic-angle solid-state NMR-spectroscopy: simultaneous acquisition of multidimensional spectra of biomacromolecules. Angew Chem Int Ed 51:1–6

Gopinath T, Veglia G (2012b) 3D DUMAS: simultaneous acquisition of three-dimensional magic angle spinning solid-state NMR experiments of proteins. J Magn Reson 220:79–84

Habenstein B, Wasmer C, Bousset L, Sourigues Y, Schutz A, Loquet A, Meier BH, Melki R, Böckmann A (2011) Extensive de novo solid-state NMR assignments of the 33 kDa C-terminal domain of the Ure2 prion. J Biomol NMR 51:235–243

Herbst C, Riedel K, Ihle Y, Leppert J, Ohlenschläger O, Görlach M, Ramachandran R (2008) MAS solid state NMR of RNAs with multiple receivers. J Biomol NMR 41:121–125

Herbst C, Herbst J, Kirschstein A, Leppert J, Ohlenschläger O, Görlach M, Ramachandran R (2009a) Design of high-power, broadband 180° pulses and mixing sequences for fast MAS solid state chemical shift correlation NMR spectroscopy. J Biomol NMR 43:51–61

Herbst C, Herbst J, Kirschstein A, Leppert J, Ohlenschläger O, Görlach M, Ramachandran R (2009b) Recoupling and decoupling of nuclear spin interactions at high MAS frequencies: numerical design of CN_n^v symmetry-based RF pulse schemes. J Biomol NMR 44:175–184

Herbst C, Herbst J, Leppert J, Ohlenschläger O, Görlach M, Ramachandran R (2009c) Numerical design of RN_n^v symmetry-based RF pulse schemes for recoupling and decoupling of nuclear spin interactions at high MAS frequencies. J Biomol NMR 44:235–244

Herbst C, Herbst J, Leppert J, Ohlenschläger O, Görlach M, Ramachandran R (2010) Broadband 15N–13C dipolar recoupling via symmetry-based RF pulse schemes at high MAS frequencies. J Biomol NMR 47:7–17

Herbst C, Herbst J, Leppert J, Ohlenschläger O, Görlach M, Ramachandran R (2011) Chemical shift correlation at high MAS frequencies employing low-power symmetry-based mixing schemes. J Biomol NMR 50:277–284

Hou G, Yan S, Sun S, Han Y, Byeon IL, Ahn J, Concel J, Samoson A, Gronenborn AM, Polenova T (2010) Spin diffusion driven by R-symmetry sequences: applications to homonuclear correlation spectroscopy in MAS NMR of biological and organic solids. J Am Chem Soc 133:3943–3953

Huang KY, Siemer AB, McDermott AE (2011) Homonuclear mixing sequences for perdeuterated proteins. J Magn Reson 208: 122–127

Kehlet C, Bjerring M, Sivertsen AC, Kristensen T, Enghild JJ, Glaser SJ, Khaneja N, Nielsen NC (2007) Optimal control based NCO and NCA experiments for spectral assignments in biological solid-state NMR spectroscopy. J Magn Reson 188:216–230

Knight MJ, Webber AL, Pell AJ, Guerry P, Barbet-Massin E, Bertini I, Felli IC, Gonnelli L, Pierattelli R, Emsley L, Lesage A, Herrmann T, Pintacuda G (2011) Fast resonance assignment and fold determination of human superoxide dismutase by high-resolution proton-detected solid-state MAS NMR spectroscopy. Angew Chem Int Ed 50:11697–11701

Leppert J, Heise B, Ohlenschläger O, Görlach M, Ramachandran R (2003) Broadband RFDR with adiabatic inversion pulses. J Biomol NMR 26:13–24

Levitt MH (2002) Symmetry-based pulse sequences in magic-angle spinning solid-state NMR. In: Grant DM, Harris RK (eds) Encyclopedia of nuclear magnetic resonance. Wiley, Chichester

Lewandowski JR, Dumaz JN, Akbey U, Lange S, Emsley L, Oschkinat H (2011) Enhanced resolution and coherence lifetimes in the solid-state NMR spectroscopy of perdeuteraded proteins under ultrafast magic-angle spinning. J Phys Chem Lett 2: 2205–2211

Linser R (2011) Side-chain to backbone correlations from solid-state NMR of perdeuterated proteins through combined excitation and long-range magnetization transfers. J Biomol NMR 51:221–226

Linser R (2012) Backbone assignment of perdeuterated proteins using long-range H/C transfers. J Biomol NMR 52:151–158

Linser R, Bardiaux B, Higman V, Fink U, Reif B (2011) Structure calculation from unambiguous long-range amide and methyl 1H–1H distance restraints for a microcrystalline protein with MAS solid-state NMR spectroscopy. J Am Chem Soc 133: 5905–5912

Loening NM, Bjerring M, Nielsen NC, Oschkinat H (2012) A comparison of NCO and NCA transfer methods for biological solid-state NMR spectroscopy. J Magn Reson 214:81–90

Nielsen AB, Bjerring M, Nielsen JT, Nielsen NC (2009) Symmetry-based dipolar recoupling by optimal control: band-selective experiments for assignment of solid-state NMR spectra of proteins. J Chem Phys 131:025101

Paulson EK, Morcombe CR, Gaponenko V, Dancheck B, Byrd RA, Zilm KW (2003) High-sensitivity observation of dipolar exchange and NOEs between exchangeable protons in proteins by solid-state NMR spectroscopy. J Am Chem Soc 125: 14222–14223

Schuetz A, Wasmer C, Habenstein B, Verel R, Greenwald J, Reik R, Böckmann A, Meier BH (2010) Protocols for the sequential solid-state NMR spectroscopic assignments of a uniformly labeled 25 kDa protein:HET-s(1–227). ChemBioChem 11: 1543–1551

Shi L, Kawamura I, Jung KH, Brown LS, Ladizhansky V (2011) Conformation of a seven- helical transmembrane photosensor in the lipid environment. Angew Chem Int Ed 50:1302–1305

Sperling LJ, Berthold DA, Sasser TL, Jeisy-Scott V, Rienstra CM (2010) Assignments strategies for large proteins by magic-angle spinning NMR: the 21-kDa disulfide-bond forming enzyme Dsba. J Mol Biol 399:268–282

States DJ, Haberkorn RA, Ruben DJ (1982) A two-dimensional nuclear Overhauser experiment with pure absorption phase in four quadrants. J Magn Reson 48:286–292

Takegoshi K, Nakamura S, Terao T (2001) 13C–1H dipolar-assisted rotational resonance in magic-angle spinning NMR. Chem Phys Lett 344:631–637

Ward ME, Shi L, Lake E, Krishnamurthy S, Hutchins H, Brown LS, Ladizhansky V (2011) Proton-detected solid-State NMR reveals intramembrane polar networks in a seven-helical transmembrane protein proteorhodopsin. J Am Chem Soc 133:17434–17443

Zhou DH, Rienstra CM (2008) High-performance solvent suppression for proton detected solid- state NMR. J Magn Reson 192: 167–172

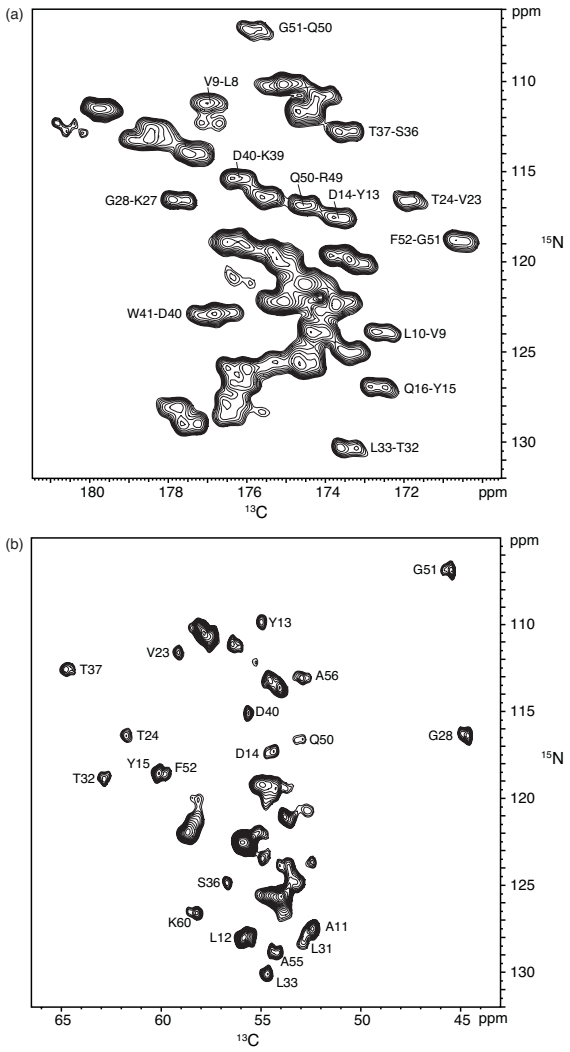
Zhou DH, Shea JJ, Nieuwkoop AJ, Franks WT, Wylie BJ, Mullen C, Sandoz D, Rienstra CM (2007) Solid-State Protein-Structure Determination with Proton-Detected Triple-Resonance 3D Magic-Angle-Spinning NMR Spectroscopy. Angew Chem Int Ed 46:8380–8383

Supplementary material

for the manuscript

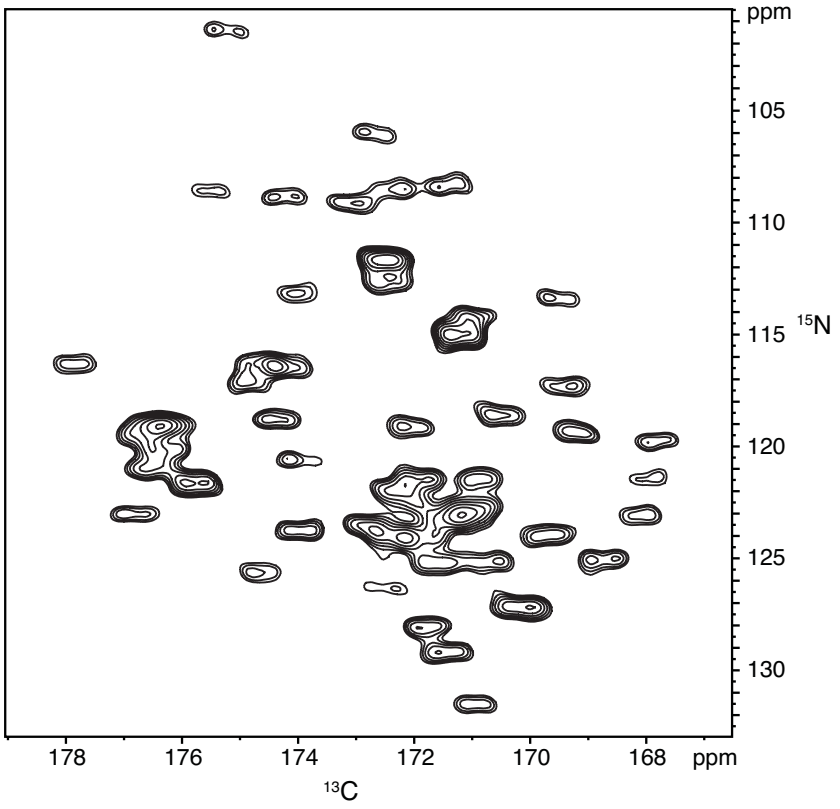
Solid state NMR of proteins at high MAS frequencies: symmetry-based mixing and simultaneous acquisition of chemical shift correlation spectra

By Peter Bellstedt et al.



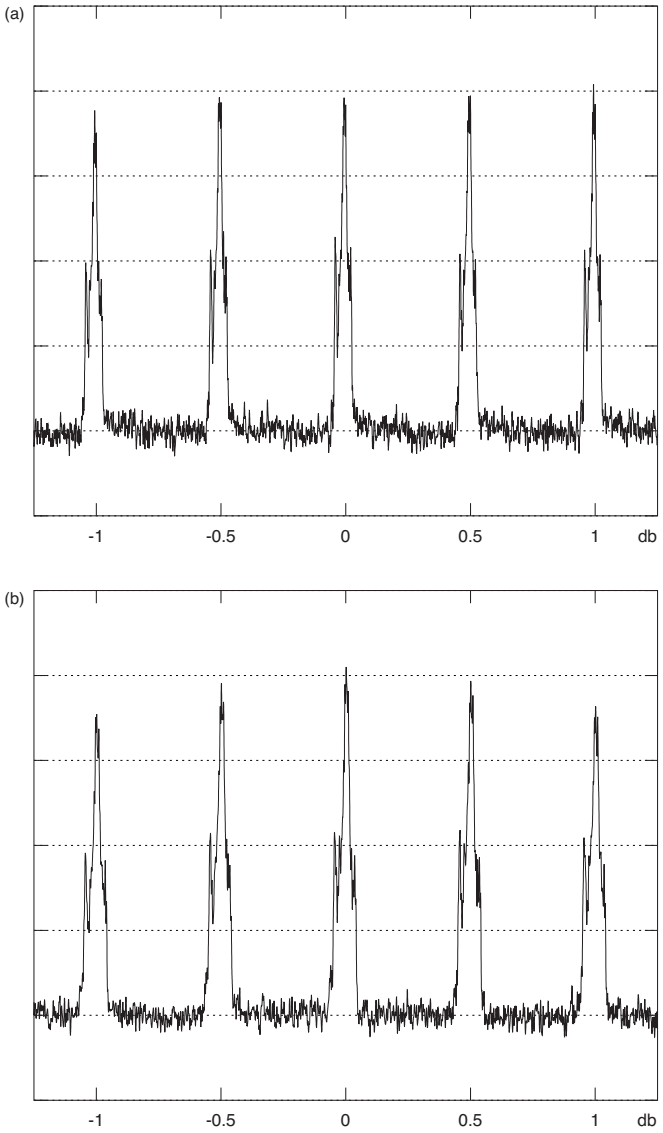
Supporting Figure 1

2D ^{15}N-^{13}CO **(a)** and ^{15}N-^{13}CA **(b)** band-selective chemical shift correlation spectra of the SH3 domain recorded at 500 MHz and a spinning speed of 20 kHz . The $R16_{49}^{-5,4}$ symmetry with the corresponding numerically optimised (scaled) *R* elements, were employed with the ^{13}C RF carrier kept at (a) 175 and (b) 55 ppm and the ^{15}N carrier at 120 ppm. Spectra were acquired using a CP contact time of 1 ms, ^{15}N-^{13}C mixing time of 2.45 ms, ^{13}C and ^{15}N mixing RF field strengths of ~13 kHz, 128 transients per t_1 increment, 64 t_1 increments, spectral width in the indirect dimension of 2500 Hz and a recycle time of 2.0 s were used. Assignments were taken from the literature (Castellani et al. 2003).



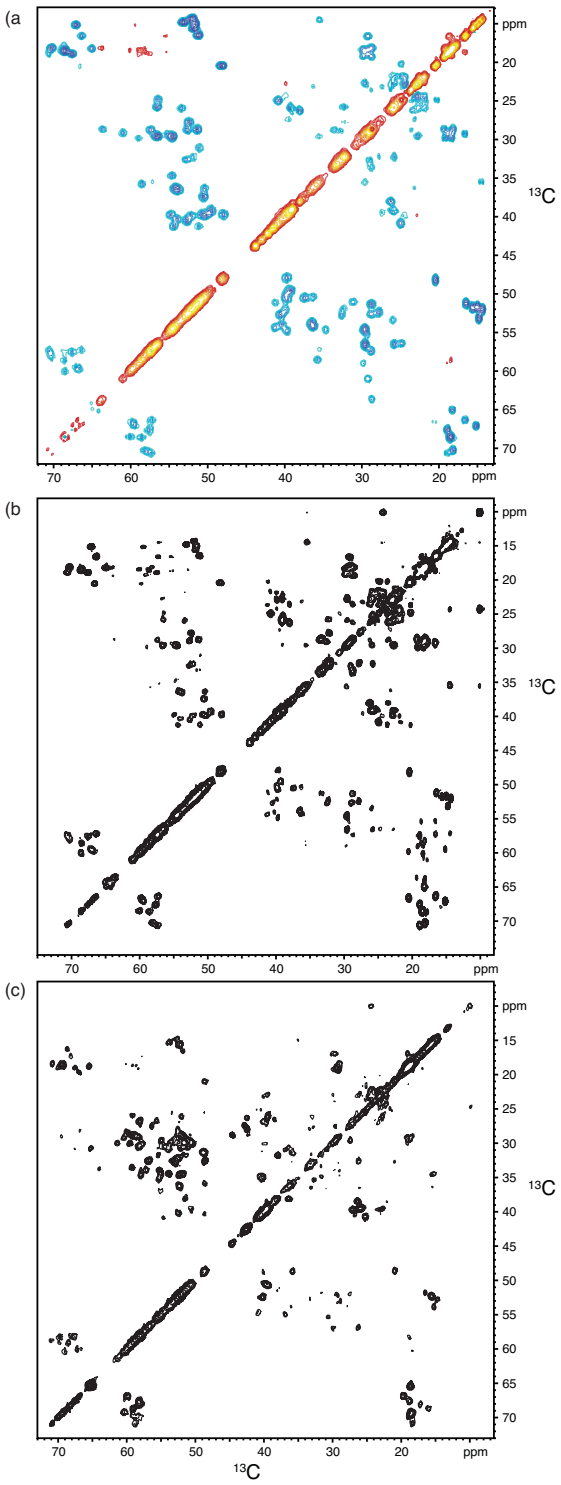
Supporting Figure 2

2D ^{15}N-^{13}CO band-selective chemical shift correlation spectrum of GB1 (crystal form B) recorded at 500 MHz and at a spinning speed of 33.333 kHz. The $R16_{49}^{-5,4}$ symmetry with the corresponding numerically optimised (scaled) *R* elements was used with the ^{13}C and ^{15}N RF carriers kept at 175 ppm and 120 ppm, respectively. The spectrum was acquired using a CP contact time of 0.9 ms, ^{15}N-^{13}C mixing time of 2.94 ms, ^{13}C and ^{15}N mixing RF field strengths of ~22 kHz, 128 transients per t1 increment, 48 t1 increments, spectral width in the indirect dimension of 2026 Hz and a recycle time of 2.0 s.

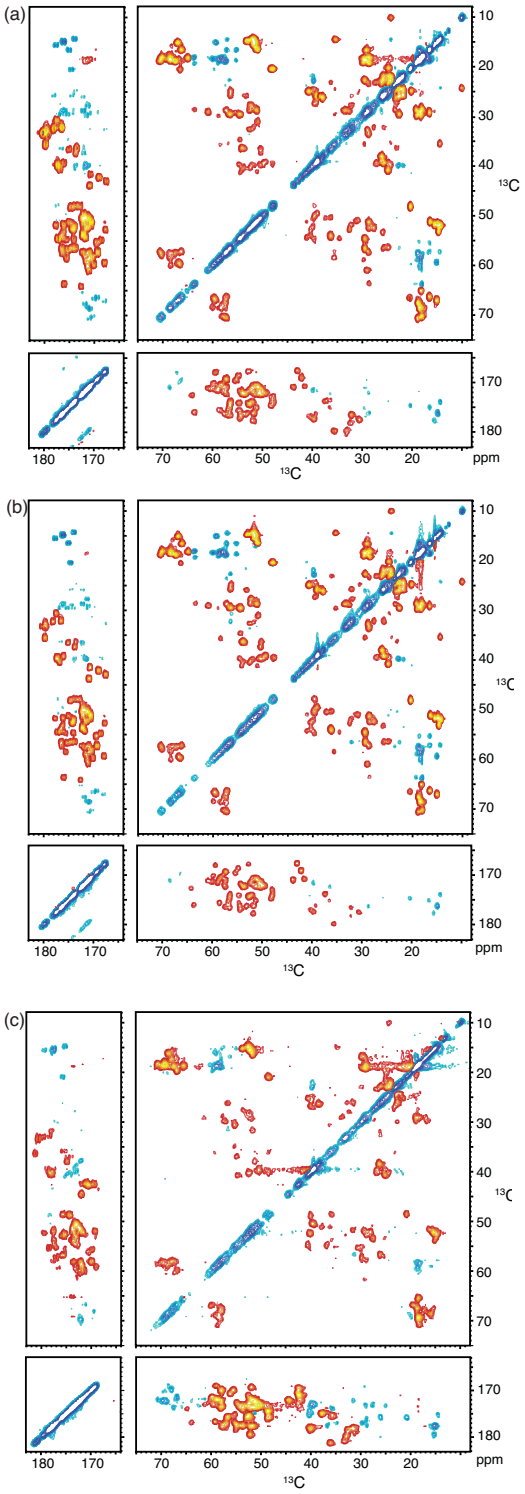


Supporting Figure 3

Effect of 15N **(a)** and 13C **(b)** RF field strength variations on the observed ^{15}N->^{13}CO transfer at 500 MHz. These GB1 (crystal form B) spectra were recorded at a MAS frequency of 33.333 kHz using the optimised $R16_{49}^{-5,4}$ symmetry.



Supporting Figure 4 (a) 2D ^{13}C-^{13}C dipolar coupling mediated band-selective chemical shift correlation spectrum of GB1 (crystal form B) acquired via longitudinal magnetisation exchange at 500 MHz, at a spinning speed of 24 kHz, with ^{13}C RF field strength of ~ 14.4 kHz during mixing and ^{13}C-^{13}C mixing time of 1.25 ms. The spectrum was recorded with the $C7_{30}^{6}$ symmetry-based RF pulse scheme, using the basic *C* element (scaled) reported earlier (Herbst et al. 2011), 64 transients per t_1 increment, 150 t_1 increments, spectral width in the indirect dimension of 9427 Hz, recycle time of 2 s and with the RF carrier kept at 38.5 ppm. Blue and red colours indicate opposite phases of the peaks. **(b)** 2D ^{13}C-^{13}C scalar coupling mediated chemical shift correlation spectrum (zoomed plot) of GB1 (crystal form B) obtained at 500 MHz with a spinning speed of 24 kHz, with ^{13}C RF field strength of ~18.8 kHz during mixing and ^{13}C-^{13}C mixing time of 8.625 ms. The spectrum was recorded using the $C9_{69}^{1}$ symmetry with the scaled basic *C* element reported earlier (Herbst et al. 2011, supplementary material). Other experimental parameters were similar to that in (a). **(c)** 2D ^{13}C-^{13}C scalar coupling mediated chemical shift correlation spectrum of perdeuterated GB1 (NH:ND 10:90, crystal form A) recorded at a spinning speed of 33.333 kHz, with ^{13}C RF field strength of ~15 kHz during mixing and ^{13}C-^{13}C mixing time of 7.2 ms. The spectrum was acquired at 500 MHz with the $C9_{120}^{1}$ symmetry, using the basic *C* element reported earlier (Herbst et al. 2011). CP contact time of 3 ms, 96 transients per t_1 increment, 128 t_1 increments, spectral width in the indirect dimension of 9000 Hz, recycle time of 2 s and with the RF carrier kept at 40 ppm were used without ^{1}H decoupling during mixing.



Supporting Figure 5 (a) 2D ^{13}C-^{13}C broadband dipolar coupling mediated chemical shift correlation spectrum of GB1 (crystal form B) at 500 MHz recorded at a spinning speed of 24 kHz, with ^{13}C mixing RF field strength of ~48 kHz and ^{13}C-^{13}C mixing time of 1.75 ms. The $R16_{14}^{-7}$ symmetry, with the scaled basic *R* element reported earlier (Herbst et al. 2009c), 64 transients per t_1 increment, 350 t_1 increments, spectral width in the indirect dimension of 25140 Hz, recycle time of 2 s and with the RF carrier kept at 100 ppm were used. Blue and red colours indicate opposite phases of the peaks. **(b)** 2D ^{13}C-^{13}C broadband dipolar coupling mediated chemical shift correlation spectrum of GB1 (crystal form B) at 500 MHz recorded at a spinning speed of 24 kHz, with ^{13}C mixing RF field strength of ~43 kHz and ^{13}C-^{13}C mixing time of 1.25 ms. The $C8_{10}^{3}$ symmetry, with the scaled basic element reported earlier (Herbst et al. 2009b), 64 transients per t_1 increment, 350 t_1 increments, spectral width in the indirect dimension of 25140 Hz, recycle time of 2 s and with the RF carrier kept at 100 ppm were employed. Blue and red colours indicate opposite phases of the peaks. **(c)** 2D ^{13}C-^{13}C broadband dipolar coupling mediated chemical shift correlation spectrum of perdeuterated GB1 (crystal form A) recorded at 500 MHz and a spinning speed of 33.333 kHz, with ^{13}C mixing RF field strength of ~60 kHz and ^{13}C-^{13}C mixing time of 1.2 ms. The $C8_{10}^{3}$ symmetry, with the scaled basic element reported earlier (Herbst et al. 2009b), a CP contact time of 3 ms, 64 transients per t_1 increment, 350 t_1 increments, spectral width in the indirect dimension of 25140 Hz, recycle time of 2 s and with the RF carrier kept at 100 ppm were used in the absence of ^{1}H decoupling during mixing. Blue and red colours indicate opposite phases of the peaks.

2.3 Publication 3

Title: Systematic unlabeling of amino acids as useful tool in the NMR resonance assignment of a challenging protein

Authors: **P.Bellstedt**, T.Seiboth, S.Häfner, H.Kutscha, R.Ramachandran, M.Görlach

Contributions: PB conceived of and performed the experiments, analyzed the GB1 data and wrote the manuscript. TS and PB assigned the APTX backbone resonances. SH and PB expressed and purified proteins. HK performed ThermoFluor assay. PB and RR adjusted the pulse sequence. MG supervised the work and corrected the manuscript.

Status: In revision for *Journal of Biomolecular NMR*; Initially submitted: 08.05.2013

Summary: This manuscript describes the specific and individual ^{14}N,^{12}C unlabeling of amino acids to supplement sparse NMR data sets in order to achieve resonance assignment. The model system GB1, for which assignment data is available, was used to validate the efficiency of the approach and to analyze potential interconversion of amino acids ("scrambling") in the standard expression host *E. coli* BL21(DE3). The information obtained were applied to the HIT-ZnF domain and led to an unambiguous assignment of 78% of the observed resonances to their respective backbone nuclei. Missing assignments are due to the non-observability of at least one of the resonances for a given amino acid required for unambiguous data analysis.

Journal of Biomolecular NMR manuscript No.
(will be inserted by the editor)

Systematic unlabeling of amino acids as useful tool in the NMR resonance assignment of a challenging protein

Peter Bellstedt · Thomas Seiboth · Sabine Häfner · Henriette Kutscha · Ramadurai Ramachandran · Matthias Görlach

Received: date / Accepted: date

Abstract NMR structure determination of a protein requires the assignment of resonances as indispensable first step. Even though heteronuclear through-bond correlation methods are available for that purpose, challenging situations may arise in cases where the protein in question only yields samples of limited concentration and/or stability. In such cases it is advantageous to use sensitive NMR experiments with a minimal number of coherence transfers. This, however, may pose the problem of ambiguous assignments already for medium-sized proteins due to the limited dispersion of CA and CO chemical shifts. Here we present a strategy based upon specific individual unlabeling of all 20 standard amino acids in conjunction with an HN(CO) experiment, to achieve unambiguous backbone assignments for two proteins, GB1 and human aprataxin with a "minimalistic" set of NMR experiments. In addition, this approach delivered insight into metabolic interconversion ("scrambling"') of NH and CO groups in a standard *E. coli* expression host.

Keywords resonance assignment · unlabeling · selective isotopic labeling · reverse labeling · aprataxin

1 Introduction

Assignment of NMR resonances and here assignment of backbone nuclei constitutes the essential first step in the process of biomolecular structure determination. Frequently, heteronuclear triple resonance experiments like the HNCACB (Wittekind and Mueller 1993) in combination with the CBCA(CO)NH (Grzesiek and Bax 1992) are performed to identify and assign intra- and interresidue resonances. In addition, approaches for linking adjacent amide residues based on NH(COCA)NH type experiments (Bracken et al 1997; Kumar et al 2010) are available. To ameliorate the inherent sensitivity problem of NMR spectroscopy, improved hardware (higher field strength, cryogenic probes) and new pulse sequences yielding higher s/n ratios (Mori et al 1995; Lescop et al 2007; Felli and Brutscher 2009; Marion 2010) and/or requiring less acquisition time through non-uniform sampling and/or projection reconstruction techniques (Kupče and Freeman 2004; Coggins et al 2010; Qiang 2011) are continuously developed. All these approaches require isotope labeling and, as second prerequisite, a protein sample of relatively high concentration and considerable stability. In cases where stability or concentration of the sample is limited, experiments with a reduced number of coherence transfer steps (*e.g.* HSCQ, HNCO, HNCA), which are inherently more sensitive and, therefore, need shorter recording times, may still be feasible. Such experiments, however, may pose a problem with respect to resonance assignment due to the limited dispersion of CA and CO chemical shifts. This in turn results in potentially ambiguous assignments already for medium sized proteins. Such ambiguity may be resolvable in part by using experiments selective for amino acid type (Schubert et al 1999; Schubert 2001) or the limited information linked to characteristic resonance "regions" (*e.g.* glycine in [^{1}H,^{15}N]-based correlation experiments) together with a given protein sequence as the latter restricts the potential amino acid *i, i+1* pairs. Furthermore, specific isotope labeling in an unlabeled background allows to identify the residue type. Specific isotopic labeling of many

P.Bellstedt · T.Seiboth · S.Häfner · H.Kutscha · R.Ramachandran · M.Görlach
Biomolecular NMR Spectroscopy
Leibniz Institute for Age Research - Fritz Lipmann Institute
Beutenbergstr. 11, 07745 Jena
E-mail: pbell@fli-leibniz.de

amino acids, however, is a costly proposition and may preclude the NMR analysis of a challenging protein altogether. On the other hand, specific incorporation of [^{14}N, ^{12}C]-amino acids in an uniformly labeled background is much less costly and can easily be achieved using standard bacterial systems for heterologous protein expression. Such "unlabeling" approaches have so far been employed sporadically and were restricted to a limited subset of amino acids (Kelly et al 1999; Krishnarjuna et al 2011; Banigan et al 2013). The complicating issue in specific (un)labeling, however, constitutes the scrambling (interconversion) of the label into the intentionally unlabeled amino acids and *vice versa*. Even though genetically modified strains can be used to avoid scrambling (Waugh 1996; O'Grady et al 2012), such bacterial strains are not considered standard recombinant systems and would *e.g.* require re-engineering to allow for T7 polymerase based (Studier and Moffatt 1986) overexpression. Furthermore, genetic engineering of metabolic pathways to minimize scrambling is restricted and affects a limited number of amino acids only (Rasia et al 2012). Virtually scrambling-free synthesis of a target protein may be achievable in cell-free protein synthesis systems (Morita et al 2004; Su et al 2011). Eukaryotic systems have also been used in combination with amino acid selective (un)labeling (Strauss et al 2003; Tanio et al 2009), but frequently suffer from lower yield and higher overall costs. Here we systematically evaluated the unlabeling approach for all 20 natural amino acids and for two recombinant proteins expressed in the standard T7 polymerase based *E. coli* BL21(DE3) system. The first immunoglobulin binding domain of protein G (GB1) served as a model system to assess the feasibility of this strategy. In a second step we successfully applied this approach to the 23 kDa catalytic domain of human aprataxin. Aprataxin mutations are linked to the neurodegenerative disease ataxia with oculomotor apraxia 1 (AOA1; Moreira et al (2001); Date et al (2001)). Subsequent biochemical and cell biological work has shown that aprataxin is involved in DNA repair (Gueven et al 2004; Ahel et al 2006; Rass et al 2007, 2008) and that it might even constitute a novel drug target in colorectal cancer therapy (Dopeso et al 2010). Despite its biomedical significance, the NMR spectroscopic analysis of human aprataxin was severely hampered by its propensity to denature in short time and to aggregate already at low protein concentrations. Based on our systematic unlabeling approach, and by including the information of the amino acid type(s) during the assignment process, here we present an assignment strategy based on linking CO and CA resonances and a partial backbone resonance assignment for aprataxin's catalytically active domain in order to provide vital chemical shift information for *e.g.* NMR-based compound screening.

2 Materials and methods

2.1 Expression and purification of GB1

Uniformly [^{15}N, ^{13}C]-labeled GB1 (T2Q mutant) was expressed in *E. coli* BL21(DE3) and M9 media containing 1 g/l $^{15}NH_4Cl$ and 2 g/l [^{1}H, ^{13}C]-glucose. For selectively unlabeled samples of GB1, 1 g/l (except for Cys: 0.1 g/l) of the respective [^{14}N, ^{12}C]-amino acid (Sigma Aldrich) was added to the medium 15 min prior to induction with 0.3 mM IPTG for 3 h at 30°C. Based on published protocols (Franks et al 2005) the harvested cells were disrupted by heating to 80 °C for 15 min in phosphate-buffered saline (200 mM NaCl, 50 mM KH_2PO_4/K_2HPO_4, pH 7), debris was removed by centrifugation (16.000×g, 4°C, 30 min) and the supernatant was subjected to size exclusion chromatography (Sephadex 75, GE Heathcare; 100 mM NaCl, 50 mM KH_2PO_4/K_2HPO_4, pH 7). GB1 containing fractions were combined and supplemented with 10% D_2O concentrated to 500 μM using a Vivaspin 20 concentrator (3.5 kDa cut-off, GE Healthcare).

2.2 Expression and purification of aprataxin

The pET-15b expression plasmid coding for residues 161-356 of human aprataxin (UniProt id: Q7Z2E3-1), comprising its enzymatically active histidine triad (HIT) domain and zinc finger motif, was transformed into *E. coli* BL21(DE3). The His_6-tagged protein was expressed in M9 media supplemented with 1 g/l $^{15}NH_4Cl$ and 1.5 g/l [^{1}H, ^{13}C]-glucose and a final concentration of 10 μM $ZnSO_4$. Unlabeling and protein expression were induced as for GB1 with 0.3 mM IPTG for 3 h at 30°C. Harvested cells were disrupted by French press and ultrasonification, debris was removed by centrifugation (15.400×g, 4°C, 30 min) and the supernatant was subjected to a affinity chromatography on Ni-NTA agarose (Qiagen) followed by thrombin cleavage and dialysis (50 mM NaCl, 10 mM Tris/HCl pH 7.5, 4°C over night). Pure protein was obtained as flow through of the final anion exchange chromatography (DEAE fast-flow sepharose; GE Heathcare). The final NMR samples were concentrated and exchanged into NMR buffer (150 mM NaCl, 10 mM dTris/HCl pH 7.5, 10% D_2O) using a Vivaspin 20 concentrator (10 kDa cut-off, GE Healthcare). A scheme of the purification process is given in Figure S4.

2.3 NMR experiments

Chemical shift correlation experiments were performed on a Bruker 750 MHz Avance III NMR system equipped with a triple resonance probe. Sample temperature was set to 25°C (GB1) or 29°C (aprataxin), respectively. For backbone resonance assignment of aprataxin 3D HNCO, 3D HN(CA)CO, 3D HNCA and 3D HN(CO)CA were collected with freshly prepared 0.5 mM samples for each of the mentioned experiments. The protein concentration after data collection was below 0.1 mM. 3D HNCO and 3D HN(CA)CO data for GB1 were obtained from a 2.3 mM uniformly [^{15}N,^{13}C]-labeled sample.

Unlabeling: [^{1}H,^{15}N]-HSQC spectra were recorded using a standard sequence (Bruker: "hsqcfpf3gpphwg"). [^{13}CO]-edited [^{1}H,^{15}N]-correlation spectra were collected as 2D versions of a 3D HNCO experiment (Bruker: "hncogpwg3d"). To achieve higher resolution in t2, normally restricted by the constant time character, this pulse sequence was modified to allow for free evolution in t2. Water suppression was achieved via presaturation (Jesson et al 1973). The modified HNCO sequence is referred to as HN(CO) here.

2.4 Data analysis

Amide resonance assignments of GB1 are based on literature data (Franks et al 2005) and were reproduced with the CCPN software package (Vranken et al 2005) using 3D HNCO and 3D HN(CA)CO data. Backbone resonances of aprataxin were analyzed using the same software package and 3D HNCO, 3D HN(CA)CO, 3D HNCA and 3D HN(CO)CA data. Unlabeling: Absolute peak intensities were extracted from [^{1}H,^{15}N]-HSQC and HN(CO) data using TOPSPIN V2.1. All spectra were recorded and processed identically as described in the respective figure legends (Figure S1, S2, S7).

3 Results and discussion

In an ideal case, the incorporation of a specific [^{14}N,^{12}C]-amino acid in an otherwise uniformly [^{15}N, ^{13}C]-labeled protein renders the respective nuclei unobservable in heteronuclear chemical shift correlation experiments. By comparing [^{1}H,^{15}N]-HSQC data of a uniformly labeled sample with data of samples where only one individual amino acid was present in its unlabeled form during expression, one can link each of the amide signals to the respective amino acid type (example shown in Figure 1a). A ^{13}CO editing step was used to additionally eliminate amide signals arising from *i+1* ^{15}N amide directly following a [^{14}N, ^{12}C]-amino acid (Figure 1 b). The main prerequisite for a straightforward analysis is that the respective [^{14}N,^{12}C]-amino acid used for unlabeling is not interconverted to other amino acids (also referred to as "isotope scrambling"), which obviously influences the (un)labeling pattern. However, in biological systems used for heterologous protein expression (*e.g. E. coli*), this precondition is met only for some amino acids (Muchmore et al 1989). The amount of isotope scrambling strongly depends on the type of amino acid and on the chemical group to be looked at (*e.g.* NH, CO, CA, side chain).

unlabeled target only	target + one more	distinct group	uniform scrambling
Arg *	Ala	Ile, Leu, Val	Asp
Asn	Gly	Phe, Tyr	Glu
Cys*		Thr	Ser*
Gln			Trp
His*			
Lys			
Met			

Table 1 Classification of amino acids with respect to amount of ^{14}N scrambling based on the analysis of [^{1}H,^{15}N]-HSQC peak intensities for GB1 (Figure S1). *: respective amino acid is not present in GB1, but unlabeling data was assessed for scrambling with respect to other amino acid types.

Here, we first systematically evaluate the amount of ^{14}NH and ^{12}CO scrambling for a number of amino acids for the *E. coli* BL21(DE3) standard expression system and GB1 as a model protein. Secondly we present an approach for sequential resonance assignment applied to the 23 kDa catalytic domain of human aprataxin, for which essentially only HN, CO and CA chemical shift information could be obtained.

3.1 Classification of amino acids based on amount of ^{14}N scrambling

We analyzed the levels of ^{14}N scrambling for each of the 20 standard amino acids. We compared the signal intensity of each of the amide signals of uniformly labeled GB1 with the signal intensity of the respective peak in each of the 20 samples with a single [^{14}N,^{12}C]-amino acid added prior to protein expression (referred to as unlabeled samples). Peak intensities were extracted from [^{1}H,^{15}N]-HSQC data obtained with equal protein concentration, equal experimental and processing parameters. Peak intensities for the same type of amino acid were grouped and averaged to analyze if and how unlabeling of a respective amino acid also effects other ("undesired") amino acids (Figure S1). Based on the resulting normalized, grouped and averaged ^{15}N peak intensities we classified each amino acid with respect to the amount of amide group scrambling (Table 1).

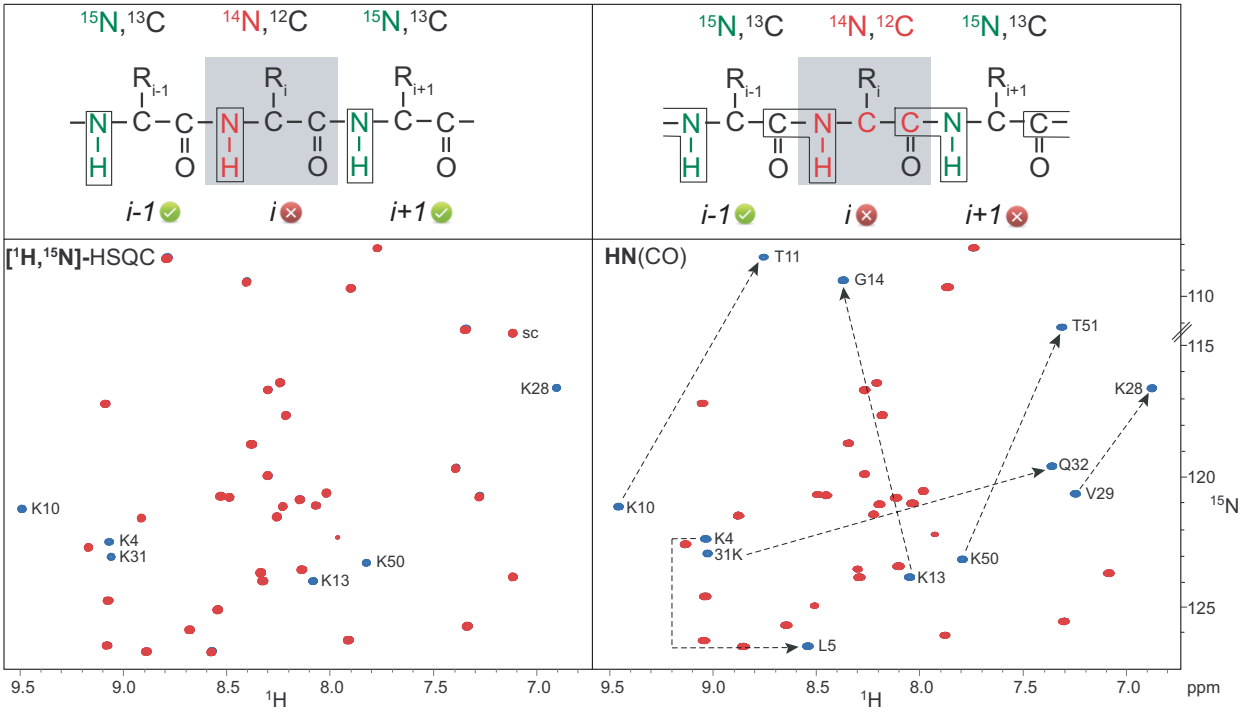


Fig. 1 Identification of unlabeled amino acid i via [^{1}H,^{15}N]-HSQC and of i and $i+1$ via HN(CO) using [^{14}N, ^{12}C]-lysine in otherwise uniformly [^{15}N,^{13}C]-labeled GB1 protein. Resonances of the reference spectrum are colored in blue, resonances from the specific [^{14}N, ^{12}C]-lysine sample in red. Missing amide signals in the [^{1}H,^{15}N]-HSQC result from specific unlabeling of lysine amide groups. In the HN(CO) spectra also amide groups are not observable, which directly follow a lysine. The respective i and $i+1$ pairs are indicated with dotted arrows. Assignment is based on standard 3D NMR data (as stated in materials and method). sc: side chain, not observable in HN(CO) spectra.

Arg, Cys, Ser and His are not present in GB1, but were used for unlabeling to analyze the effect to other amino amino acids. For 7 out of the 20 amino acids tested (Table 1, 1^{st} column), no scrambling for amide groups was found, allowing a straightforward and unambiguous identification with the outlined unlabel approach. Specific unlabeling of Ala and Gly, respectively, additionally results in a approx. 50 % signal reduction for Val and Trp, respectively (Figure S1). Since the reduction of signal intensity is significantly different (complete loss of signal for Ala and Gly *versus* approx. 40-50% reduction in signal intensity for Val and Trp), this finding can safely be exploited to identify both of the respective amino acids using one singly unlabeled protein sample only. In fact, in the case of the [^{14}N]-Gly sample where also Trp can be determined this is quite important since specific unlabeling of Trp itself leads to uniform scrambling which renders identification of the targeted amide impossible. It is worth emphasizing, that although unlabeling of Gln leads to a general scrambling (Figure S1), the intensity drop is significantly greater for Gln amide signals as compared to the all other amino acids. Therefore although Gln is considered to be metabolized by the bacteria, in our hands a quantitative analysis allowed a clear identification of Gln amide signals in GB1. The amide groups of 6 out of 20 amino acids (Table 1, 3^{rd} column) are scrambled among a smaller group of other amino acids and, therefore, only permit the identification of a distinct group of amino acids instead of a single one. Amide signal intensities of Ile, Leu and Val are reduced to the same extent irrespective of which of the three amino acids was present in unlabeled form during protein expression (Figure S1). Likewise, the amide groups of Phe and Tyr are scrambled among each other. Although a clear identification of the desired and originally unlabeled amino acid within the latter two groups is not possible, information from that spectra is quite useful to narrow down the possible amino acid types. This in turn can be essential if used in conjunction with sequence information to resolve ambiguities during the assignment process. Adding unlabeled Asp, Glu, Ser and Trp (Table 1, 4^{th} column) results in equal reduction in signal intensity indicative of uniform scrambling and precluding spectral identification (Figure S1).

no scrambling	some scrambling	not analysed[1]
Arg*, Cys*	Ala, Gly	Asp, Asn
His*, Ile	Thr, Pro*	Gln, Glu
Leu, Lys	Ser*	Trp
Met, Phe		
Tyr, Val		

Table 2 Classification of amino acids with respect to amount of ^{12}CO scrambling based on the analysis of HN(CO) peak intensities for GB1 (Figure S2). [1]: not analysed, see text for explanation. *: respective amino acid is not present in GB1, but unlabeling data was assessed for scrambling with respect to other amino acid types.

3.2 Classification of amino acids based on amount of ^{12}CO scrambling

A labeling scheme for specific ^{13}CO *labeling* of amino acids has already been presented by Takeuchi et al. However, from the experimental setup and the biochemical point of view specific *labeling* is different from specific *unlabeling*. Unlabeling is achieved by adding the desired [^{14}N,^{12}C]-amino acid only to the expression media. In contrast, for specific *labeling* mostly not only the targeted amino acid (or a precursor) is added to the expression media in its labeled form, but also the other amino acids in their unlabeled form to minimize scrambling through influencing regulation of the amino acid metabolism. To analyze the effect of specific *unlabeling* on the labeling state of carbonyls among the different amino acids, we compared the peak intensities in HN(CO) data recorded with a fully labeled reference sample with data recorded with 20 differently unlabeled samples in the same way as described for the analysis of amide scrambling. Since the resonance intensity of the HN(CO) is not only affected by ^{12}C incorporation in the carbonyl positions, but also by the amide of the following amino acid, a perfect analysis would require an unlabeling approach with [^{15}N, ^{12}C]-amino acids, which is far beyond the intent of this study to provide a cost effective way to obtain additional information for the assignment process. For this reason we only analyzed carbonyls of amino acids *i*, which are followed by an *i+1* amide group, of which signal intensity reduction was less than 20% as determined from [^{1}H,^{15}N]-HSQC data. With that restriction we were able to evaluate 15 out of 20 amino acids (all except Asp, Asn, Gln, Glu, Trp) with respect to CO scrambling (Figure S2, summary listed in Table 2). Most of the amino acids analyzed are not prone to ^{12}CO scrambling confirming the observation (Takeuchi et al 2007), that in general, carbonyls are less affected by scrambling than amide groups. Whereas *e.g.* the amide group of Ile, Leu and Val scrambles among these 3 amino acids (Table 1, 3rd column, Figure S1), the carbonyls are *not* affected (Figure S2), permitting unambiguous identification of the following *i+1* amide in the HN(CO) spectrum.

3.3 Backbone resonance assignment strategy

We have shown that by quantitative analysis of peak intensities for [^{1}H,^{15}N]-HSQC and HN(CO) spectra in combination with amino acid-selective unlabeling, in many cases one can extract information about the amino acid type(s) of the respective resonance. Likewise, it is possible to identify peaks that originate from *i+1* amide groups directly following an unlabeled amino acid *i* by analysis of HN(CO) spectra. In cases where concentration and/or stability of a protein in question restricts the NMR data collection to experiments with a minimal number of coherence transfer steps, additional information has to be included to unambiguously assign these backbone resonances. Here we applied this strategy to the backbone resonance assignment of human aprataxin, the accessibility of which for NMR structure determination is severely limited by its aggregation propensity even after buffer optimization (Figure S5). Therefore, the residue specific assignment process of aprataxin was essentially based on NH, CO and CA chemical shifts only. However, owing to limited spectral dispersion and the resulting overlap, it was essential to employ the unlabeling approach to achieve that unambiguous backbone resonance assignment for this 23 kDa protein domain.

The strategy to include information obtained via amino acid selective unlabeling during the assignment process is presented in Figure 2 and consists of 4 principal steps. *Step 1:* prediction of amino acid type of the query amino acid *i* via analysis of [^{1}H,^{15}N]-HSQC spectra of reference and specifically unlabeled samples, *step 2:* prediction of amino acid type of potentially matching *i-1* amino acids in the same way, *step 3:* check protein sequence if combination of *i* and *i-1* occurs, if not exclude respective *i-1*; *step 4:* cross check in HN(CO) of unlabeled *i-1* candidate as predicted in step 2, if the NH of *i* is vanished as consequence of ^{13}C editing. Where available additional HNCA and sparse HNCACB data were included in the analysis as well. Following our approach, 78 % of the backbone nuclei of the the catalytic domain of human aprataxin could be assigned successfully (Table S2, Figure S6). Missing assignments result from non-observability of at least one of the resonances for a given amino acid required for unambiguous data analysis. Together with our observation of aprataxin's propensity to aggregate and to precipitate we attribute this to the highly unfavorable dynamics of the protein

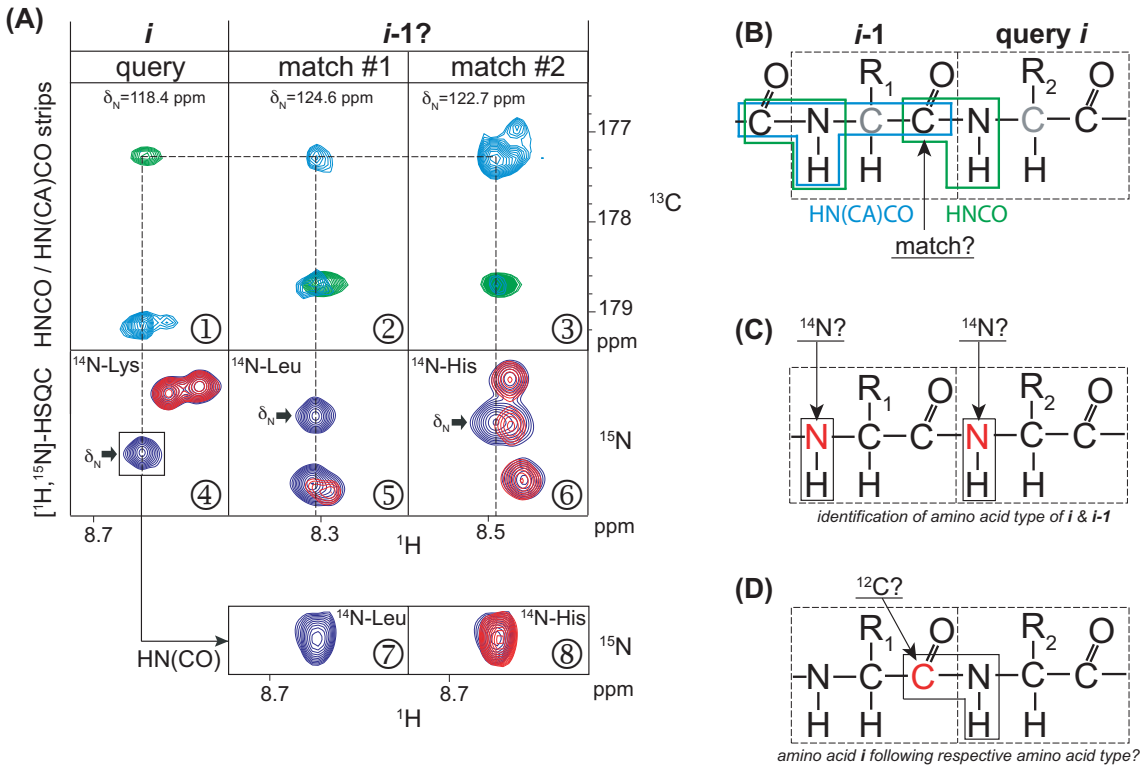


Fig. 2 Assignment strategy for apratxin supported by amino acid type information obtained from specific unlabeling. This strategy is illustrated here for one assignment step: one "query" amino acid *i* **(A, box 1)** could be linked to different potential *i-1* "matches" with equally well corresponding carbon chemical shifts **(A, box 2 and 3)**. Resonances in light blue represent HN(CA)CO, resonances in green HNCO data. HN(CA)CO is *i, i-1* specific, HNCO *i-1* specific (see scheme in **B**). The 20 different specifically unlabeled samples of apratxin have to be evaluated together with the sequence information to predict the amino acid type of the chosen query amino acid *i*. Here, the amide peak intensity is reduced only in the [^{14}N]-Lys sample **(A, box 4)** and, therefore, the chosen query amino acid *i* is most probably a lysine. Blue contours represent data from the fully labeled reference sample, contours in red data from the unlabeled sample which were recorded and processed identically. Next, the amino acid types of the potentially matching amino acids have to be identified in the same way (scheme in **C**). In **A, box 5 and 6** show the region of [^{15}N]-HSQC spectra where the reduction of signal intensity is the highest observed for the amide group to be evaluated in all 20 unlabeled samples. The particular amino acid used for the respective unlabeling is indicated inside the panels. Potentially matching amino acid #1 is most probably a leucine **(A, box 5)**, matching amino acid #2 a histidine **(A, box 6)**. The obtained information about amino acid type of the query (*i*) as well as potentially matching amino acids (*i-1*) are used to assess if combinations of *i* and *i-1* are present in the protein sequence. This is the first step, at which *i-1* candidates can be excluded. Here, based on the protein sequence (Figure S3 b), His can be excluded from further consideration as *i-1* candidate, since no His-Lys combination is present in the catalytic domain of apratxin. Finally, the HN(CO) spectrum for the specific unlabeled amino acid is used to cross check if the initially chosen query *i* is following a Leu residue (scheme in **D**). Only in the [^{14}N]-Leu sample there is a complete reduction in signal intensity in the HN(CO) spectrum **(A, box 7** *versus* **box 8)**, indicating that the chosen *i* is most probably adjacent to a Leu and therefore the query should be linked to matching amino acid #1, delivering an unambiguously assigned Leu-Lys fragment.

backbone. The assigned chemical shifts have been deposited in the BMRB (BioMagResBank (Ulrich et al 2007)) under accession number 19182.

As, in contrast to GB1 (Table S1), each of the 20 natural amino acids is present in apratxin, we used this work to provide for a more detailed analysis of ^{14}N scrambling. As expected, the labeling scheme of amide groups following amino acid-type specific unlabeling of apratxin (Figure S7) are in agreement with the results extracted from the GB1 data. Strikingly, specific [^{14}N]-Gly unlabeling leads also to complete unlabeling of Cys and Ser amide groups not present in GB1. This nicely illustrates the underlying bacterial amino acid metabolism (Figure S8).

In summary we provide a detailed analysis of ^{14}N and ^{12}CO scrambling using *E. coli* BL21(DE3) as a standard system for heterologous protein expression. In addition, we present a cost efficient strategy to include the information of amino acid type(s) during the assignment process to resolve potential ambiguities if analysis is restricted to a "minimalistic" NMR data set.

References

Ahel I, Rass U, El-Khamisy SF, Katyal S, Clements PM, Mckinnon PJ, Caldecott KW, West SC (2006) The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates. Nature 443(7112):713–716

Banigan JR, Gayen A, Traaseth NJ (2013) Combination of 15N reverse labeling and afterglow spectroscopy for assigning membrane protein spectra by magic-angle-spinning solid-state NMR: application to the multidrug resistance protein EmrE. Journal of Biomolecular NMR 55(4):391–399

Bracken C, Palmer AG, Cavanagh J (1997) (H)N(COCA)NH and HN(COCA)NH experiments for 1H-15N backbone assignments in 13C/15N-labeled proteins. Journal of Biomolecular NMR 9(1):94–100

Coggins BE, Venters RA, Zhou P (2010) Radial sampling for fast NMR: Concepts and practices over three decades. Progress in Nuclear Magnetic Resonance Spectroscopy 57(4):381–419

Date H, Onodera O, Tanaka H, Iwabuchi K, Uekawa K, Igarashi S, Koike R, Hiroi T, Yuasa T, Awaya Y, Sakai T, Takahashi T, Nagatomo H, Sekijima Y, Kawachi I, Takiyama Y, Nishizawa M, Fukuhara N, Saito K, Sugano S, Tsuji S (2001) Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. Nature genetics 29(2):184–188

Dopeso H, Mateo-Lozano S, Elez E, Landolfi S, Ramos Pascual FJ, Hernández-Losa J, Mazzolini R, Rodrigues P, Bazzocco S, Carreras MJ, Espín E, Armengol M, Wilson AJ, Mariadason JM, Ramon Y Cajal S, Tabernero J, Schwartz S, Arango D (2010) Aprataxin tumor levels predict response of colorectal cancer patients to irinotecan-based treatment. Clinical cancer research : an official journal of the American Association for Cancer Research 16(8):2375–2382

Felli IC, Brutscher B (2009) Recent advances in solution NMR: fast methods and heteronuclear direct detection. ChemPhysChem 10(9-10):1356–1368

Franks WT, Zhou DH, Wylie BJ, Money BG, Graesser DT, Frericks HL, Sahota G, Rienstra CM (2005) Magic-Angle Spinning Solid-State NMR Spectroscopy of the 1 Immunoglobulin Binding Domain of Protein G (GB1): 15N and 13C Chemical Shift Assignments and Conformational Analysis. J Am Chem SOC 127(35):12,291–12,305

Grzesiek S, Bax A (1992) Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. Journal of the American Chemical Society 114(16):6291–6293

Gueven N, Becherel OJ, Kijas AW, Chen P, Howe O, Rudolph JH, Gatti R, Date H, Onodera O, Taucher-Scholz G, Lavin MF (2004) Aprataxin, a novel protein that protects against genotoxic stress. Human molecular genetics 13(10):1081–1093

Jesson JP, Meakin P, Kneissel G (1973) Homonuclear decoupling and peak elimination in Fourier transform nuclear magnetic resonance. Journal of the American Chemical Society 95(2):618–620

Kelly MJ, Krieger C, Ball LJ, Yu Y, Richter G, Schmieder P, Bacher A, Oschkinat H (1999) Application of amino acid type-specific 1H- and 14N-labeling in a 2H-, 15N-labeled background to a 47 kDa homodimer: potential for NMR structure determination of large proteins. Journal of Biomolecular NMR 14(1):79–83

Krishnarjuna B, Jaipuria G, Thakur A, D'Silva P, Atreya HS (2011) Amino acid selective unlabeling for sequence specific resonance assignments in proteins. Journal of Biomolecular NMR 49(1):39–51

Kumar D, Paul S, Hosur RV (2010) BEST-HNN and 2D-(HN)NH experiments for rapid backbone assignment in proteins. Journal of magnetic resonance (San Diego, Calif : 1997) 204(1):111–117

Kupče E, Freeman R (2004) ProjectionReconstruction Technique for Speeding up Multidimensional NMR Spectroscopy. J Am Chem SOC 126(20):6429–6440

Lescop E, Schanda P, Brutscher B (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. Journal of magnetic resonance (San Diego, Calif : 1997) 187(1):163–169

Marion D (2010) ScienceDirect - Journal of Magnetic Resonance : Combining methods for speeding up multidimensional acquisition. Sparse sampling and fast pulsing methods for unfolded proteins. Journal of Magnetic Resonance

Moreira MC, Barbot C, Tachi N, Kozuka N, Uchida E, Gibson T, Mendonça P, Costa M, Barros J, Yanagisawa T, Watanabe M, Ikeda Y, Aoki M, Nagata T, Coutinho P, Sequeiros J, Koenig M (2001) The gene mutated in ataxia-ocular apraxia 1 encodes the new HIT/Zn-finger protein aprataxin. Nature genetics 29(2):189–193

Mori S, Abeygunawardana C, Johnson MO, Vanzijl PCM (1995) Improved Sensitivity of HSQC Spectra of Exchanging Protons at Short Interscan Delays Using a New Fast HSQC (FHSQC) Detection Scheme That Avoids Water Saturation. Journal of Magnetic Resonance, Series B 108(1):94–98

Morita EH, Shimizu M, Ogasawara T, Endo Y, Tanaka R, Kohno T (2004) A novel way of amino acid-specific assignment in 1H- 15N HSQC spectra with a wheat germ cell-free protein synthesis system. Journal of Biomolecular NMR 30(1):37–45

Muchmore DC, McIntosh LP, Russell CB, Anderson DE, Dahlquist FW (1989) Expression and nitrogen-15 labeling of proteins for proton and nitrogen-15 nuclear magnetic resonance. In: Methods in . . . , Elsevier, pp 44–73

O'Grady C, Rempel BL, Sokaribo A, Nokhrin S, Dmitriev OY (2012) One-step amino acid selective isotope labeling of proteins in prototrophic Escherichia coli strains. Analytical Biochemistry 426(2):126–128

Qiang W (2011) Signal enhancement for the sensitivity-limited solid state NMR experiments using a continuous, non-uniform acquisition scheme. Journal of magnetic resonance (San Diego, Calif : 1997) 213(1):171–175

Rasia RM, Brutscher B, Plevin MJ (2012) Selective Isotopic Unlabeling of Proteins Using Metabolic Precursors: Application to NMR Assignment of Intrinsically Disordered Proteins. Chembiochem : a European journal of chemical biology 13(5):732–739

Rass U, Ahel I, West SC (2007) Actions of Aprataxin in Multiple DNA Repair Pathways. Journal of Biological Chemistry 282(13):9469–9474

Rass U, Ahel I, West SC (2008) Molecular mechanism of DNA deadenylation by the neurological disease protein aprataxin. The Journal of biological chemistry 283(49):33,994–34,001

Schubert M (2001) MUSIC, Selective Pulses, and Tuned Delays: Amino Acid Type-Selective 1H–15N Correlations, II. Journal of Magnetic Resonance 148(1):61–72

Schubert M, Smalla M, Schmieder P, Oschkinat H (1999) MUSIC in triple-resonance experiments: amino acid type-

selective (1)H-(15)N correlations. Journal of magnetic resonance (San Diego, Calif : 1997) 141(1):34–43

Strauss A, Bitsch F, Cutting B, Fendrich G, Graff P, Liebetanz J, Zurini M, Jahnke W (2003) Amino-acid-type selective isotope labeling of proteins expressed in Baculovirus-infected insect cells useful for NMR studies. Journal of Biomolecular NMR 26(4):367–372

Studier F, Moffatt B (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. Journal of molecular biology 189(1):113–130

Su XC, Loh CT, Qi R, Otting G (2011) Suppression of isotope scrambling in cell-free protein synthesis by broadband inhibition of PLP enymes for selective 15N-labelling and production of perdeuterated proteins in H2O. Journal of Biomolecular NMR 50(1):35–42

Takeuchi K, Ng E, Malia TJ, Wagner G (2007) 1-13C amino acid selective labeling in a 2H15N background for NMR studies of large proteins. Journal of Biomolecular NMR 38(1):89–98

Tanio M, Tanaka R, Tanaka T, Kohno T (2009) Amino acid-selective isotope labeling of proteins for nuclear magnetic resonance study: Proteins secreted by Brevibacillus choshinensis. Analytical Biochemistry 386(2):156–160

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2007) BioMagResBank. Nucleic Acids Research 36(Database):D402–D408

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins: Structure, Function, and Bioinformatics 59(4):687–696

Waugh D (1996) Genetic tools for selective labeling of proteins with ?-15N-amino acids. Journal of Biomolecular NMR 8(2):184–192

Wittekind M, Mueller L (1993) HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha-and beta-carbon resonances in proteins. Journal of Magnetic Resonance, Series B 101(2):201–205

Supplementary material

for the manuscript

Systematic unlabeling of amino acids as useful tool in the NMR resonance assignment of a challenging protein

By Peter Bellstedt et al.

Table S1 Top: Sequence information for GB1 (Swiss Prot accession P19909), residues 303-357, with mutation T2Q. Bottom: Number of residues grouped by amino acid type. Amino acids labeled in red are not present in GB1, for Met the amide is not detectable.

```
         10          20          30          40          50
MQYKLILNGK TLKGETTTEA VDAATAEKVF KQYANDNGVD GEWTYDDATK TFTVTE
```

Amino acid	Number of residues
Ala	6
Arg	0
Asn	3
Asp	5
Cys	0
Gln	2
Glu	5
Gly	4
His	0
Ile	1
Leu	3
Lys	6
Met	1
Phe	2
Pro	0
Ser	0
Thr	10
Trp	1
Tyr	3
Val	4

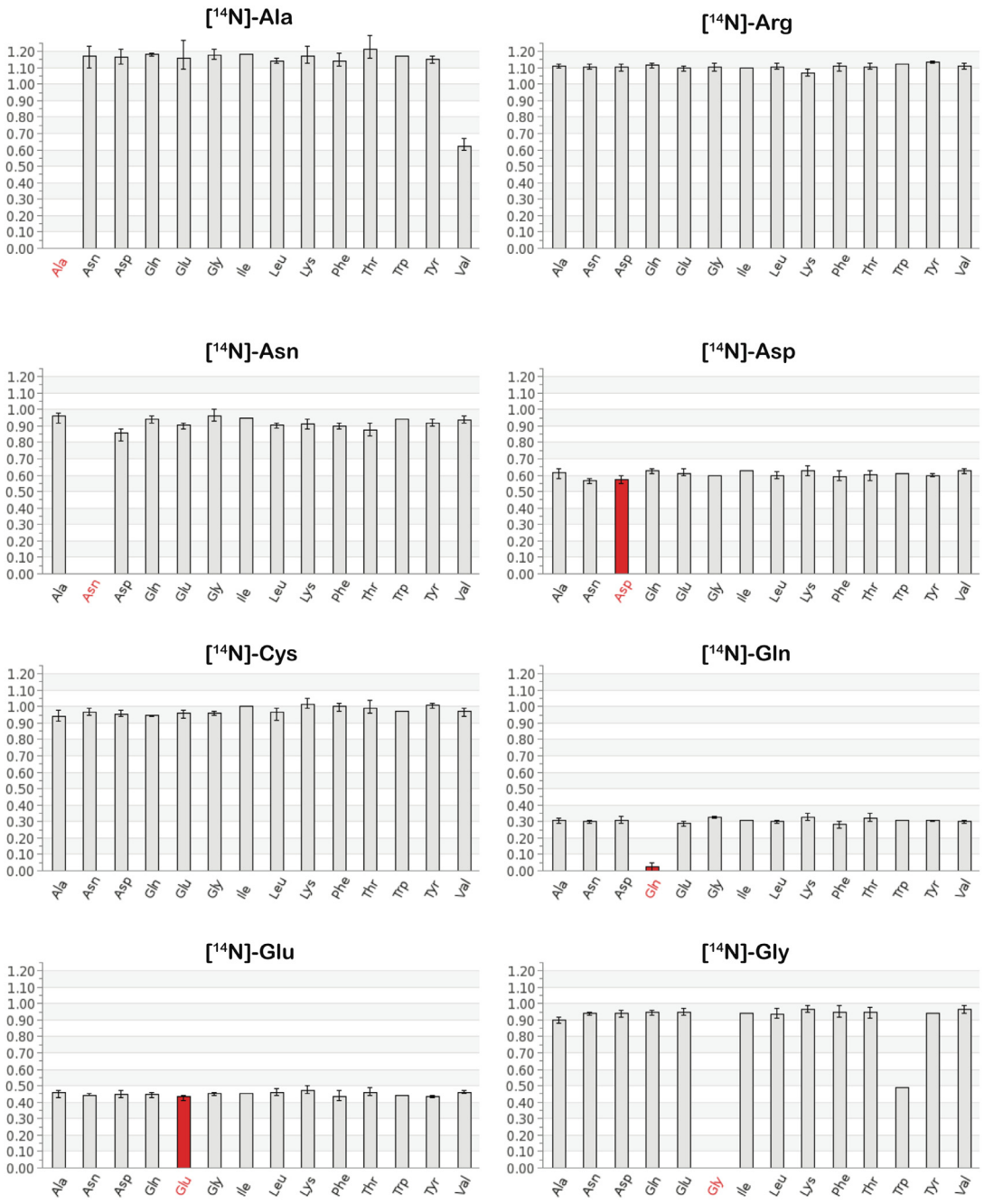


Figure S1 [^{1}H,^{15}N]-HSQC peak intensities of GB1 following unlabeling. GB1 was expressed in M9 media containing $^{15}NH_4Cl$ and [^{13}C]-glucose in the presence of unlabeled [^{14}N,^{12}C]-amino acids as indicated on top of the bar graph panels. The intensities of each individual peak was extracted from the respective [^{1}H,^{15}N]-HSQC and normalized against the corresponding peak intensity extracted from the reference [^{1}H,^{15}N]-HSQC of the fully labeled sample. Peak intensities for individual signals of each amino acid type were averaged (bar graph). Error bars indicate min and max values, respectively. Even though

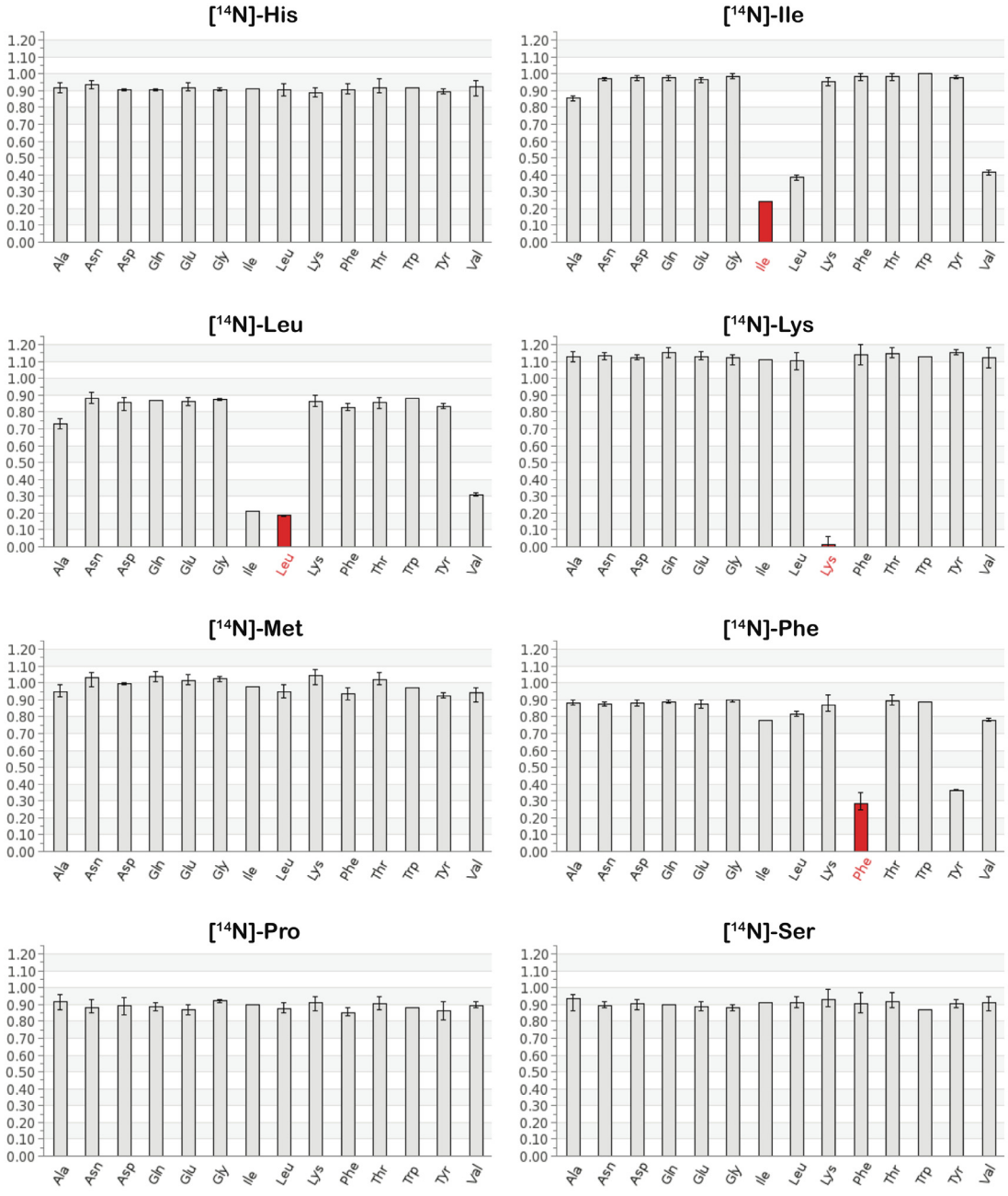


Figure S1 (continued)

Arg, Cys, His, Ser are not present in the sequence, unlabeling with these allows for assessment of scrambling with respect to other amino acid types. Spectra were recorded identically at 750 MHz, 25°C, with a sample concentration of 0.5 mM using 2048 t_2 increments; 512 t_1 increments; a spectral width of 12019 Hz (t_2), 2812 Hz (t_1); 4 transients per t_1 increment; keeping the ^{1}H and ^{15}N RF carrier at 4.7 ppm and 118.5 ppm, respectively. Recorded spectra were processed identically and the absolute peak intensities were extracted using TOPSPIN V2.1. Figure continued on next page.

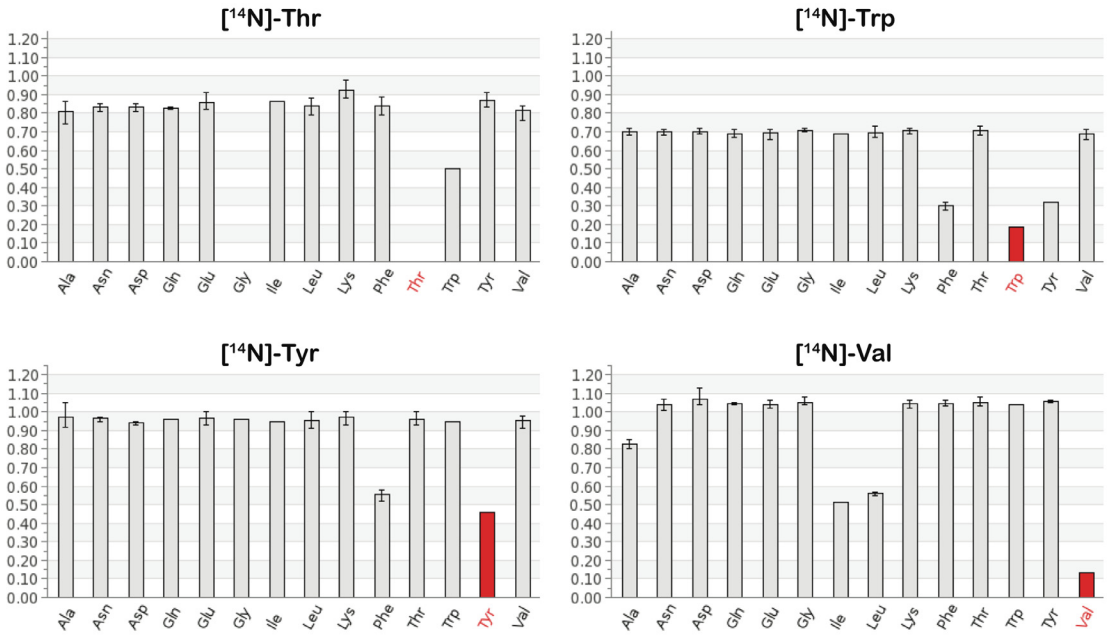


Figure S1 (continued)

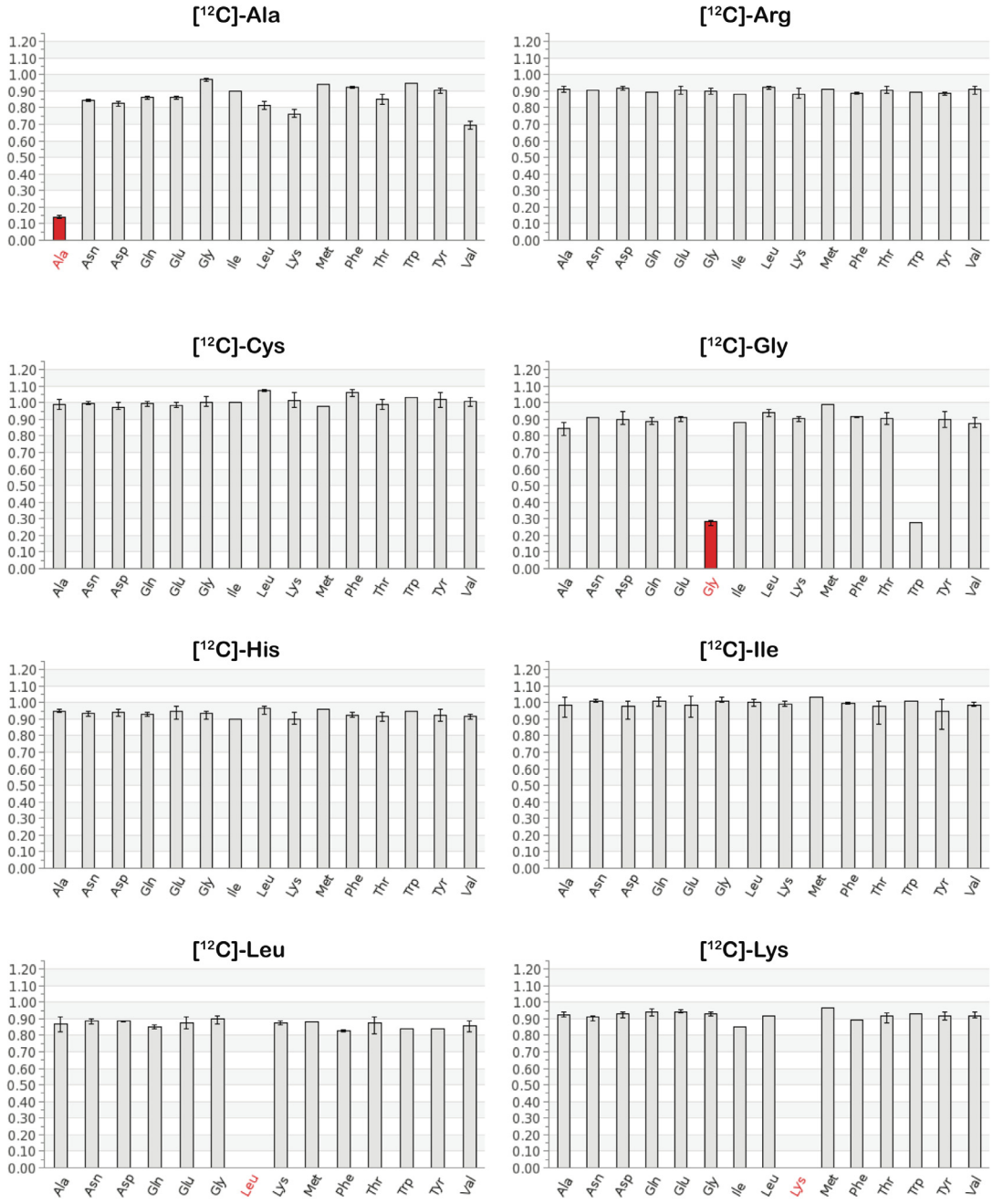


Figure S2 HN(CO) peak intensities of GB1 following unlabeling. GB1 was expressed M9 media containing $^{15}NH_4Cl$ and [^{13}C]-glucose in the presence of unlabeled [^{14}N,^{12}C]-amino acids as indicated on top of the bar graphs. The intensity of each individual peak was extracted from the respective HN(CO) and normalized against the peak intensity extracted from the reference HN(CO) of the fully labeled sample. Peak intensities for individual signals of each amino acid type were averaged (bar graph). Error bars indicate min and max values, respectively. Even though Arg, Ser are not present in the sequence, unlabeling with these allows for assessment of scrambling with respect to other amino acid types. Spectra were recorded identically at 750 MHz, 25°C, with a sample concentration of 0.5 mM using 2048 t_2 increments, 512 t_1 increments; a spectral width of

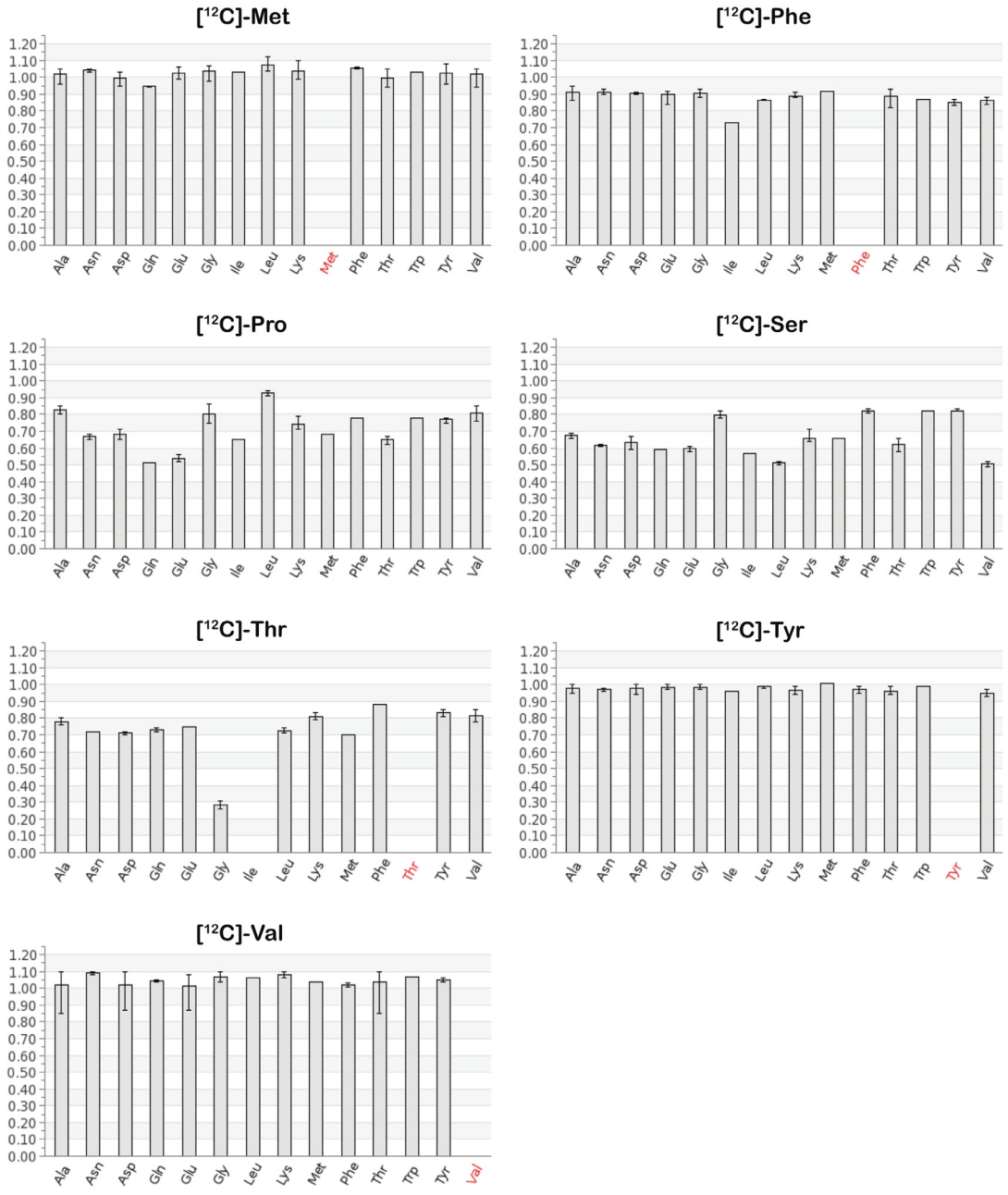


Figure S2 (continued)

12019 Hz (t_2), 2812 Hz (t_1); 12 transients per t_1 increment; keeping the ^{1}H,^{15}N and ^{13}C carrier at 4.7 ppm,118.5 ppm and 173 ppm, respectively. Recorded spectra were processed identically and the absolute peak intensities were extracted using TOPSPIN V2.1.

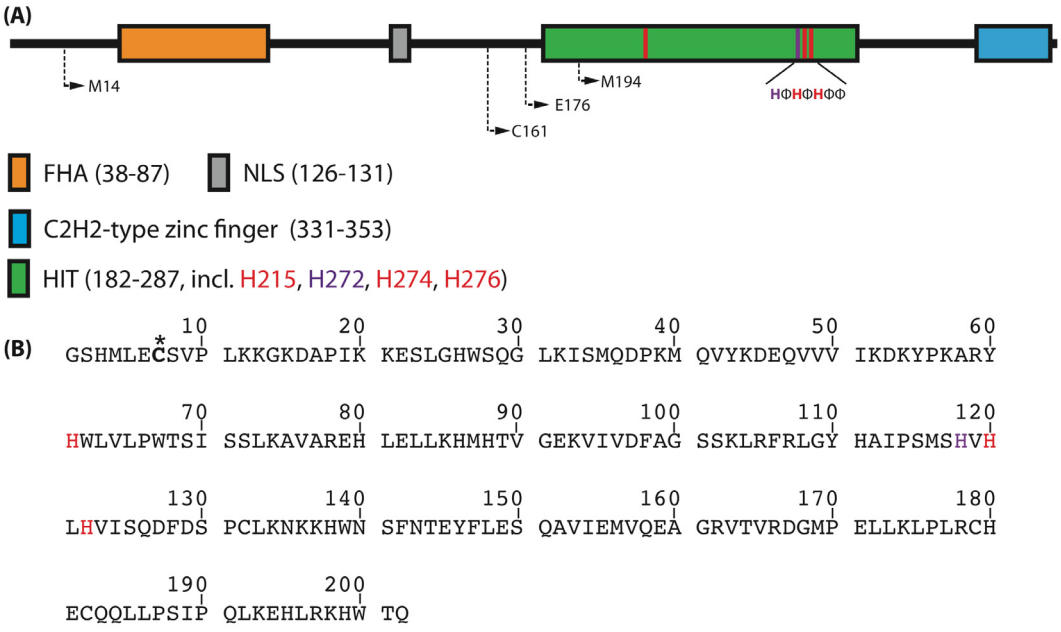


Figure S3 (A) Domain organization of human aprataxin according to Uniprot Q7Z2E3-1. Out of 4 constructs tested starting as indicated at M14, C161, E176 and M194, respectively, C161 turned out to be the most feasible one with respect to protein stability. FHA: forkhead associated domain; NLS: nuclear localization sequence; HIT: histidine triad (HΦHΦHΦΦ, whereas Φ is any hydrophobic amino acid), residues indicated in red are essential for catalytic activity[1]. **(B)** Protein sequence of the aprataxin construct C161-Q356 used in this study. The first 6 amino acids are coded by the pET15b vector. First amino acid of the native aprataxin sequence (C161) is highlighted by a star.

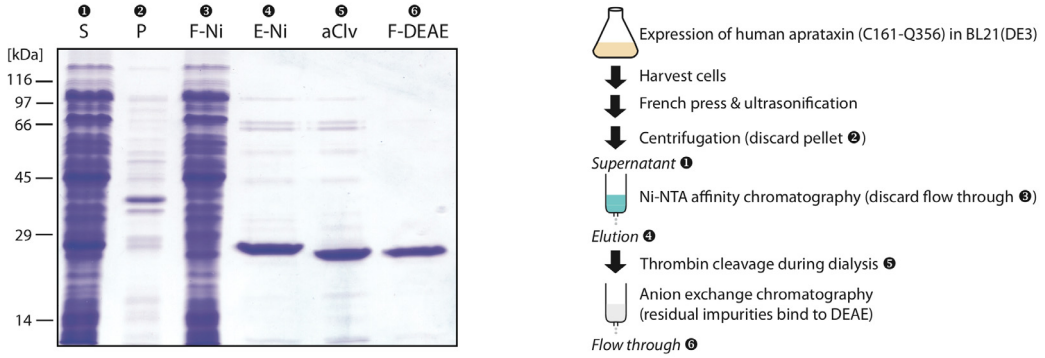
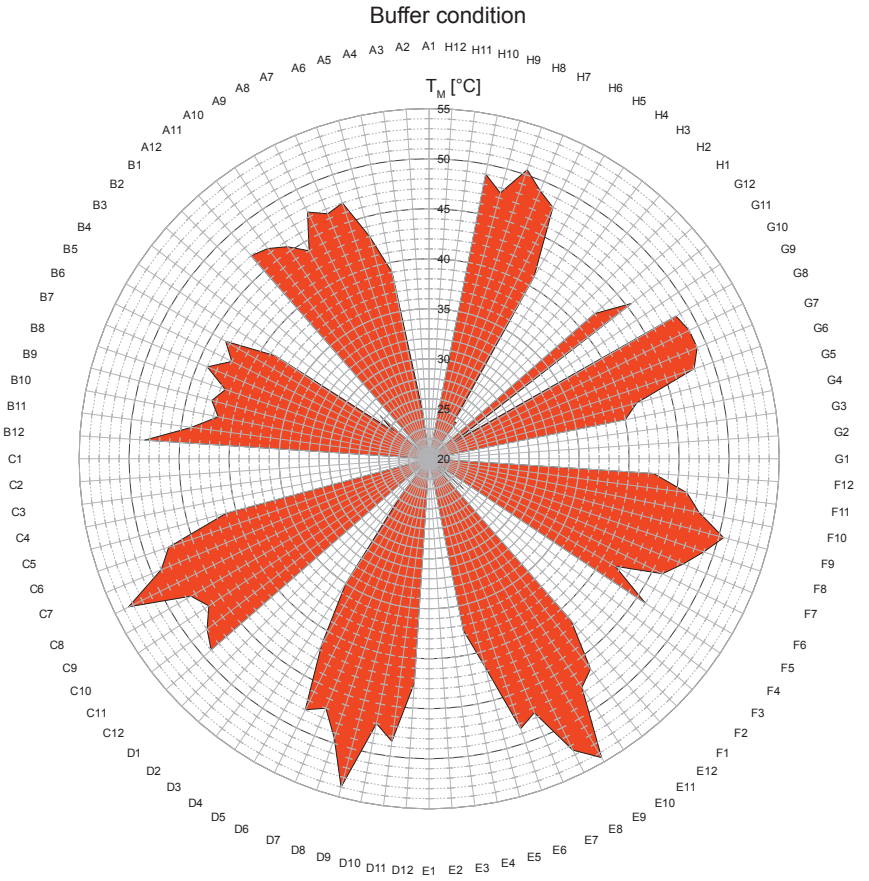


Figure S4 SDS-PAGE and purification of aprataxin C161-Q356. Harvested *E.coli* BL21(DE3) cells were disrupted by French press and ultrasonification following centrifugation (15.400×g, 4°C, 30 min). Pelleted debris (P) was discarded and the supernatant (S) was subjected to affinity chromatography on Ni-NTA agarose, flow through (F-Ni) was discarded. After washing, His_6-tagged aprataxin was eluted from Ni-NTA (E-Ni) and the tag was removed by thrombin cleavage (5 U thrombin per mg aprataxin) during dialysis (overnight at 4°C; 10 mM Tris/HCl pH 7.5, 50 mM NaCl). Untagged aprataxin (after cleavage, aClv) was finally subjected to DEAE anion exchange chromatography to remove Impurities. DEAE flow through (F-DEAE) was collected as pure protein.



	1	2	3	4	5	6	7	8	9	10	11	12
A	50 mM NaAcetate pH 4.0	50 mM NaAcetate pH 4.4	50 mM Citrate pH 5.4	50 mM NaCacodylate pH 6.0	50 mM NaCacodylate pH 6.4	50 mM HEPES pH 7.0	50 mM HEPES pH 7.4	50 mM Tris/HCl pH 8.0	50 mM Tris/HCl pH 8.4	50 mM CHES pH 9.0	50 mM CHES pH 9.4	50 mM CHES pH 10.0
B	50 mM NaAcetate pH 4.0	50 mM NaAcetate pH 4.4	50 mM Citrate pH 5.4	50 mM NaCacodylate pH 6.0	50 mM NaCacodylate pH 6.4	50 mM HEPES pH 7.0	50 mM HEPES pH 7.4	50 mM Tris/HCl pH 8.0	50 mM Tris/HCl pH 8.4	50 mM CHES pH 9.0	50 mM CHES pH 9.4	50 mM CHES pH 10.0
C	50 mM NaAcetate pH 4.0 150 mM NaCl	50 mM NaAcetate pH 4.4 150 mM NaCl	50 mM Citrate pH 5.4 150 mM NaCl	50 mM NaCacodylate pH 6.0 150 mM NaCl	50 mM NaCacodylate pH 6.4 150 mM NaCl	50 mM HEPES pH 7.0 150 mM NaCl	50 mM HEPES pH 7.4 150 mM NaCl	50 mM Tris/HCl pH 8.0 150 mM NaCl	50 mM Tris/HCl pH 8.4 150 mM NaCl	50 mM CHES pH 9.0 150 mM NaCl	50 mM CHES pH 9.4 150 mM NaCl	50 mM CHES pH 10.0 150 mM NaCl
D	50 mM NaAcetate pH 4.0 150mM NaCl	50 mM NaAcetate pH 4.4 150 mM NaCl	50 mM Citrate pH 5.4 150 mM NaCl	50 mM NaCacodylate pH 6.0 150 mM NaCl	50 mM NaCacodylate pH 6.4 150 mM NaCl	50 mM HEPES pH 7.0 150 mM NaCl	50 mM HEPES pH 7.4 150 mM NaCl	50 mM Tris/HCl pH 8.0 150 mM NaCl	50 mM Tris/HCl pH 8.4 150 mM NaCl	50 mM CHES pH 9.0 150mM NaCl	50 mM CHES pH 9.4 150 mM NaCl	50 mM CHES pH 10.0 150 mM NaCl
E	50 mM NaAcetate pH 4.0 200 mM NaCl	50 mM NaAcetate pH 4.4 200 mM NaCl	50 mM Citrate pH 5.4 200 mM NaCl	50 mM NaCacodylate pH 6.0 200 mM NaCl	50 mM NaCacodylate pH 6.4 200 mM NaCl	50 mM HEPES pH 7.0 200 mM NaCl	50 mM HEPES pH 7.4 200 mM NaCl	50 mM Tris/HCl pH 8.0 200 mM NaCl	50 mM Tris/HCl pH 8.4 200 mM NaCl	50 mM CHES pH 9.0 200 mM NaCl	50 mM CHES pH 9.4 200 mM NaCl	50 mM CHES pH 10.0 200 mM NaCl
F	50 mM NaAcetate pH 4.0 200 mM NaCl	50 mM NaAcetate pH 4.4 200 mM NaCl	50 mM Citrate pH 5.4 200 mM NaCl	50 mM NaCacodylate pH 6.0 200 mM NaCl	50 mM NaCacodylate pH 6.4 200 mM NaCl	50 mM HEPES pH 7.0 200 mM NaCl	50 mM HEPES pH 7.4 200 mM NaCl	50 mM Tris/HCl pH 8.0 200 mM NaCl	50 mM Tris/HCl pH 8.4 200mM NaCl	50 mM CHES pH 9.0 200 mM NaCl	50 mM CHES pH 9.4 200 mM NaCl	50 mM CHES pH 10.0 200 mM NaCl
G	50 mM NaAcetate pH 4.0 500 mM NaCl	50 mM NaAcetate pH 4.4 500 mM NaCl	50 mM Citrate pH 5.4 500 mM NaCl	50 mM NaCacodylate pH 6.0 500 mM NaCl	50 mM NaCacodylate pH 6.4 500 mM NaCl	50 mM HEPES pH 7.0 500 mM NaCl	50 mM HEPES pH 7.4 500 mM NaCl	50 mM Tris/HCl pH 8.0 500 mM NaCl	50 mM Tris/HCl pH 8.4 500mM NaCl	50 mM CHES pH 9.0 500 mM NaCl	50 mM CHES pH 9.4 500 mM NaCl	50 mM CHES pH 10.0 500 mM NaCl
H	50 mM NaAcetate pH 4.0 500 mM NaCl	50 mM NaAcetate pH 4.4 500 mM NaCl	50 mM Citrate pH 5.4 500 mM NaCl	50 mM NaCacodylate pH 6.0 500 mM NaCl	50 mM NaCacodylate pH 6.4 500 mM NaCl	50 mM HEPES pH 7.0 500 mM NaCl	50 mM HEPES pH 7.4 500 mM NaCl	50 mM Tris/HCl pH 8.0 500 mM NaCl	50 mM Tris/HCl pH 8.4 500 mM NaCl	50 mM CHES pH 9.0 500 mM NaCl	NO DYE BUT PROTEIN (NEG. CONTROL)	DYE & WATER (NEG. CONTROL)

Figure S5 ThermoFluor[2] based screening of melting temperature of aprataxin to identify optimal pH and NaCl concentration. The fluorophore SYPRO Orange (Invitrogen) was added to the respective sample in a 96 well plate (bottom: plate setup), the plate was heated and increasing fluorescence indicating denature/unfolding was detected using a real time PCR machine (iCycler, Biorad). The total volume of each sample in a well was 50 µl comprising 5 µl 10x SYPRO Orange dye and 5 µl protein. Final protein concentration was 5 µM. Melting temperature T_M of aprataxin samples under different buffer conditions (top) was extracted from the inflection point of the respective fluorescence signal .

Table S2 Partial resonance assignment for backbone nuclei of human aprataxin residues 161-356. Numbering of residues (#Res.) according to UniProt (ID: Q7Z2E3-1). Residues 156-160 are pET15b vector coded residues. Red: histidine residues involved in catalysis[1].

#Res.					
156	Ser				
157	His				
158	Met				
159	Leu				
160	Glu				
161	Cys	-	-	-	-
162	Ser	-	-	173.51	58.32
163	Val	7.96	122.66	174.31	59.87
164	Pro	-	-	176.59	63.01
165	Leu	8.24	122.94	177.29	55.16
166	Lys	8.25	122.61	176.12	56.03
167	Lys	-	-	-	-
168	Gly	-	-	-	-
169	Lys	-	-	176.32	56.58
170	Asp	8.33	120.24	175.28	54.14
171	Ala	7.95	124.82	175.35	50.57
172	Pro	-	-	176.78	63.07
173	Ile	8.08	120.52	176.18	61.15
174	Lys	8.28	125.39	176.44	56.10
175	Lys	-	-	-	-
176	Glu	-	-	-	-
177	Ser	-	-	-	-
178	Leu	-	-	-	-
179	Gly	-	-	-	-
180	His	-	-	-	-
181	Trp	-	-	-	-
182	Ser	-	-	-	-
183	Gln	-	-	-	-
184	Gly	-	-	-	-
185	Leu	-	-	177.79	57.70
186	Lys	7.94	117.68	178.55	59.16
187	Ile	7.55	117.85	178.97	63.41
188	Ser	7.91	117.56	175.18	62.08
189	Met	8.07	116.52	175.51	58.39
190	Gln	7.20	112.63	175.05	55.30
191	Asp	7.45	122.32	175.53	51.08
192	Pro	-	-	179.01	64.61
193	Lys	8.28	115.23	178.17	58.05
194	Met	7.91	115.19	176.60	54.21
195	Gln	7.05	117.55	175.89	55.79
196	Val	8.95	124.41	175.16	62.89
197	Tyr	7.38	118.79	172.59	57.34
198	Lys	7.43	125.87	172.90	56.01
199	Asp	9.29	122.78	175.80	53.27
200	Glu	8.92	116.12	176.96	58.79
201	Gln	9.56	117.36	177.30	58.32
202	Val	8.94	114.71	171.56	57.92
203	Val	8.56	117.57	171.93	59.06
204	Val	8.68	125.12	175.66	59.75
205	Ile	8.76	119.83	175.47	58.35
206	Lys	8.83	121.96	176.34	57.08
207	Asp	7.67	124.01	176.63	55.71
208	Lys	7.31	126.05	175.20	58.05
209	Tyr	8.91	118.51	172.63	53.62
210	Pro	-	-	178.90	62.11
211	Lys	11.56	126.35	171.19	54.16
212	Ala	7.10	114.10	175.96	50.22
213	Arg	9.74	120.91	176.38	60.06
214	Tyr	7.98	115.28	173.00	56.48
215	His	7.71	122.07	174.18	53.73
216	Trp	9.66	126.45	173.30	57.64
217	Leu	9.25	118.94	175.96	52.44
218	Val	9.49	122.09	175.09	61.41
219	Leu	9.24	126.81	-	52.02
220	Pro	-	-	177.05	61.22
221	Trp	7.64	120.42	178.62	59.49
222	Thr	8.17	114.99	172.76	62.42
223	Ser	8.24	120.07	177.05	58.40
224	Ile	-	-	-	-
225	Ser	-	-	-	-
226	Ser	-	-	-	-
227	Leu	-	-	-	-
228	Lys	-	-	175.99	58.33
229	Ala	7.28	119.86	176.90	51.33
230	Val	6.71	118.80	174.98	62.84
231	Ala	11.41	137.49	177.72	50.45
232	Arg	8.42	119.91	179.00	60.51
233	Glu	9.03	117.23	176.83	58.39
234	His	8.08	119.32	175.39	57.68
235	Leu	7.86	121.43	177.59	59.49
236	Glu	8.54	116.21	179.48	59.97
237	Leu	8.05	124.17	178.25	58.11
238	Leu	8.53	119.89	181.98	58.49
239	Lys	8.38	119.96	178.70	60.30
240	His	8.49	122.72	177.30	60.20
241	Met	9.07	117.51	176.38	60.66
242	His	7.48	117.72	176.28	60.67
243	Thr	8.18	114.96	177.45	66.49
244	Val	8.46	123.24	177.18	66.54
245	Gly	8.15	108.32	173.75	48.00
246	Glu	7.83	117.14	178.86	59.75
247	Lys	7.66	121.88	177.48	58.76

#Res.		chemical shift [ppm]			
		H^N	N^H	C'	CA
248	Val	7.95	118.55	177.15	65.99
249	Ile	6.86	118.56	178.03	65.85
250	Val	6.92	117.83	178.52	66.12
251	Asp	8.38	119.14	178.60	56.78
252	Phe	8.19	114.49	176.32	59.05
253	Ala	8.28	123.33	179.29	51.81
254	Gly	7.94	109.60	175.28	47.59
255	Ser	-	-	-	-
256	Ser	-	-	-	-
257	Lys	-	-	176.04	56.71
258	Leu	7.72	120.45	176.11	54.32
259	Arg	8.08	120.22	174.53	55.08
260	Phe	8.69	119.55	176.76	57.34
261	Arg	8.42	116.53	173.65	56.05
262	Leu	9.02	121.81	176.82	54.26
263	Gly	8.97	111.12	169.42	46.80
264	Tyr	8.75	116.42	175.86	56.65
265	His	8.55	119.50	177.01	55.14
266	Ala	8.47	122.67	176.43	54.32
267	Ile	6.09	113.11	-	56.85
268	Pro	-	-	176.05	64.32
269	Ser	6.78	109.07	173.16	56.20
270	Met	-	-	-	-
271	Ser	-	-	-	-
272	His	-	-	-	-
273	Val	-	-	-	-
274	His	-	-	170.65	54.34
275	Leu	8.45	124.25	175.86	53.84
276	His	9.47	125.37	175.47	57.22
277	Val	9.35	123.47	174.97	62.11
278	Ile	8.94	123.25	174.50	58.35
279	Ser	8.75	128.54	173.92	58.57
280	Gln	7.48	110.71	175.57	55.15
281	Asp	9.17	117.01	175.26	54.69
282	Phe	5.91	106.88	173.01	60.55
283	Asp	9.05	117.66	175.11	52.03
284	Ser	7.28	112.69	174.57	52.97
285	Pro	-	-	176.84	63.35
286	Cys	8.09	113.38	174.47	54.66
287	Leu	7.19	122.42	174.98	55.81
288	Lys	10.29	127.20	176.07	57.01
289	Asn	8.09	116.75	174.75	51.87
290	Lys	-	-	176.92	59.90
291	Lys	8.03	120.83	179.21	60.36
292	His	8.08	118.73	178.23	57.95
293	Trp	7.63	115.75	179.07	60.41
294	Asn	9.12	116.31	177.07	54.91
295	Ser	7.98	117.21	171.84	60.67
296	Phe	6.47	112.62	175.05	59.00
297	Asn	6.77	114.05	173.88	54.46
298	Thr	7.14	107.87	176.09	60.95
299	Glu	9.26	119.90	173.30	58.58
300	Tyr	7.81	119.63	173.27	60.13
301	Phe	7.84	117.49	173.88	56.40
302	Leu	9.06	128.98	176.33	53.85
303	Glu	8.80	123.43	177.46	57.67

#Res.		chemical shift [ppm]			
		H^N	N^H	C'	CA
304	Ser	9.10	120.14	177.49	61.75
305	Gln	9.16	114.24	177.62	58.88
306	Ala	6.92	120.21	180.39	54.74
307	Val	8.08	121.57	177.15	67.54
308	Ile	8.38	119.88	177.38	66.57
309	Glu	7.93	117.70	179.07	59.73
310	Met	8.14	119.30	178.15	59.99
311	Val	8.04	119.04	178.36	66.85
312	Gln	8.71	119.75	178.02	59.66
313	Glu	8.30	117.22	177.63	58.51
314	Ala	8.59	119.81	178.74	52.12
315	Gly	8.23	108.70	171.66	46.02
316	Arg	7.19	112.50	173.29	53.97
317	Val	9.60	122.48	175.51	61.96
318	Thr	7.81	125.31	172.80	62.60
319	Val	8.31	127.14	175.72	60.93
320	Arg	8.56	127.41	175.48	56.40
321	Asp	8.31	119.96	177.09	55.08
322	Gly	-	-	-	-
323	Met	-	-	-	-
324	Pro	-	-	179.36	66.17
325	Glu	7.76	115.39	179.90	58.51
326	Leu	7.97	120.53	178.29	56.98
327	Leu	7.04	111.30	178.75	56.00
328	Lys	7.33	117.56	176.31	55.81
329	Leu	6.96	122.71	174.41	54.71
330	Pro	-	-	176.92	62.81
331	Leu	8.47	122.32	174.94	55.43
332	Arg	7.59	121.96	175.55	53.75
333	Cys	8.57	126.64	176.64	61.41
334	His	-	-	176.50	58.58
335	Glu	9.16	125.56	177.87	57.50
336	Cys	8.27	117.15	175.51	58.55
337	Gln	7.47	114.91	174.31	58.52
338	Gln	7.85	121.54	174.97	57.39
339	Leu	8.31	125.31	177.22	54.87
340	Leu	8.36	126.62	176.32	49.43
341	Pro	-	-	-	-
342	Ser	-	-	-	-
343	Ile	-	-	-	-
344	Pro	-	-	179.94	66.61
345	Gln	7.15	114.01	178.69	58.95
346	Leu	8.31	124.56	177.27	58.78
347	Lys	8.65	118.38	179.13	60.75
348	Glu	7.37	117.35	178.21	58.94
349	His	7.95	119.86	176.65	59.94
350	Leu	8.61	117.67	177.95	58.07
351	Arg	6.69	114.35	178.75	58.17
352	Lys	7.35	116.70	177.94	57.84
353	His	7.39	114.93	175.69	55.70
354	Trp	7.36	119.98	176.66	55.91
355	Thr	8.01	114.16	173.68	61.82
356	Gln	7.96	127.44	180.51	57.60

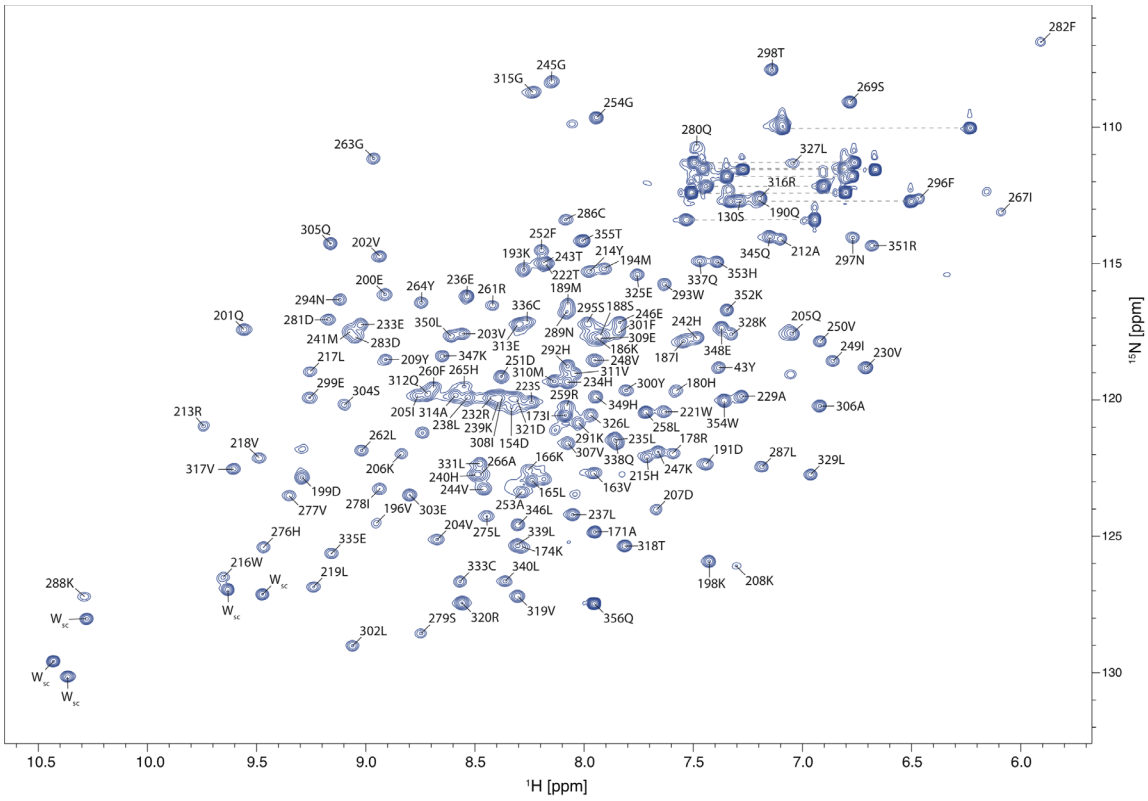


Figure S6 [^{1}H,^{15}N]-HSQC of human aprataxin (residues 161-356 according to UniProt Q7Z2E3-1). Uniformly [^{15}N,^{13}C]-labeled aprataxin was purified from *E. coli* BL21(DE3). The [^{1}H,^{15}N]-HSQC was recorded at 750 MHz and 29°C using a 0.5 mM sample in NMR buffer (10 mM dTris/HCl pH 7.5; 150 mM NaCl; 0.05% NaN_3 ; 10% D_2O). The expected number of backbone resonances is 190 (202 residues - 11x Pro - N-term). 30 backbone amide resonances (6%) are not observable.160 resonances for backbone amides are observable, of which 150 (94% of observed, 79% of expected) have been assigned. Side chain amides of Gln & Asn are indicated by horizontal lines. All 5 amide groups of Trp side chains were observed (labeled with W_{SC}).

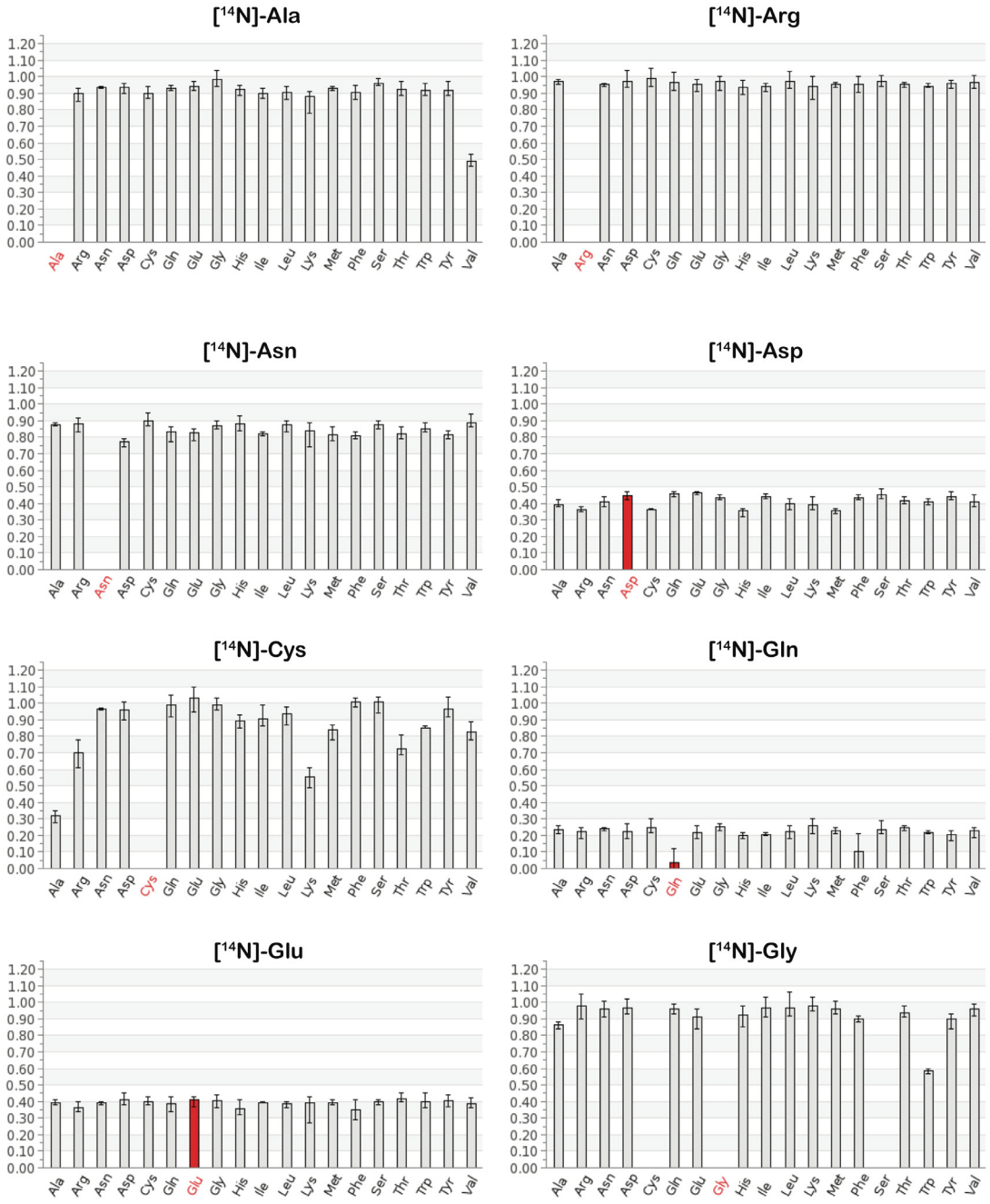


Figure S7 Normalized [^{1}H,^{15}N]-HSQC peak intensities of human aprataxin following unlabeling. Aprataxin was expressed M9 media containing $^{15}NH_4Cl$ and [^{13}C]-glucose in the presence of unlabeled [^{14}N,^{12}C]-amino acids as indicated on top of the bar graphs. The intensity of each individual peak was extracted from the respective [^{1}H,^{15}N]-HSQC and normalized against the peak intensity extracted from the reference [^{1}H,^{15}N]-HSQC of the fully labeled sample. Normalized peak intensities for same amino acid types were grouped and averaged (bar graph). Error bars indicate min and max values, respectively. Spectra were recorded identically at 750 MHz, 29°C, with a sample concentration of 0.38 mM using 2048 t_2 increments; 512 t_1 increments; a spectral width of 10870 Hz (t_2), 2584 Hz (t_1); 12 transients per t_1 increment; keeping the ^{1}H and ^{15}N RF carrier at

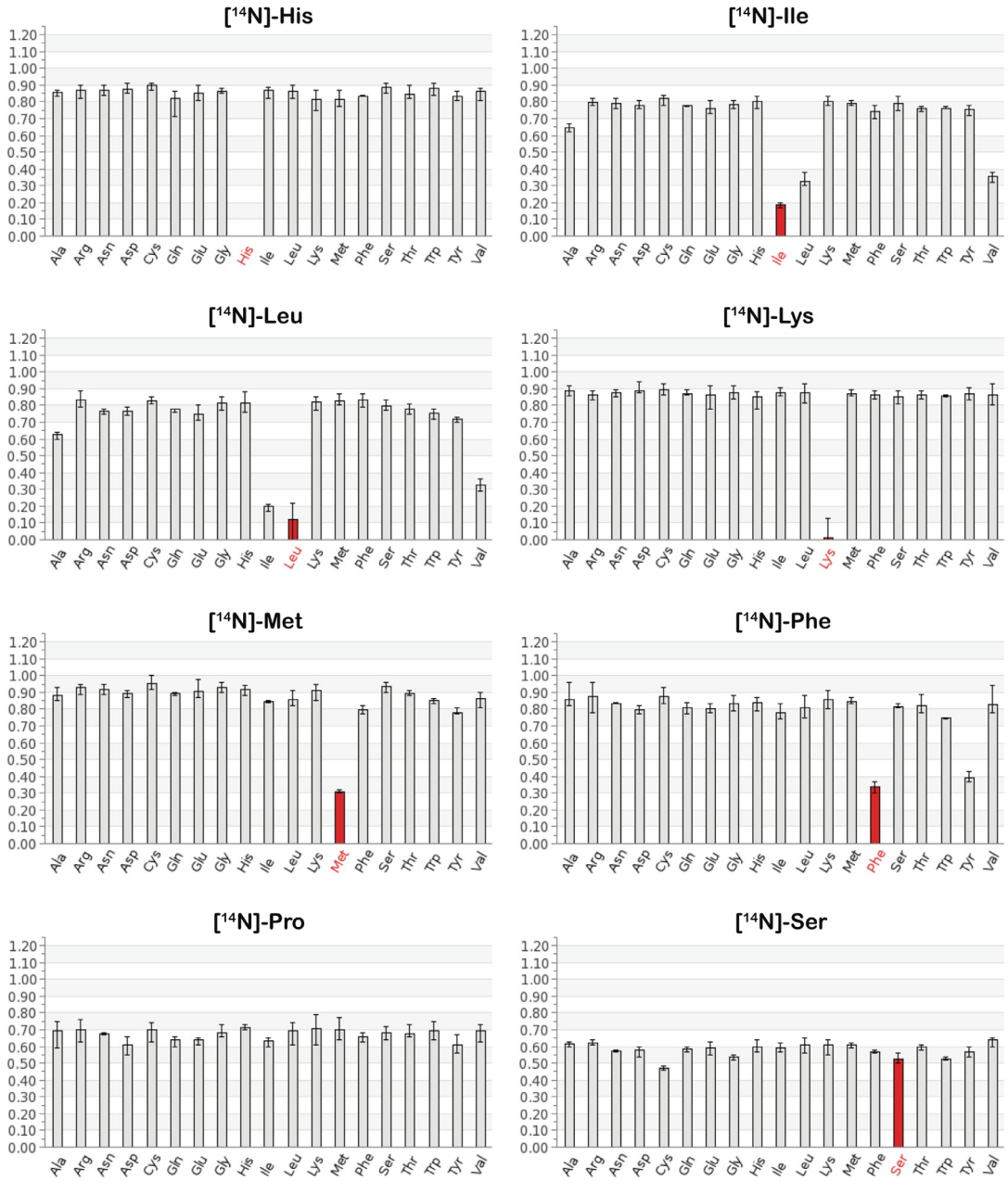


Figure S7 (continued)

4.7 ppm and 122.0 ppm, respectively. Recorded spectra were processed identically and the absolute peak intensities were extracted using TOPSPIN V2.1. Figure continued on next page.

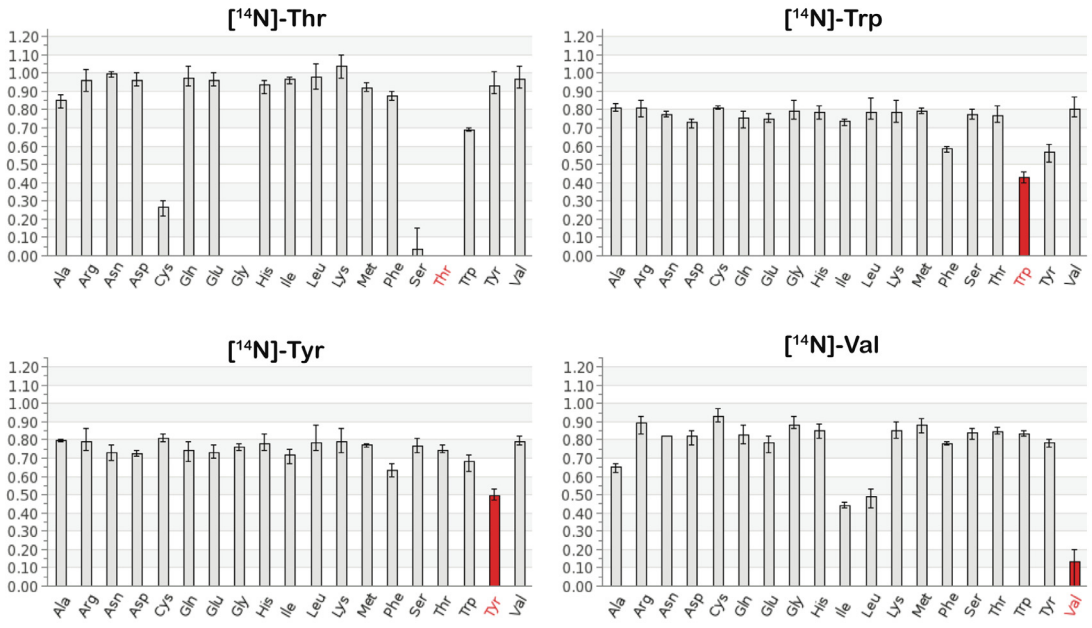


Figure S7 (continued)

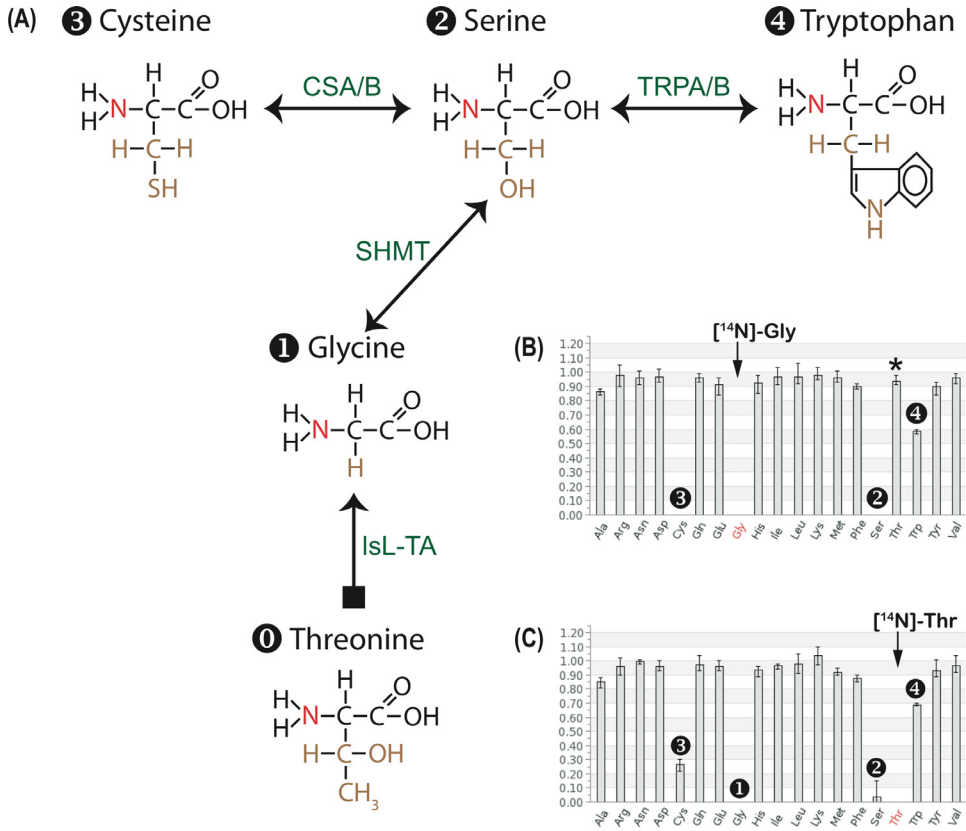


Figure S8 (A) Metabolic pathways involving the conversion of Thr, Gly, Ser, Cys and Trp respectively. Glycine (1) can be interconverted into serine by the bacterial enzyme SHMT (Serine hydroxymethyltransferase; EC 2.1.2.1). Serine (2) in turn can be utilized to produce both cysteine (3) and trypthophan (4). Whereas the first reaction is catalyzed by the cysteine synthase complex consisting of CSA and CSB (Cysteine synthase A and B, respectively; EC 2.5.1.47), the latter one is catalyzed by TRPA/B (Tryptophan synthase alpha and beta chain, respectively; EC 4.2.1.20).
Threonine (0) can be cleaved to Glycine by the bacterial enzyme IsL-TA (Low specificity L-threonine aldolase; EC 4.1.2.48). Please note that all enzymes, except IsL-TA, catalyze the respective conversion in both directions. **(B, C)** Normalized [^{1}H,^{15}N]-HSQC peak intensities of human aprataxin (taken from Fig. S8) following unlabeling with [^{14}N]-Gly and [^{14}N]-Thr, respectively, illustrating the underlying metabolism depicted in (A). Of note, the unidirectionality of the reaction catalyzed by IsL-TA causes threonine to be converted into glycine (C, number 1), but glycine unlabeling does not affect the threonine labeling state (B, asterisk).
All enzymes involved in the depicted reactions where extracted from the respective amino acid metabolism scheme found in the Kyoto encyclopedia of genes and genomes[3]. Respective EC numbers and protein names for *E. coli* (strain K12) were identified using UniProt[4].

References

(1) Rass, U., Ahel, I., & West, S. C. (2008). Molecular mechanism of DNA deadenylation by the neurological disease protein aprataxin. *The Journal of biological chemistry*, *283*(49), 33994–34001.

(2) Pantoliano, M. W., Petrella, E. C., Kwasnoski, J. D., Lobanov, V. S., Myslik, J., Graf, E., et al. (2001). High-Density Miniaturized Thermal Shift Assays as a General Strategy for Drug Discovery. *Journal of Biomolecular Screening*, *6*(6), 429–440.

(3) Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *27*(1), 29–34.

(4) The UniProt Consortium. (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, *35*(Database), D193–D197.

3 Nucleolytic activities of APTX

3.1 Production of hAPTX, *Mth Ligase*, hPARP1

Aprataxin: Human APTX cDNA was expressed in the T7-based *E. coli* BL21(DE3) system as full-length protein and as isolated domains, including the FHA domain and several N-terminal variants and point mutants of the catalytic HIT-ZnF domain (Figure 3.1 A).

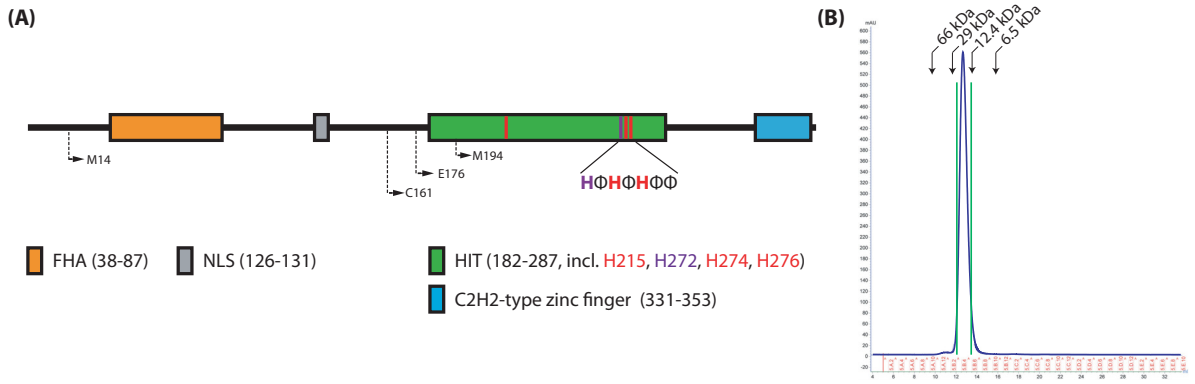


Figure 3.1: Domain organization, construct overview and final purification of hAPTX. (A) Domain organization of human Aprataxin according to Uniprot Q7Z2E3-1. Out of 4 constructs tested starting as indicated, C161 turned out to be the most feasible one with respect to protein stability. Residues indicated in red are essential for catalytic activity and of note, the first histidine of the histidine tirade (HIT) domain is dispensable for deadenylation activity [80]. FHA: forkhead associated domain; NLS: nuclear localization sequence; Φ: hydrophobic residue. **(B)** For enzymatic tests, the final purification step consisted of a size exclusion chromatography (details given in text). Only the central peak fractions (between green bars) were used for further biochemical characterization. The elution volume of different reference proteins is indicated at the top: 9.7 ml for 66 kDa (albumin), 11.6 ml for 29 kDa (carbonic anhydrase), 13.5 ml for 12.4 kDa (cytochrome c) and 15.8 ml for 6.5 kDa (aprotinin), respectively.

The purified FHA domain and most HIT-ZnF constructs were highly prone to aggregation and exquisitely temperature sensitive. The C161-Q356 construct, hereafter termed HIT-ZnF, turned out the most stable and was purified as described in the supplementary material (Figure S3) of *Publication 3*. To minimize the risk for potentially co-purifying contaminants, an additional purification step *via* size exclusion chromatography on a G75 10/300 Superose FF column (equilibrated in 10 mM Tris/HCl pH 7.5, 250 mM NaCl) was performed for all APTX

constructs including mutants prior to activity tests. The elution volume (12.6 ml) for HIT-ZnF corresponds to the expected molecular weight of a HIT-ZnF monomer (approx. 23 kDa). Only the central peak fractions (Figure 3.1 B) were utilized for biochemical characterization and activity assays. The numbering of residues here is based on the longest isoform of human Aprataxin comprising of 356 amino acids. In contrast, most of the literature data refer to the 342 amino acid form of APTX [24], which was used for functional characterization before the long (356 a.a.) isoform was discovered [40].

Mth ligase: For the production of 5′-adenylated DNA, an archeal *Mth* RNA ligase was utilized. It is stable, works also on single-stranded DNA and produces higher yields of adenylated DNA than the T4 enzyme [99, 109]. The ORF for *Mth* ligase was custom-synthesized (Eurofins MWG GmbH), cloned into pET-15b and recombinantly expressed in *E. coli* BL21(DE3). Protein expression in LB media [8] was induced with 0.3 mM IPTG at OD_{600} of 0.3. After protein expression overnight at 20°C (Figure 3.2 A), cells were harvested and stored at -80°C. For protein purification, harvested cells were disrupted by French Press and ultrasonification and the His_6-tagged ligase was purified via Ni-NTA affinity, ion exchange and size exclusion chromatography (Figure 3.2 B). Functional tests of our *Mth* RNA ligase confirmed the capability of adenylating a 5′-phosphorylated single-stranded DNA oligo with the same activity as observed for the commercial available enzyme. One liter LB media yielded approx. 40 mg of pure protein. The same protein from a commercial source (NEB) has a value of about 2000 euro per mg. Due to the high amounts needed for preparative scale substrate preparation, in-house production of this ligase was necessary.

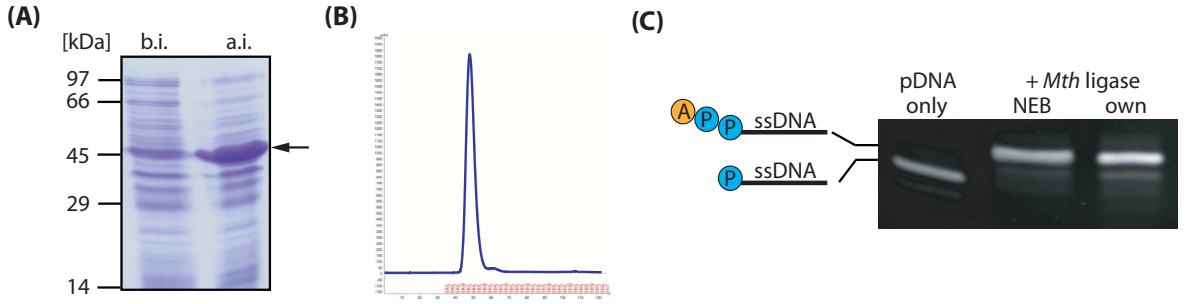


Figure 3.2: Expression and Purification of *Mth* RNA ligase. **(A)** SDS-PAGE of recombinant expression in *E. coli* BL21(DE3). The comparison between before induction (b.i.) and before harvesting the cells after induction (a.i.) of protein expression shows the overexpression of the ligase with the expected size of approx. 45 kDa. **(B)** Final purification step using size exclusion chromatography (SEC) on a G75 10/300 Superose FF column. Only the central peak fractions were used in adenylation reactions. **(C)** Adenylation of 5′-phosphorylated ssDNA (pDNA) utilizing the commercially available enzyme (NEB) and pure protein fractions from B confirmed the functional integrity of our protein. Details on adenylation reaction are provided in section 3.2.

PARP1: His_6-tagged full-length PARP1 was expressed in *E. coli* BL21(DE3) RIPL using an pET-28 expression plasmid kindly provided by J. Pascal (Jefferson University, Philadelphia, PA, USA) [57]. Purification was achieved through Ni-NTA affinity chromatography and SEC, according to published protocols [3, 57]. PARP1 auto-PARylating activity was assessed to ascertain functional integrity of our recombinant PARP1 (Figure 3.3).

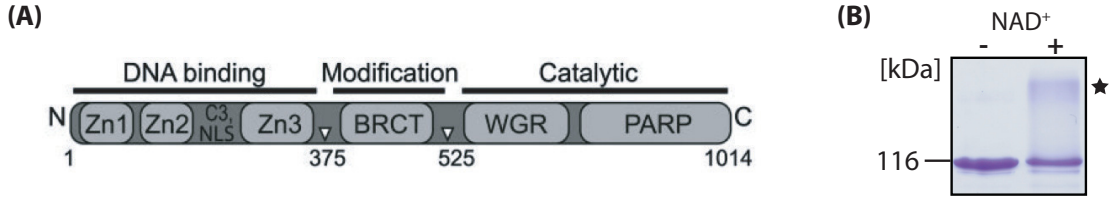


Figure 3.3: Domain organization and auto-PARylation of recombinant hPARP1. **(A)** Domain organization of human PARP1 depicting its modular architecture. While the zinc fingers (Zn1-Zn3) are responsible for DNA binding and the recognition of DNA damage, the BRCT (BRCA1 C terminus fold) is PARylated as a consequence of DNA damage induced activation. The C-terminal PARP1 domain harbors the NAD^+ binding site and the catalytic residues involved in PARylation. WGR denotes a sequence in a highly conserved region. C3: cleavage site of caspase-3 for apoptose-induced inactivation of PARP1. NLS: nuclear localization sequence. Scheme taken from Langelier et al. [57]. **(B)** Auto-PARylating activity in the presence of NAD^+ and nick-mimicking DNA 5′-overhang oligonucleotides is detected on SDS gels as a smear (marked by asterisk), well above the position of the purified recombinant PARP1. Details on PARylation reaction are provided in next section.

3.2 Production of adenylated DNA, of RNA and of poly(ADP-ribose)

Adenylated ssDNA: 5′-adenylated single-stranded (ss) DNA (Figure 3.4 A) was produced according to the manual supplied with the commercial *Mth* RNA Ligase (NEB) with the exception that our recombinant RNA ligase was used. The final DNA adenylation buffer contained 50 mM sodium acetate pH 6.0, 10 mM $MgCl_2$, 5 mM DTT, 0.1 mM EDTA and was supplemented with 0.5 mM rATP. Due to the relative low affinity for DNA as compared to RNA, the enzyme was added in equimolar amounts to the DNA oligo. The DNA including the 5′-phosphorylation (5′-pGTTCCGATAGTGACTACA-3′) was custom-synthesized and HPLC-purified by Eurofins MWG GmbH. The final adenylation reaction was incubated for 1.5 h at 65°C, followed by inactivation of the enzyme (15 min, 95°C). Inactivated protein was sedimented by centrifugation at 16 ×g, and the adenylated DNA in the supernatant was precipitated with 3-fold excess of ethanol in the presence of 10% (v/v) 3.3 M sodium acetate overnight at -20°C. Residual salt was removed during multiple dialysis against water at 4°C and finally adenylated ssDNA was stored lyophilized until used for *de*adenylation assays. For the generation of (2′-AP)ppDNA (Figure 3.4 B), the analogue 2′-aminopurin-riboside-5′-triphosphate (Jena Bioscience GmbH) was used in the reaction instead of rATP.

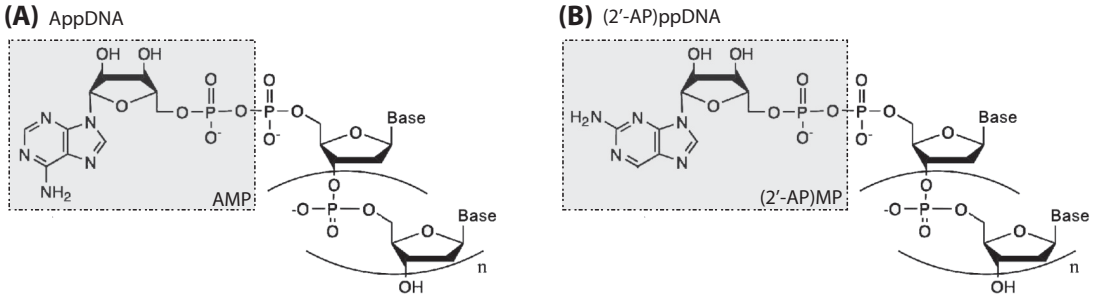


Figure 3.4: Chemical structure of 5′-adenylated and 5′-(2′-AP)-modified ssDNA. (A) Adenylated DNA (AppDNA) comprises an adenosine group linked *via* a characteristic 5′-5′ diphosphate bridge diphosphate bridge to the 5′-end of the DNA. **(B)** (2′-AP)ppDNA differs in the position of the amino group of the purine base only. Figure modified from Krishnakumar and Kraus [53]

RNA: Single-stranded (ss) RNA oligo was produced by *in vitro* transcription using purified T7 RNA polymerase [70, 94, 108] . A linearized pUC18 plasmid coding for the desired RNA oligo and a self-splicing hammerhead ribozyme sequence (Figure 3.5 A) was used to assure homogenous length of the desired oligo after transcription [78, 89]. To obtain RNAs with a 5′-cap structure, the transcription reaction was complemented with different cap analogs, which are incorporated by the T7 RNA polymerase at the first position of the transcribed RNA

[22, 38, 74]. The composition of typical transcription reactions is listed in Table 3.1. All RNA oligos were HPLC-purified using a Vydac C18 218TP54 column at 60°C, a flow rate of 2 ml/min and a linear gradient of buffer A (50 mM potassium phosphate pH 5.9 + 2 mM tetrabutylammonium hydrogen sulfate) to 50% buffer B (buffer A + 60% acetonitrile) and a gradient slope of 0.3%/min [94, 108]. The desired RNA oligo eluted at approx. 38% of buffer B and could be separated from the hammerhead RNA, which eluted at 40% of buffer B.

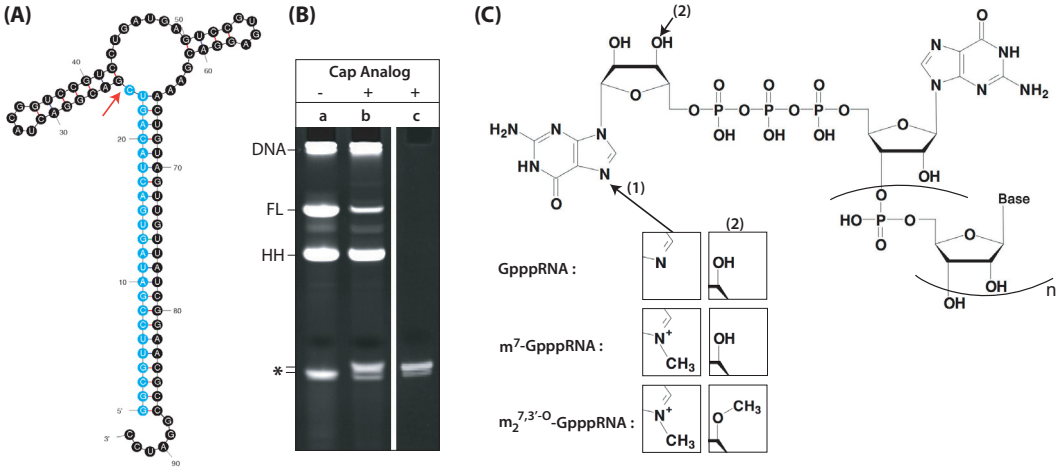


Figure 3.5: Preparation of unmodified and 5′-capped RNA. (A) Structure of uncapped RNA. The desired 24 nt ssRNA (blue) is auto-catalytically liberated by the hammerhead ribozyme RNA in *cis* (black) at the position indicated (red arrow). **(B)** End product analysis of the *in vitro* transcription on a denaturing Urea-PAA gel. Uncapped (a) and capped target-RNA (b) differ in size by approx. 1 nucleotide. The RNA oligo (*) can be separated from the hammerhead-RNA (HH), residual full-length transcript (FL) and plasmid-DNA (DNA) by HPLC yielding a purified RNA (c). **(C)** Chemical structure of capped RNAs. Depending on the cap analog used during the *in vitro* transcription, the guanine ribonucleotide of the 5′-cap structure is either unmethylated (GpppRNA) or monomethylated (m^7-GpppRNA) at the position indicated (1). The 3′-hydroxyl group of the ribose (2) is additionally methylated in the case of $m_2^{7,3'-O}$-GpppRNA. Composition of transcription reactions are listed in Table 3.1.

Table 3.1: *In vitro* transcription for the production of different RNA substrates. BamHI linearized pUC18 plasmid coding for the RNA is incubated with T7-polymerase (T7-Pol) for 4 h at 37°C in the presence of NTPs. For the production of 5′-capped RNAs, the transcription reaction was complemented with GpppG (Cap-Analog 1, Jena Bioscience GmbH), m^7-GpppG (Cap-Analog 2, NEB) or $m_2^{7,3'-O}$-GpppG (Cap-Analog 3, NEB), respectivley.

	pRNA	**GpppRNA**	**m^7-GpppRNA**	**$m_2^{7,3'-O}$-GpppRNA**
1 M Tris/Glu pH 8.1	100 μl	100 μl	40 μl	25 μl
200 mM Spermidine	5.0 μl	5.0 μl	2.0 μl	1.25 μl
300 mM MgAcetate	72 μl	72 μl	28.8 μl	18 μl
100 mM ATP	50 μl	50 μl	20 μl	12.5 μl
100 mM CTP	50 μl	50 μl	20 μl	12.5 μl
100 mM UTP	50 μl	50 μl	20 μl	12.5 μl
100 mM GTP	50 μl	10 μl	4 μl	2.5 μl
100 mM Cap-Analog 1	-	40 μl	-	-
100 mM Cap-Analog 2	-	-	16 μl	-
100 mM Cap-Analog 3	-	-	-	10 μl
1 M DTT	10 μl	10 μl	4.0 μl	2.5 μl
pUC18 plasmid (2.6 μg/μl)	38.5 μl	38.5 μl	15.4 μl	9.62 μl
T7-Pol (34 mg/ml)	5.0 μl	5.0 μl	2.0 μl	1.25 μl
H_2O	68.5 μl	5.0 μl	27.8 μl	17.38 μl
Final volume	**500 μl**	**500 μl**	**200 μl**	**125 μl**

Poly(ADP-ribose) (PAR): PAR is a polymer (Figure 3.6 A) with a ADP-ribose chain length of approx. up to 200 units [14]. Auto-modification of PARP1 is predominantly induced by DNA single-strand breaks [26]. For the *in vitro* production of PAR, PARP1 was activated with nick-mimicking double-stranded DNA with a 5′ overhang obtained *via* annealing of two complementary single-stranded DNA oligos (Oligo 1: 5′-pGTTCC GATAG TGACT ACA-3′ , Oligo 2: 5′-TGTAG TCACT ATCGG AACTC GGGCG ACACG GATAT G-3′).

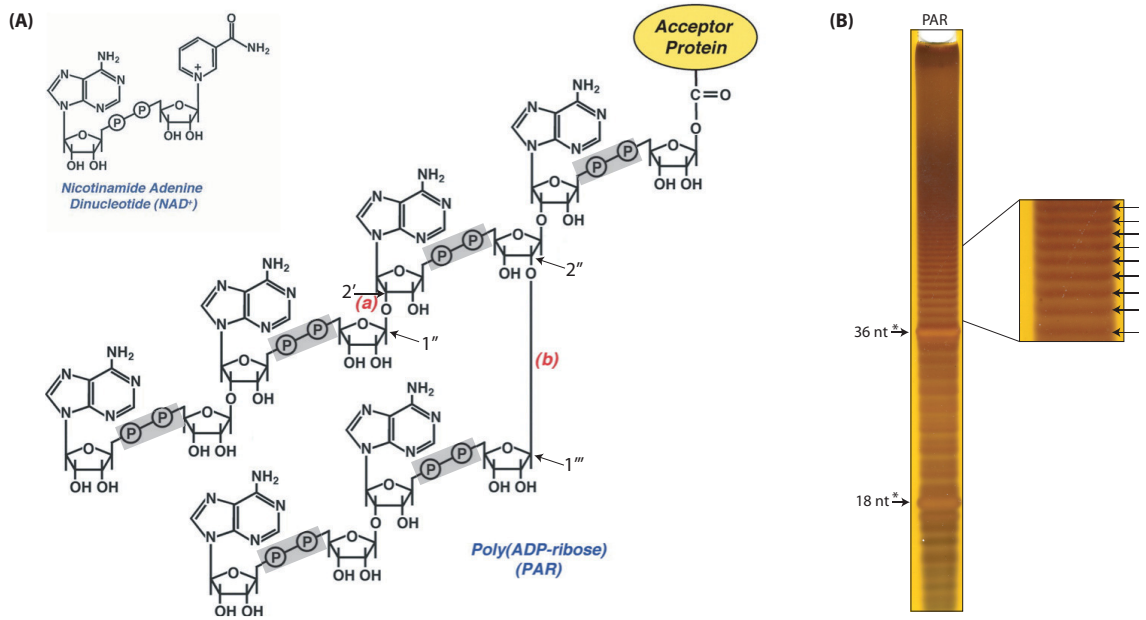


Figure 3.6: Chemical structure of NAD$^+$ and poly(ADP-ribose). (A) Poly(ADP-ribose) (PAR) is a branched polymer synthesized on acceptor proteins by poly(ADP-ribose)-polymerases (PARPs) using NAD$^+$ as donor of monomer units. The ADP-ribose monomers are linked either linearly *via* $1'' \rightarrow 2'$ ribose-ribose glycosidic bond (a) or in a branched mode (b) *via* a $1''' \rightarrow 2''$ ribose-ribose glycosidic bond, respectively. Grey boxes: diphosphate bridges linking an adenine and a ribose unit, similar to the situation found in adenylated DNA (Figure 3.4 A). Figure and figure caption modified from Krishnakumar and Kraus [53]. **(B)** Poly(ADP-ribose) product of PARP1 auto-PARylation separated on a denaturing Urea-PAA gel and visualized by silver staining. PAR of different length units can be discriminated as indicated by the arrows. *: DNA oligos used for PARP1 activation to produce PAR.

The final reaction buffer contained 40 mM Tris/HCl pH 7.5, 100 mM NaCl, 10 mM $MgCl_2$, 150 ng/μl DNA and 1 mM NAD$^+$. The final auto-PARylation reaction was incubated for 4 h at 37°C following heat inactivation of PARP1 (95°C, 30 min). Residual NAD$^+$ was removed by dialysis against water (4°C, overnight) and precipitated protein was sedimented by centrifugation at 16 $\times$g. The PAR-containing supernatant (Figure 3.6 B) was stored at -80°C until further use.

3.3 APTX not only cleaves adenylated DNA

In order to demonstrate that the purified, recombinant HIT-ZnF construct is catalytically active, a deadenylation assay was performed using double-stranded (ds), adenylated DNA construct (Figure 3.7 A). In addition to cleaving the "canonical" substrate, which confirms already published results [80], APTX interestingly also acts on *single-stranded* (ss), adenylated DNA (Figure 3.7 B). The possibility of co-purified impurities, responsible for the activity observed, can be essentially excluded, as the active-site mutant (H274A), in which the catalytically important central histidine of the HIT motif (Figure 3.1 A) is replaced by Ala, shows no deadenylation activity (Figure 3.7 B). To test for substrate specificity, the 5′-(2′-aminopurine)-DNA analog was prepared, which differs in the position of the amino group of the adenine base only (see also 3.4). Interestingly, the HIT-ZnF is also capable of removing this non-canonical (2′-AP)-modification from ssDNA (Figure 3.7 C).

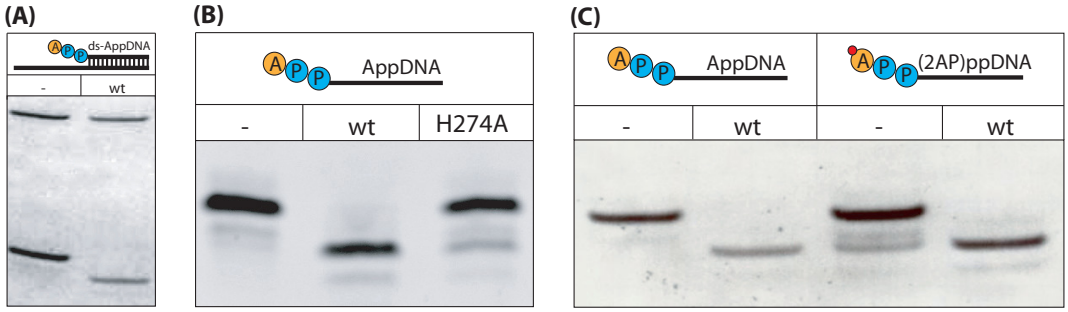


Figure 3.7: Activity of purified recombinant hAPTX HIT-ZnF on adenylated DNA. (A) Double-stranded, adenylated DNA was incubated in absence (-) or presence of wt HIT-ZnF (wt). Products were analysed on denaturing Urea-PAA gels. After incubation with wt HIT-ZnF, the DNA migrates faster, indicating the removal of the 5′-adenylate modification. ds-AppDNA: The 18mer ssDNA oligonucleotides carrying a 5′-adenylate (AppDNA) was annealed to an complementary, unmodified 36mer. **(B)** The same activity is seen with adenylated, single-strand DNA (AppDNA). The active-site mutant H274A does not remove the 5′-adenylate.**(C)** HIT-ZnF also deadenylates single-stranded DNA carrying a 5′-2-aminopurine modification ((2AP)ppDNA). Of note, the gels also attests to the activity of the purified recombinant *Mth* ligase to covalently attach the 5′-rAp and 5′-(2′-AP)p moiety to the ssDNA oligonucleotide.

Apart of this characterization of substrate specificity, there was evidence [41], that the highly conserved Ser 182 of APTX is phosphorylated by the serine/threonine-protein kinase ATR induced upon DNA damage [1, 29]. Furthermore, the existing DNA complex structure of the *S. pombe* APTX homolog [100] shows, that this Ser is involved in the binding of the adenylated DNA and here in particular the 5'-phosphate group of the DNA. I speculated, whether this phosphorylation may regulate the activity of the deadenylation reaction. However, utilizing a Ser to Asp mutation mimicking phosphorylation at this position, no influence on the deadenylation was observed (data not shown), indicating that this putative phosphorylation

may not be of functional significance here or would require the presence of the FHA domain to elicit an effect.

3.3.1 APTX acts upon chimeric RNA/DNA-oligonucleotides

Given the apparent lack of absolute specificity towards the adenylate, I analyzed if HIT-ZnF would remove also terminal single *ribonucleotides* from DNA oligonucleotides. Contrary to the situation in adenylated ssDNA, the ribonucleotide here is attached via regular *single* phosphodiester bond. Nevertheless, 5′-rNMP nucleotides are removed with a rG < rA ≈ rU < rC preference (Figure 3.8 A).

In contrast, no activity was detected for removing single ribonucleotides at the 3′terminus of a DNA oligonucleotide (Figure 3.8 B). Furthermore, APTX was tested for endonucleolytic activity by utilizing a DNA oligomer with an internally incorporated single ribonucleotide. This chimeric DNA/RNA-oligomer was designed in such a way, that the rNMP inside the ssDNA oligomer is in an off-center position. A cleavage at the ribonucleotide would therefore result in the generation of two smaller DNA fragments of different size, which can easily be discriminated on a denaturing Urea-PAA gel. Here, APTX also cuts chimeric single-stranded RNA/DNA-oligomers with a single embedded ribonucleotide, but not if they are annealed to a complementary DNA strand (Figure 3.8 C). In order to identify the exact place of cleavage within the modified ssDNA, the products were separated together with “putative” end products as references (Figure 3.8 D). This clearly indicates, that the longer DNA fragment is unphosphorylated at its 5′-end, consistent with the observation for the 5′-exonucleolytic activity. The latter activity also generates DNA without a 5′-phosphate group after cleavage of the attached 5′-ribonucleotide (Figure 3.8 A). The nature of the short fragment can not be identified conclusively, as the migration distance does not exactly match with one of the reference oligonucleotides. The short fragment migrates faster than the 19 nt 3′-rC-DNA, but slower than the unphosphorylated 18 nt DNA reference oligo. However, the migration behavior in the gel system used depends upon the size:charge ratio only. Therefore, one can speculate that the short fragment corresponds to a 19 nt 3′-rC-DNA carrying one additional negative charge. This negative charge can only be attributed to an additional 3′ phosphate group leading to a putative cleavage mode as depicted in Figure 3.8 E. In consequence, the 3′-ribonucleotide remains on the 5′-ssDNA (18 nt) cleavage product.

Surprisingly, the functional mutant of the deadenylating catalytic site (H274A), virtually inactive in deadenylation (Figure 3.7 C), still removes the 5′-rNMP (Figure 3.9 A). In contrast, two other residues (K288, H292) just outside the canonical catalytic pocket and mutated on

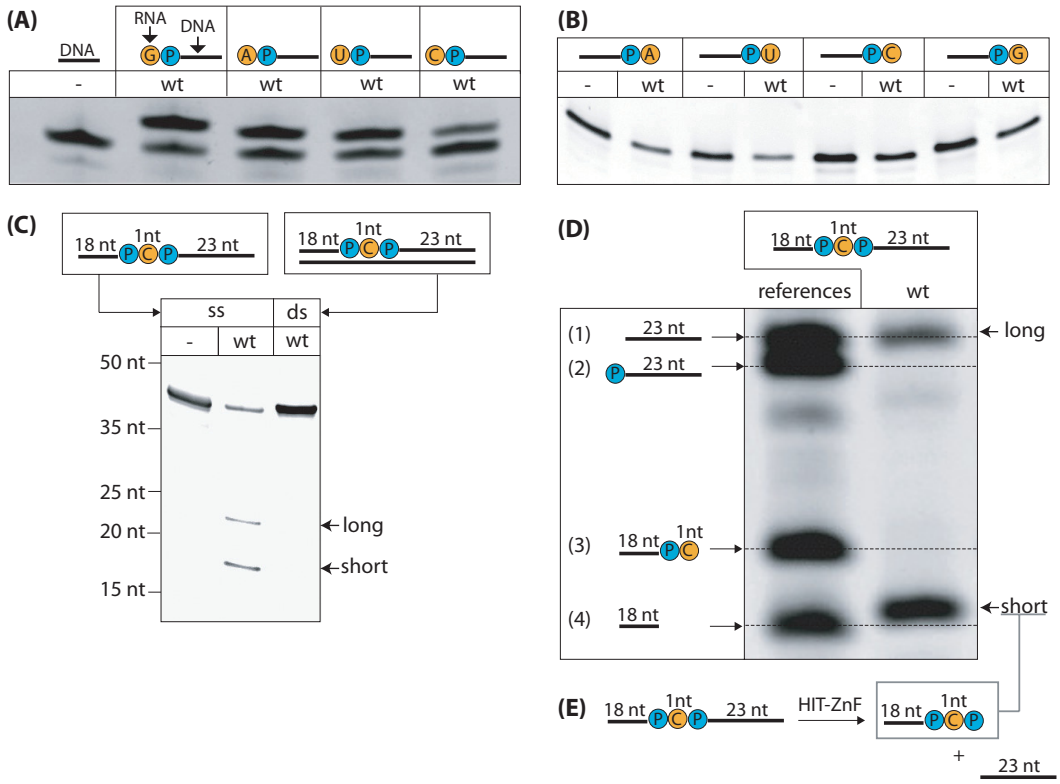


Figure 3.8: RNAse-like activity of HIT-ZnF. DNA with 5′, 3′ or single internally incorporated single ribonucleotides was incubated in absence (-) or presence of wt HIT-ZnF (wt). Products were analysed on denaturing Urea-PAA gels. **(A)** After incubation with wt HIT-ZnF, the DNA migrates faster, indicating the removal of the ribonucleotides from the 5′-end of ssDNA with a order of preference of rG<rA≈rU<rC. DNA: ssDNA, lacking the 5′-rNMP, shown for size comparison. **(B)** No activity is seen with 3′-attached ribonucleotides. **(C)** ssDNA (ss) with an internally incorporated single rC-ribonucleotide (depicted in orange) is cleaved resulting in a long and a short DNA fragment. No activity is observed, if the ribonucleotide containing DNA is annealed to a complementary DNA strand prior to the reaction (ds). **(D)** For size comparison, different modified and unmodified reference oligomers were separated together with the long and short DNA fragment generated by the cleavage of ssDNA with an internal rC. (1): unphosphorylated ssDNA, (2): 5′-phosphorylated ssDNA, (3): ssDNA with 3′-attached rC (18+1 nt = 19 nt), (4): unphosphorylated ssDNA. **(E)** Putative mode of cleavage, based on observation presented in (A-C) . Modified and unmodified DNA oligonucleotides were custom synthesized by MWG Eurofins GmbH.

the basis of the existing DNA complex structure of the *S. pombe* APTX homologue [100] appear largely responsible for this activity. Their mutation to Glu yields significantly lower activity towards 5′-rC-DNA (Figure 3.9 A). Importantly, the K288E and H292 mutants *not* cleaving 5′-rNMP-DNA, conclusively indicate, that this RNAse-like activity of purified HIT-ZnF is *not* due to a contaminant as in such case *all* HIT-ZnF constructs, which were all purified exactly in the same way, should show this nucleolytic activity. Interestingly and essential for the following observations, a commercial ribonuclease inhibitor selectively interferes with the HIT-ZnF RNAse-like activity, but does not affect the deadenylation activity (Figure 3.9 B).

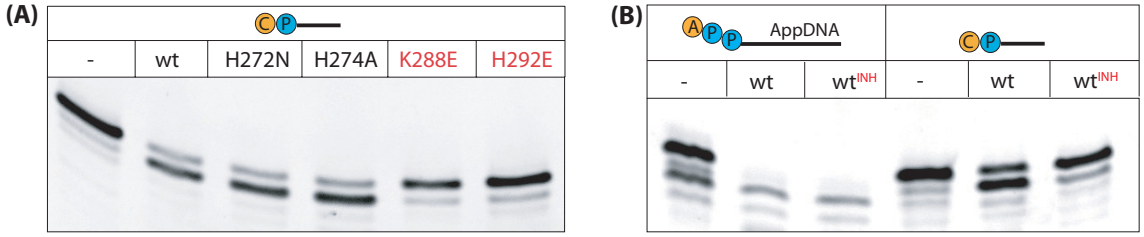


Figure 3.9: Inactivation/Inhibition of RNAse-like activity. **(A)** Wild-type HIT-ZnF (wt) and deadenylating active site mutant (H274A) are active in removing the single rC from ssDNA but the K288E and H292E mutants are much less active. The H272N substitution is found in APTX orthologs in yeast and is apparently not linked to the RNase-like activity. (-): 5′-rCMP-modified ssDNA for size comparison. **(B)** Adenylated ssDNA and 5′-rC-DNA was incubated in the absence (-) or presence of wild-type HIT-Znf (wt). The presence of the commercial ribonuclease inhibitor (SUPERase.In; Life technologies GmbH) during the reaction (wt^{INH}) prevents the rC-cleavage, but does not interfere with the deadenylation activity of HIT-ZnF.

Together with the K288E / H292E mutants, this suggests, that this RNAse-like activity is an intrinsic HIT-ZnF property, which however is sensitive to ribonuclease inhibitors.

3.3.2 APTX is able to decap RNA

It was reported that APTX binds dsDNA and ds*RNA* with comparable affinity, and hairpin DNA structures with even higher affinity than linear DNA [51]. Eukaryotic mRNA is protected against degradation from the 5′-end *via* an m^7-Gppp modification [82]. Furthermore, this 5′-cap structure facilitates recruitment of mRNAs to ribosomes *via* the cap-binding initiation factors and fulfills essential functions in *e.g.* nuclear export signaling [5]. The 5′-capped RNA (Figure 3.5 C) shares some similarities with the adenylated DNA ($m^7\underline{GpppN}$ vs. $\underline{AppN}$). Interestingly, the RNase-like activity of the HIT-ZnF domain is not restricted to ribonucleotides connected to DNA, but is also able to completely degrade even capped RNA (Figure 3.10 A).

To our surprise, the K288E and H292E mutants, which were essentially inactive in removing the single ribonucleotide from ssDNA, are both still capable of degrading the capped RNA (data not shown). This indicates, that the RNAse-like activity is not solely mediated by these two residues alone. Importantly, this degradation of RNA can be suppressed by the application of the ribonuclease inhibitor, which enabled me to separate experimentally this RNAse-like activity from all other activities of HIT-ZnF. Still, when this RNAse-like activity was inhibited, HIT-ZnF is able to decap m^7-GpppRNA (Figure 3.10 B). This decapping reaction is strongly dependent on the presence of the 3′-hydroxyl group of the ribose unit in the cap structure, as the replacing of the 3′-hydroxyl group by a 3′-O-methyl group abolishes the decapping activity (Figure 3.10 B). This hydroxyl group is in close proximity to the first phosphate group of the

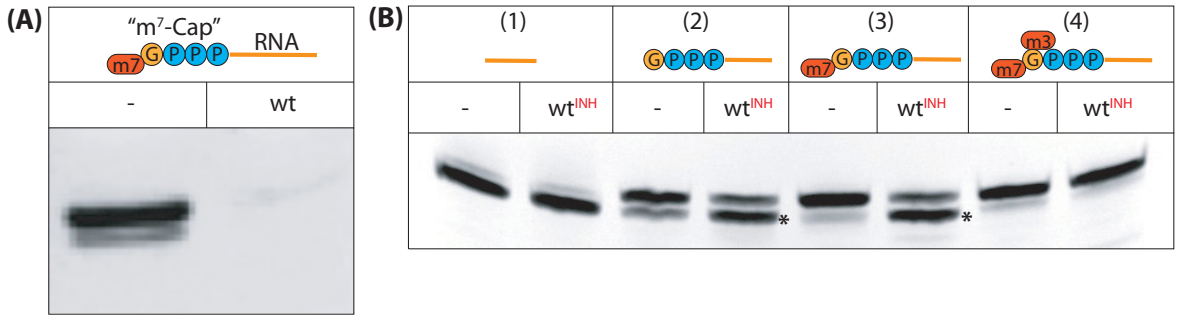


Figure 3.10: Decapping activity of HIT-ZnF. (A) RNA m^7-GpppRNA was incubated in the absence (-) and presence of wild-type HIT-ZnF (wt). In the latter case, no band is observable on the denaturing Urea-PAA gel, indicating the complete degradation of the RNA. **(B)** Incubation of m^7-capped RNA with HIT-ZnF in the presence of ribonuclease inhibitor (wt^{INH}) results in faster migration of the RNA (*) indicating decapping. The methylation of the 3′-hydroxyl group of the ribose in the cap structure ($m_2^{7,3'-O}$-GpppRNA) prevents HIT-ZnF from decapping. (-): capped RNA for size comparison. Of note: The ribonuclease inhibitor *per se* does not harbor any decapping activity (data not shown).

tri-phosphate bridge between the cap structure and the RNA (see also Figure 3.5). The fact, that the absence of the 3′-hydroxyl group prevents the cleavage, provides a first insight into the reaction mechanism. It strongly suggests, that the cleavage involves the 3′-hydroxyl group as a nucleophile and that this nucleophile attacks the first phosphate group of the triphosphate 5′-5′ linkage - a reaction mechanism similar to the first G-nucleotide dependent transesterification step during the course of the catalysis of group 1 intron self-splicing [9].

3.3.3 APTX degrades poly(ADP-ribose)

As the 5′-adenylate group and poly(ADP-ribose) (PAR) share principal structural similarities (Figure 3.11), HIT-ZnF was analyzed with respect to a nucleolytic activity towards PAR. Using PAR produced via recombinant PARP1 (Figure 3.6 B), it could be shown that long chain PAR is degraded completely in the presence of HIT-ZnF (Figure 3.12 A), which is most probably the most interesting and enticing discovery of my work.

The wild-type HIT-ZnF, but also the deadenylation-inactive H274A mutant, reduce long-chain PAR (Figure 3.12 B). This indicates, that the HIT-ZnF PAR cleavage activity does not solely depend upon the canonical catalytic centre of HIT-ZnF. However, the H274A mutant gives rise to PAR fragments with different electrophoretic mobility (Figure 3.12 B). Here, the PAR bands are detected "in-between" of the bands observed for the wt HIT-ZnF. This suggests a change in charge or charge distribution at the terminus of the cleavage product(s) and warrants further detailed analysis in the future. Still, longer incubation of PAR in the presence of this mutant results also in complete degradation of PAR (data not shown). This

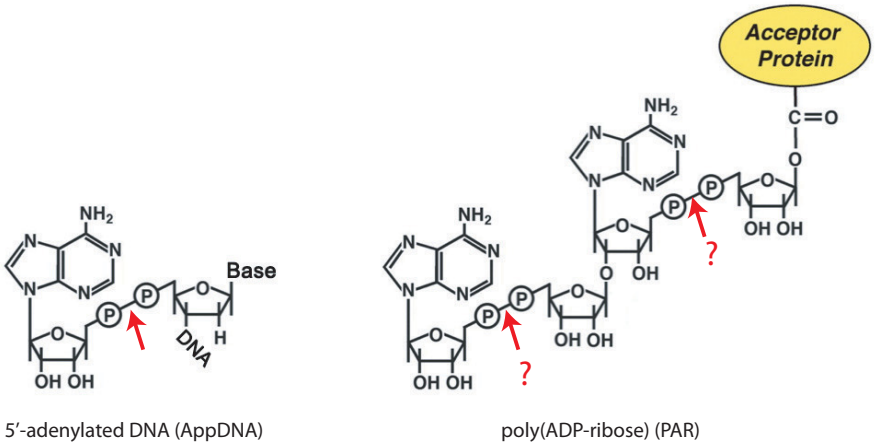


Figure 3.11: Structural similarities between 5'-adenylated DNA and poly(APD-ribose). APTX cleaves AppDNA between the diphosphate bridge connecting the adenine base to the 5'-terminal ribose unit of the DNA (red arrow). As AppDNA and PAR share structural similarities, I speculated whether HIT-ZnF is also able to recognize and cleave the diphosphate bridges of PAR. Figure modified from Krishnakumar and Kraus [53].

implies that the H274 is involved but not exclusively accountable for the PAR cleavage activity and further work has to be done in order to identify the (other) residues involved in the catalysis of PAR degradation. Importantly, the difference in the cleavage products for the wild-type HIT-ZnF and the H274A mutant (Figure 3.12 B) practically excludes a contaminant causing the PAR cleavage, as such a contaminant should *not* be sensitive to mutations in the HIT-ZnF. As neither the ribonuclease inhibitor nor the K288E+H292E mutations (inactive in the cleavage of 5'-rNMP-DNA) abolish the PAR degrading activity investigated here (Figure 3.12 C, D), I conclude, that this PAR degradation activity is also separable from the RNAse-like activity described above (section 3.3.1).

Consistent with the observation, that the S182D mutation does not affect the *deadenylation* activity of HIT-ZnF, this phosphorylation mimicking mutant is still capable of degrading PAR (Figure 3.12 E). This indicates that the putative phosphorylation of this residue by the serine/threonine-protein kinase ATR may also not be involved in the regulation of the PAR degradation activity analyzed here.

An initial characterization of the PAR cleavage reaction by HPLC yields a product with a retention time different from control nucleotides, but close to AMP (Figure 3.13 A). This suggests that wild-type APTX HIT-ZnF produces a low molecular weight cleavage product which could be similar or identical to the one shown in Figure 3.13 B. Interestingly, the retention time of the cleavage product is also different from ADP-ribose. The latter is the PAR degradation product of poly(ADP-ribose) glycohydrolase (PARG) (Figure 3.13 B), which is the major enzyme responsible for PAR degradation *in vivo* [10]. This not only provides an

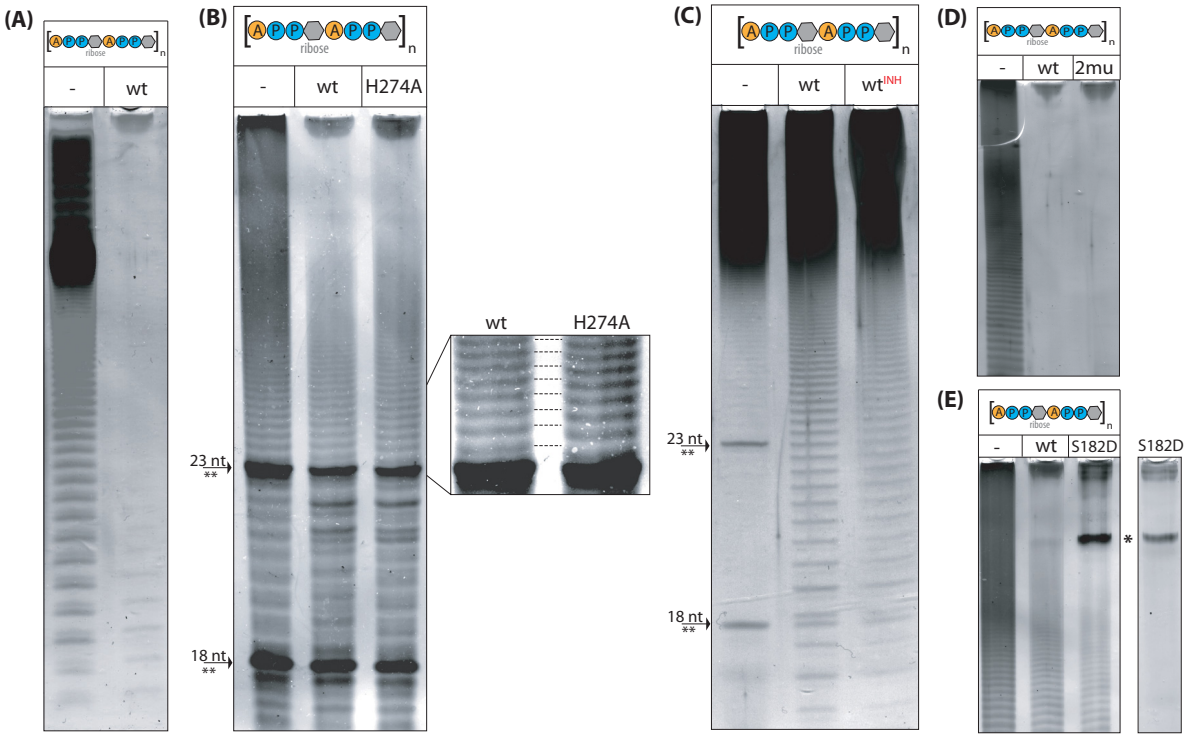


Figure 3.12: Poly(ADP-ribose) degradation by HIT-ZnF. (A) Complete PAR degradation by wild-type HIT-ZnF (wt) monitored on a denaturing Urea-PAA gel and silver staining. No degradation is observed under equal conditions but in the absence of HIT-ZnF (-). **(B)** Partial degradation of PAR. Both, the wild-type HIT-ZnF (wt) and the deadenylation inactive mutant (H274A) reduce the amount of long-chain PAR in the upper part of the gel. The cleavage products of wt and H274A differ in their electrophoretic mobility as illustrated in the zoomed region. **: DNA oligos remaining from PARP1 auto-PARylation to produce PAR. **(C)** Partial degradation of PAR in the absence (-), in the presence of HIT-ZnF (wt) and in the presence of HIT-ZnF and ribonuclease inhibitor (wt^{INH}). The presence of the inhibitor (SUPERase.In obtained from Life technologies GmbH) does *not* significantly affect the PAR degradation by HIT-ZnF. **(D)** Complete degradation of PAR by wild-type HIT-ZnF (wt) and by the double mutant K288E+H292E (2mu). The incubation of PAR without HIT-ZnF (-) does not lead to any PAR degradation. **(E)** The phosphorylation of serine 182, as evaluated by the S182D phosphorylation mimicking mutant of HIT-ZnF, does not affect PAR degradation activity. The band indicated (*) can be traced back to the staining of S182D protein itself, as the same band is observed for the protein without PAR (panel on right side). Of note: PAR used for (B) to (E) was produced *via* auto-PARylation of recombinant PARP1 as described in section 3.2. PAR utilized in (A) was kindly provided by Alexander Bürkle's lab (Konstanz).

experimental approach to distinguish the putative APTX degradation product from the PARG-produced one, but additionally raises interesting questions about potential intracellular effects induced by such a small APTX-generated molecule *e.g.* in signal transduction pathways.

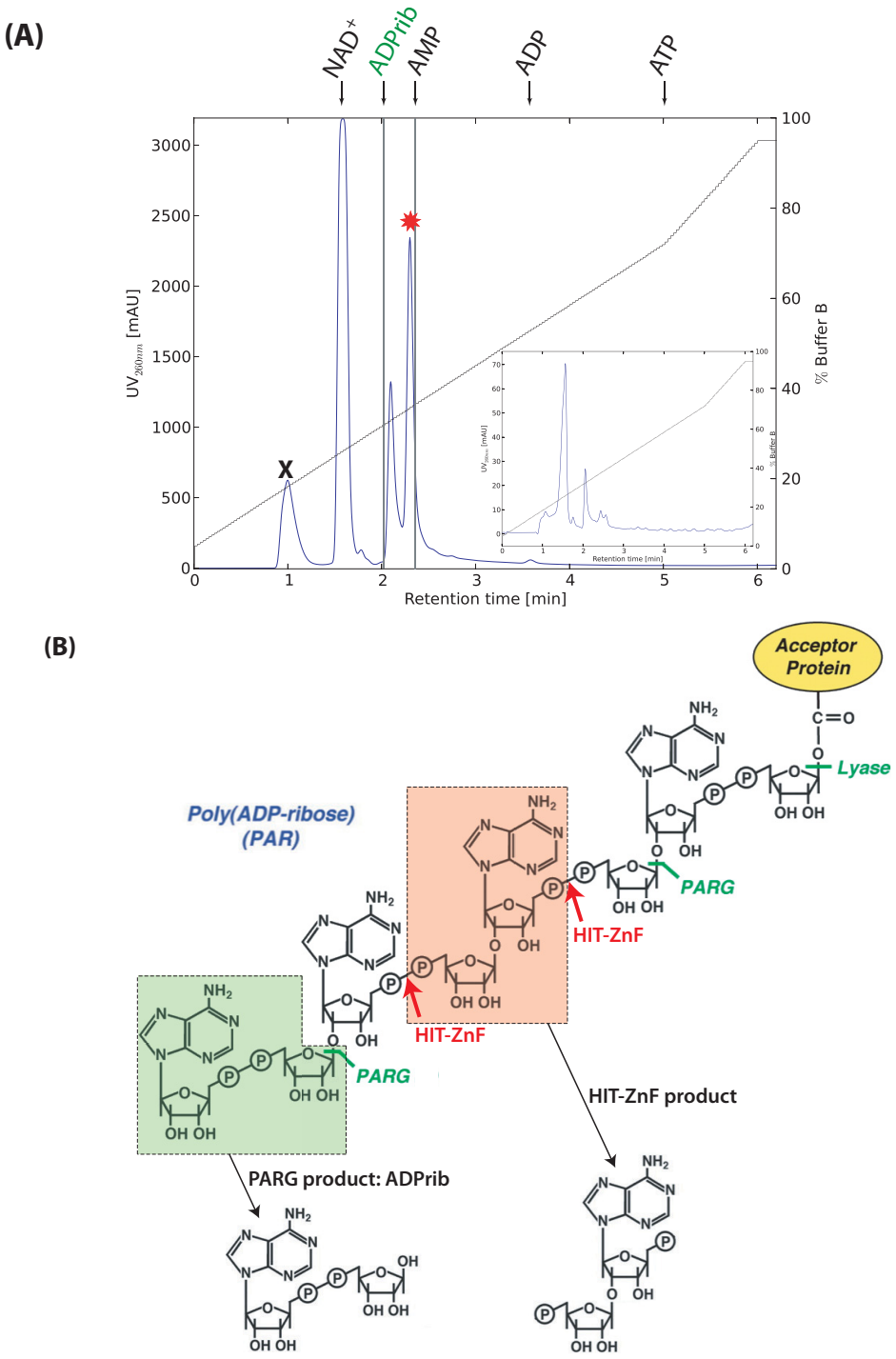


Figure 3.13: End product analysis of PAR cleavage by HIT-ZnF. (A) HPLC analysis of PAR cleavage by HIT- ZnF. PAR was incubated in the presence of HIT-ZnF. Assay was separated on a 125x4mm 4000-7 PEI column using a gradient (as depicted) of buffer A (2.5 mM Tris/phosphate pH 7.2) to 100% buffer B (2.5 mM Tris/phosphate pH 8.0, 1.5 M KCl). Retention time of reference substances indicated on top, (X) injection peak, (*) indicates the only peak not observed in presence of heat inactivated HIT-ZnF in a parallel assay (inset). Residual NAD^+ in the reaction remained from PARP1 auto-PARylation to produce PAR. ADPrib: ADP-ribose. **(B)** Degradation of PAR: The cleavage of PAR by poly(ADP-ribose) glycohydrolase (PARG) (indicated in green) results in formation of ADP-ribose. Red arrows denote the potential positions of PAR cleavage by HIT-ZnF yielding the putative cleavage product 2′-(5′-phosphoribosyl)-adenosine-5′-monophosphate. PAR structure modified from Krishnakumar and Kraus [53].

4 Discussion

4.1 Structural model of HIT-ZnF

The first part of this dissertation concerns the structural characterization of the HIT-ZnF domain of human Aprataxin and was inspired by the fact the the unique combination of a HIT domain and C_2H_2-type zinc finger is found in Aprataxin only and not in any other member of the HIT superfamily (section 1.2, [80]). However, due to the aggregating and precipitating behavior of all APTX constructs tested, the standard route of NMR spectroscopy could not be followed and different complementing approaches had to be utilized to extract information of structural properties of the HIT-ZnF. The first technique used here comprised CD spectroscopy as a well established and label-free method for characterizing the overall secondary structure content of the protein in question. The analysis of CD data depends on the availability of reference 3D structure data sets [61]. By utilizing the largest, publicly available CD data set of proteins with known 3D structures (PCDDB, [103]), we designed and I programmed a novel web-server based tool, named CAPITO, to provide a reliable estimate of secondary structure based on CD data (*Publication 1*).

Table 4.1: Result of secondary structure prediction based on CD data of HIT-ZnF. CD spectra were collected on a JASCO J-710 CD spectropolarimeter at 20°C in a 1 mm quartz cuvette. Protein concentration of HIT-ZnF was 3.55 μM in 50 mM Borat/NaOH pH 7.5 and was verified spectrophotometrically at 280 nm with an extinction coefficient of 34950 $M^{-1}cm^{-1}$ as obtained from ProtParam [35]. CAPITO uses different methods for secondary structure prediction. The result shown is the one obtained using the area difference method providing an reliable range of secondary structure content (see *Publication 2* for further details). CDSSTR, Selcon3 and CONTIN [93] are part of the CDPro software package. 'Irregular' refers to irregular structure content including turns.

	CAPITO	**CDSSTR**	**Selcon3**	**CONTIN**
α-helix	7-26%	18%	20%	21%
β-sheet	28-52%	27%	29%	25%
irregular	31-56%	55%	53%	54%

The evaluation of CD data of the HIT-ZnF employing our approach and comparison of the CAPITO result with the output of other programs, commonly used for the analysis of CD

data, is provided in Table 4.1. Interestingly, the amount of α-helix is found consistently in the range of 7-26% and is significantly smaller than the content of 41% back-calulated from the crystal structure of the APTX *S. pombe* ortholog Hnt3 [Gong et al. 37; PDB id: 3SP4]. The amount of β-sheet in HIT-ZnF is significantly elevated with 28-52% compared to 15% in Hnt3, respectively. This indicates, although both fulfill the same function with respect to the canonical deadenylation activity, that the structure of the HIT-ZnF may harbor additional structural elements and/or differs significantly in the structural organization of the HIT-ZnF domain. Furthermore, this fact supports the necessity of a refined three dimensional structure of the human HIT-ZnF domain irrespective of the availability of the *S. pombe* ortholog structure published during the course of this work [37, 100]. In particular, this refined structure is essential to provide a structural rational for the alternative activities of HIT-ZnF investigated here, that might well not be found for other organisms including *S. pombe*. The aggregating behavior of all APTX constructs in solution at moderate concentration, precluding standard liquid-state NMR approaches, was intended to be utilized by way of producing microcrystalline samples in combination with solid-state (ss) NMR spectroscopy. However, up to now only a small number of high-resolution protein structures have been solved *via* ssNMR (*e.g.* [19, 64, 107]). This is mainly due to the fact that suitable crystallization condition have to be identified and that there is currently no "straight-forward" standard route for structure determination *via* ssNMR and, therefore, often sample-specific methodology has to be developed [55]. Even if crystallization conditions are identified in high-throughput methods as common for X-ray crystallographic studies, the up-scaling towards the production of amounts of up to 5-10 mg often requires further adaptions of the already identified crystallization conditions. To establish ssNMR methodology for HIT-ZnF, I used two well characterized model systems, the protein domains GB1 and SH3. Both yielded high-resolution ssNMR spectra, that are comparable to published data (*Publication 2*, *e.g.* [33, 77]). Motivated by the fact, that the sample lifetime is often limited by method-related stress factors such as sample heating [30, 95] and that HIT-ZnF was susceptible to elevated temperatures, both proteins were used to develop a strategy to minimize data acquisition time employing the *simultaneous* acquisition of multiple NMR data sets (*Publication 2*). Despite significant efforts to find a suitable crystallization condition for the HIT-ZnF domain, only low-quality ssNMR data (Figure 4.1) were obtained, preventing further investigations using this technique. Although not applicable to APTX, I could provide a contribution to the development of ssNMR methodology as discussed at the end of the publication's manuscript.

NMR is an indirect method providing information on several chemical and structural properties of the (bio)molecule in question (*e.g.* chemical shifts, torsion angles or distances be-

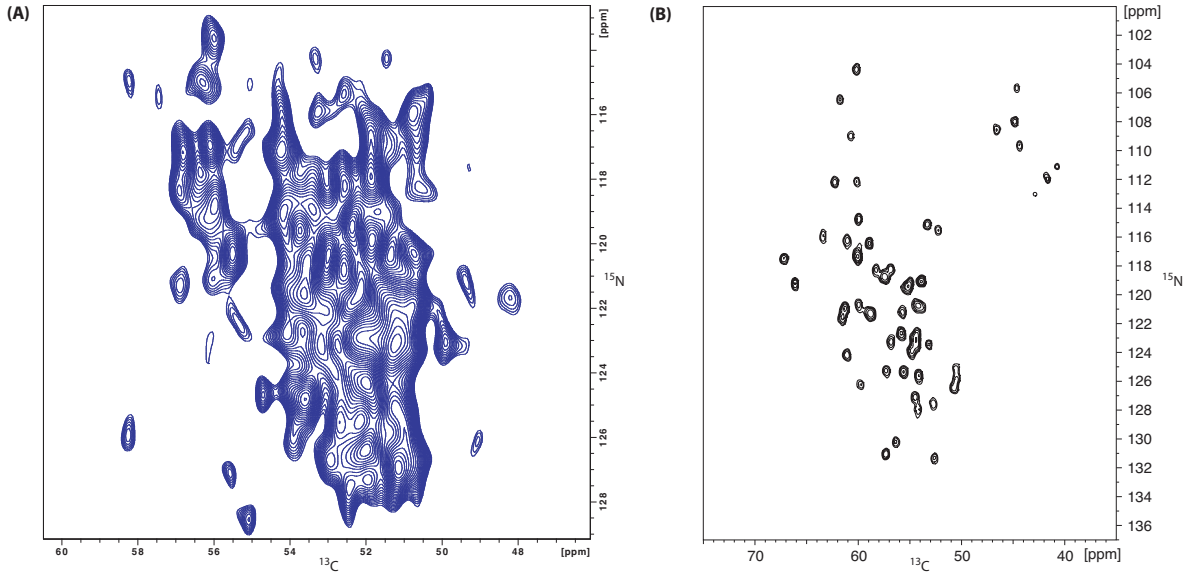


Figure 4.1: [^{13}C,^{15}N]-correlation spectra of HIT-ZnF and GB1 obtained *via* solid-state NMR and field strength of 500 MHz. (A) Microcrystals of HIT-ZnF were obtained by stepwise addition of ammonium sulfate (final concentration 1.7 M; [37]) to a HIT-ZnF solution of [^{13}C,^{15}N]-labeled pure protein (5 mg/ml) in 10 mM dTris/HCl pH 7.5, 150mM NaCl yielding the most promising but yet low-quality [^{13}C,^{15}N]-correlation spectra with an insufficient resolution and low signal-to-noise ratio. The spectrum shown was collected in a 2.5 mm rotor at a MAS frequency of 20 kHz and with 196 transients per t_1 increment. Higher MAS frequencies (up to 33.33 kHz) or higher field strength (750 MHz) did not result in better spectral quality. **(B)** For comparison, the same type of spectra obtained from microcrystalline GB1 yielding high-resolution data with only 16 transients per t_1 increment. Crystallization conditions as described in Publication 2.

tween atomic nuclei). This NMR information is used as restraints in the subsequent iterative structure calculation process. Based on the provided NMR data, structural models are generated and rated according to the degree of how well the respective structures fulfills the experimentally determined NMR constraints. The unconditional prerequisite for all NMR work is the assignment of resonances to their respective nuclei. This assignment is normally obtained *via* more complex, less sensitive and longer lasting experiment. Facing the apparent limitation of the HIT-ZnF to precipitate in less than 24 hours in solution, I decided to focus on very sensitive liquid-state (ls) NMR experiments only, and complemented these experiments with information obtained from the individual specific unlabeling of all 20 amino acids in HIT-ZnF. Following this approach (*Publication 3*), 78% of all observable resonances could be assigned to their respective backbone nuclei (Figure S6 and Table S2 of supplementary material of *Publication 3*). Missing assignments result from non-observability of at least one of the resonances that would be required for unambiguous assignment. The assigned backbone-resonance data of HIT-ZnF were directly used for the calculation of a first structural model of the human HIT-ZnF domain (Figure 4.2) utilizing the CS-ROSETTA approach. CS-ROSETTA is a ro-

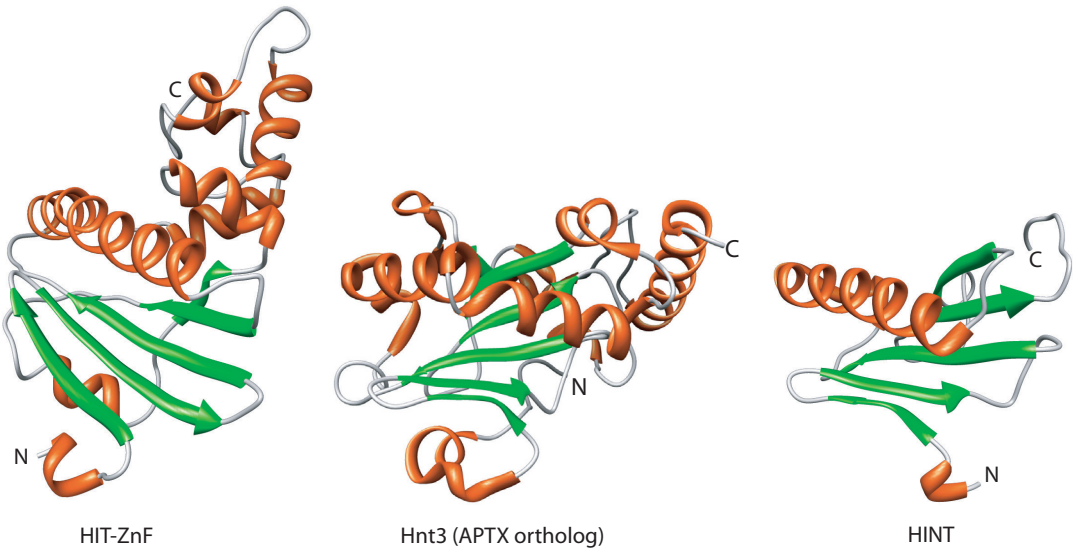


Figure 4.2: Structural model of HIT-ZnF, X-ray structures of the APTX's *S. pombe* ortholog Hnt3 and of a classical member of the HIT superfamily, HINT. HIT-ZnF (left) and its *S. pombe* ortholog Hnt3 (middle) share a sequence homology of approx. 23% identity and 34% similarity, respectively. PDB codes of the the structures shown are 3SP4 for Hnt3, 6RHN for HINT and were published by Gong et al. [37] and Brenner et al. [11], respectively. The HIT-ZnF structure was calculated using CS-ROSETTA's RASREC algorithm [56]. Further refinement is necessary before a detailed comparison is reasonable.

bust protocol for *de novo* protein structure model generation and utilizes the fact, that short peptide fragments characterized by a distinct chemical shifts often harbor a conserved local structure [56]. These local structures are combined to a global fold and optimized to fulfill inherent restraints (*e.g.* allowed torsion angles). The CS-ROSETTA model of the human APTX HIT-ZnF domain (Figure 4.2) exhibits a number of similarities to the *S. pombe* homologue and to rabbit histidine nucleotide-binding protein (HINT), another member of the HIT superfamily. Most striking is the similarity of a long α-helix connected to a four-stranded β-sheet. However, the back-calculated secondary structure content of the HIT-ZnF model yields higher α-helical content (approx. 40%) than the α-helical content calculated based on the CD values (< 26%, Table 4.1). This indicates, that the CS-ROSETTA model of HIT-ZnF needs further refinement before a detailed comparison (including r.m.s.d. values) to the *S. pombe* ortholog Hnt3 is reasonable. This refinement of the structure should be achievable by the inclusion of methyl-amide and methyl-methyl distance constraints. These distance constraints can be obtained from ^{15}N samples with specific [^{13}C]-methyl labeling [83, 90]. Such samples have ben prepared and initial data have already been collected, but still need to be evaluated[1]. However, the assignment of backbone resonances, already deposited in the publicly accessible BMRB

[1]The assignment of methyl-groups will be carried out utilizing site-directed mutants where necessary.

database[2], *per se* can be used for the characterization of interactions with substrates or small chemical compounds *via* NMR methods.

4.2 Implications of APTX's additional nucleolytic activities

As described in detail in chapter 3, I observed the following HIT-ZnF activities *in vitro*:
(1) Cleavage of canonical substrate (adenylated dsDNA) indicating functional integrity of recombinant HIT-ZnF and dispensability of FHA domain for deadenylation activity as also shown previously by [51]. The central residue of the HIT motif (H274) is essential for deadenylation activity, as the H274A mutant is inactive on adenylated dsDNA in accordance with [80].
(2) HIT-ZnF does not need double-stranded DNA for deadenylation, but also acts on single-stranded adenylated DNA.
(3) The amino group in the 6 position of the adenine base (6-aminopurine) is not essential for substrate recognition nor for cleavage, as 2-aminopurine is also recognized as substrate and cleaved.
(4) HIT-ZnF is active on single-stranded chimeric RNA-DNA: 5′-exonucleolytic cleavage of single 5′-attached ribonucleotides but no 3′-exonucleolytic activity is observed; a single internally embedded ribonucleotide is recognized and HIT-ZnF incises at 3'-position of rNMP. No activity is detected on internally embedded ribonucleotides if the chimeric RNA-DNA strand is annealed to a complementary DNA strand.
(5) K288E and H292E mutants, but not the H274A mutant, abolish RNAse-like exonuclease activity on ribonucleotides attached to the 5'-end of ssDNA.
(6) Degradation of ssRNA is not affected by the mutants tested so far, but is sensitive to ribonuclease inhibitor (SUPERase.In; Life technologies GmbH).
(7) Decapping of m^7-GpppRNA is not susceptible to ribonuclease inhibitor. Activity of mutants was not tested yet.
(8) HIT-ZnF degrades poly(ADP-ribose) (PAR). PAR-degrading activity is not significantly affected by the presence of ribonuclease inhibitor. Furthermore, PAR-degrading activity is not solely dependend on His274 as the H274A mutant (inactive in deadenylation) still degrades PAR. However, different migration of H274A-generated products is observed on denaturing Urea-PAA gels as compared to wild-type HIT-ZnF. The degradation product of PAR gener-

[2]The assigned chemical sifts of the HIT-ZnF construct were deposited in the BioMagResBank [101] under accession number 19182 and will immediately be released upon acceptance of Publication 3.

ated by wild-type HIT-ZnF is different from ADP-ribose (=PARG degradation product) and different from AMP as found in HPLC analyses.

Importantly, the RNAse-like 5′-exonucleolytic activity, as well as the PAR-degrading activity of HIT-ZnF are sensible to mutations in HIT-ZnF. Hence it is extremely unlikely, that these two activities are a result of a co-purified contaminant. The substrate used here is rather similar to the natural substrate of RNAse H proteins that typically act on RNA-DNA hybrids [73]. However, there are important differences between the activity observed here and the activities mediated by RNAse H enzymes. Firstly, RNAse *H1* does not act on *single* ribonucleotides embedded in DNA [92], where as HIT-ZnF incises 3′ of the single rNMP (section 3.3.1, Figure 3.8). Secondly, even though RNAse *H2* is active on single ribonucleotides incorporated in DNA, this activity is only present in the context of double strand, which is essential for substrate binding [92]. In contrast, HIT-ZnF does not operate on chimeric RNA-DNA if annealed to a complementary strand. Therefore, the RNAse H1-like as well as RNAse H2-like activities, both present in *E. coli* [47, 73, 81], can practically be excluded as cause for the observed RNAse H-like activity. The degradation of canonical ssRNA, however, was not sensitive to mutations tested so far. Although the purification procedure of HIT-ZnF comprises several steps (Ni-NTA affinity, ion exchange, as well as a final size exclusion chromatography), a co-purification of an *E. coli* RNAse can formally not be excluded at present. I therefore suggest, that future work should include the screening of further mutants with respect to their capability to cleave ssRNA. Non-hydrolyzable RNA analogs (*e.g.* phosphorothioate oligos), that are most probably bound but not cleaved by HIT-ZnF, can be used for the identification of binding sites *via* lsNMR methods, thereby providing a basis for targeted mutations. Once ssRNAse-inactive mutants of HIT-ZnF are available, the decapping activity, that is currently complicated by the RNAse activity, should be characterized in more detail. However, the fact, that 6-aminopurine (canonical adenylate), 2-aminopurine as well as m^7-G (RNA-cap) is recognized by HIT-ZnF as substrate, indicates a low overall substrate specificity for this enzyme. This might explain the additional unanticipated activities of HIT-ZnF on the different *ribo*nucleotide-containing substrates.

4.2.1 APTX's potential role in ribonucleotide excision repair

The (mis)incorporation of *ribo*nucleotides (rNMPs) into DNA during replication is stated to be a surprisingly frequent event with approx. 1 rNMP per 0.7 kb [92]. Due to the presence of a reactive 2′-hydroxyl group in rNMPs, this misincorporation results in a higher susceptibility to single strand nicks / breaks of the resulting DNA strand. The system to remove rNMP

embedded in DNA is called ribonucelotide excision repair (RER) and depends on RNAse H2 and FEN1 [92]. RNAse H2 incises at the 5′-position of the rNMP and FEN1 finally removes

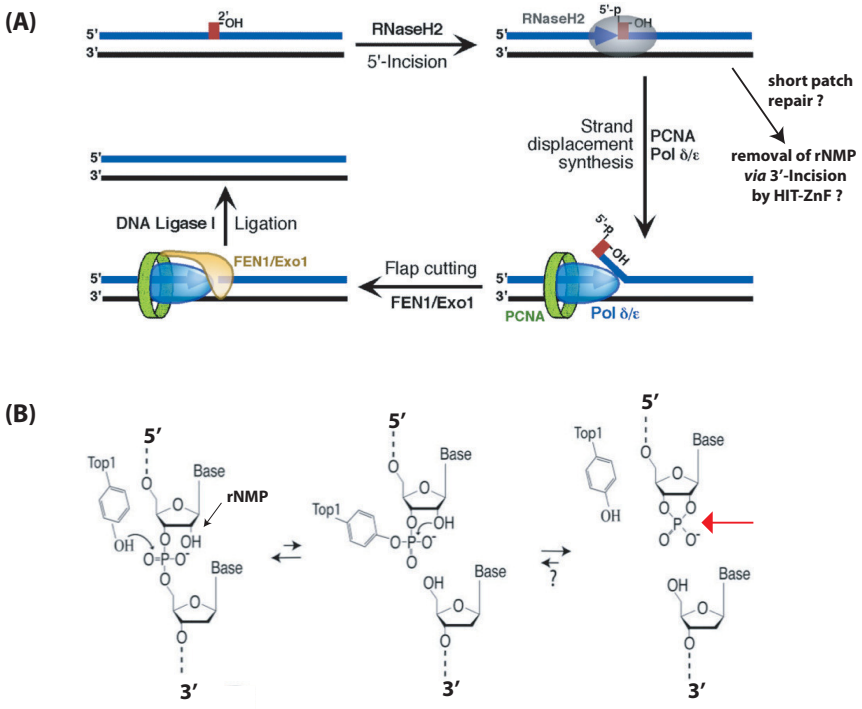


Figure 4.3: Model of ribonucleotide excision repair (RER). (A) Long patch repair pathway for removing misincorporated ribonucleotides depends on RNAse H2 and FEN1. The redundant functions of FEN1 and Exo1 as well as of Pol δ and Pol ϵ, respectively, are indicated. An alternative putative short patch repair pathway involving the HIT-ZnF domain of APTX after RNAse H2 incision might result in the removal of the rNMP. See text for further details. Figure adapted from Sparks et al. [92] **(B)** Toposiomerase 1 (Top1) mediated rNMP removal resulting in a 2′, 3′-cyclo-phosphate group 3′-attached to the ribonucleotide (red arrow). Figure modified from Sekiguchi and Shuman [87].

a longer DNA fragment ("flap cutting") including the 5′-rNMP from the DNA strand (Figure 4.3 A) during "long-pach" repair (see section 1.3). I observed single rNMP recognition in a ssDNA context and subsequent incision by recombinant HIT-ZnF *in vitro*. Contrary to the cleavage mode of RNAse H2, HIT-ZnF incises at the 3′-position of the embedded rNMP. I speculate, that the HIT-ZnF domain might be involved in RER. In particular, it might act in combination with the RNAse H2 effecting the immediate release of the rNMP from the DNA without the necessity for flap-cutting mediated by FEN1. Importantly, this would allow for a short patch repair and provide for a backup-system to correct misincorporated single ribonucleotides and to foster genomic stability. As the RNAse-H like activity of HT-ZnF *in vitro* is found in the context of single-stranded DNA only, this scenario would require an APTX-associated helicase or the presence of local structural fluctuations ("breathing") after RNAse H2 incision 5′ of the rNMP resulting in a pseudo-single stranded state of the rNMP - similar to

the situation in the adenylated dsDNA, which HIT-ZnF is able to process (section 3.3.1, Figure 3.7 A). The double-stranded RNA-DNA substrate employed for the RNAse H-like activity test here comprised of two complementary oligonucleotides - thereby not exactly matching the situation found after a 5'-incision by RNAse H2. As this experiment would point out HIT-ZnF's potential role in RER even in the context of double-strands, it will immediately be addressed in the near future. Even if HIT-ZnF is not active in the aforementioned double-stranded context after RNAse H2 incision, this unique activity should be characterized in more detail as neither RNAse H1 nor H2 act on *single-stranded* DNA (Figure Figure 4.3 A). Interestingly, besides the RNAse H2/FEN1 mediated removal of misincorporated rNMP, there is evidence for an alternative pathway involving topoisomerase 1 (Top1) [52]. There, Top1 incises at the 3-′ position of the ribonucleotide leaving a 2′, 3′-cyclo-phosphate terminus (Figure 4.3 B) [87]. Hence, Top1 and HIT-ZnF cleave at the same position in chimeric RNA-DNA, although to my knowledge the Top1 activity is restricted to double-stranded DNA, whereas HIT-ZnF is active on ssDNA. Interestingly, I postulated an DNA cleavage product with a 3′-attached ribonucleotide carrying a 3′-phosphate group (section 3.3.1, Figure 3.8 E). As this cleavage product is similar to the Top1-generated product with a 2′, 3′-cyclo-phosphate group 3'-attached to the ribonucleotide (Figure 4.3 B), I suggest to analyze the nature of the HIT-ZnF cleavage product with respect to the nature of the additional 3′-phosphate group, in particular to check for a cyclo-phosphate. The latter one can be identified on a denaturing Urea-PAA gel utilizing RNAse A and alkaline phosphatase treatment resulting in the removal of the phosphate group [87].

4.2.2 APTX's potential role in poly(ADP-ribose) degradation

Poly(ADP-ribosyl)ation (PARylation) is an essential protein modification involved in many cellular responses such as DNA repair, transcriptional regulation, centromere function or apoptosis [15]. PARylation is catalyzed by the family of poly(ADP-ribose) polymerases, which utilize NAD^+ as precursor molecule for the synthesis of (branched) polymers with an ADP-ribose chain length of approx. up to 200 units [14]. During the short patch repair of DNA single-strand breaks, PARP1 is an essential sensor molecule that is activated upon detection of a break resulting in the PARylation of its automodification domain and other nuclear proteins [26]. Importantly, this poly(ADP-ribose)ylation of PARP1 is essential for the recruitment of repair enzymes and scaffolding proteins such as XRCC1 to the sites of DNA damage (see section 1.3 for further details). Obviously, the "marking" of DNA damage sites by the presence of poly(ADP-ribose) (PAR) should only be of temporary nature and has to be removed in order to deallocate the repair factors after the repair process. PAR degradation is predomi-

nately mediated by poly(ADP-ribose)glycohydrolases (PARGs), that hydrolyze the glycosidic linkage between the ribose units of PAR (section 3.3.3, Figure 3.13) resulting in ADP-ribose units as cleavage products [10, 91]. Motivated by the functional significance of PAR in DNA damage repair and by the chemical similarities between PAR and adenylated DNA (section 3.3.3, Figure 3.11), I tested the recombinant HIT-ZnF domain for PAR-degrading activity. Here I could show, that HIT-ZnF is indeed capable of degrading PAR. However this activity is not solely dependent on the central histidine of the HIT motif (His274), as the H274A mutant still degrades PAR although the resulting cleavage products differ from the wild-type HIT-ZnF-generated cleavage product as analyzed on denaturing Urea-PAA gels (section 3.3.3, Figure 3.12 B). HPLC-analyses of the PAR-degradation by wild-type HIT-ZnF indicate, that the cleavage product is different from ADP-ribose (= cleavage product of PARGs) and different from AMP (section 3.3.3, Figure 3.13). Based on the cleavage mode of HIT-ZnF on adenylated DNA (cleavage between the two phosphate groups connecting the adenylate to the DNA) [80], I postulate that the cleavage product is a 2′-(5′-phosphoribosyl)-adenosin-5′-monophosphat (section 3.3.3, Figure 3.13). As small phosphorylated ribonucleotide-derived molecules such as cAMP carry out essential functions as second messenger in various signal transduction pathways, the exact chemical nature of this cleavage product should be immediately addressed in the future (*e.g.* by mass spectrometry or NMR). Up to my knowledge, the postulated HIT-ZnF-generated cleavage product has not yet been linked to any activity *in vivo*, and therefore, it would be really interesting to analyze a potential second messenger effect of this product on *e.g.* cell cycle regulation. To address the role of APTX's HIT-ZnF within PAR catabolism, it is essential that further work identifies the (other) residues involved in the catalysis of the cleaving reaction. The absence of APTX *in vivo* causes reduction of the level of a number of proteins involved in single strand break (SSB) repair, including PARP1 [42]. Lower level of PARP1 will probably lead to a reduced production of PAR upon DNA damage but at the same time potentially to a reduced degradation of PAR due to the missing APTX. Hence, an APTX mutant deficient in PAR degradation but proficient in its other activities would allow to analyze APTX's role in PAR metabolism in an otherwise undisturbed *in vivo* background. This in turn would validate my *in vitro* observation but also open a new route into understanding of DNA damage repair.

4.3 Suggested future work plan

My major suggestions for further characterization and confirmation of the different nucleolytic activities of HIT-ZnF are summarized below.
(1) Clarify if the ssRNAse activity is due to a contaminant of HIT-ZnF preparations: Utilize different expression systems (including cell free expression systems such as rabbit reticulocyte lysate) and different purification procedures to conclusively exclude a potential contaminant as cause for the activity observed. Perform an intensive NMR-based screening utilizing non-hydrolyzable ribonucleotides to identify residues involved in the ssRNA degradation by HIT-ZnF. Besides characterization of putative ssRNAse activity, ssRNAse-inactive mutants are essential for further investigation of the HIT-ZnF decapping activity, the analysis of which is compromised by the ssRNAse activity.
(2) Confirm the ribonucleotide excision repair activity of HIT-ZnF on "physiological" chimeric RNA-DNA substrates in a double-stranded context after RNAse H2 incision. Quantify amount of rNMP incorporation in APTX-proficient and APTX-deficient cells (already provided by K. Caldecott; University of Sussex) to substantiate APTX's role in that repair pathway.
(3) Characterize and confirm PAR degradation: Identify further residues of HIT-ZnF involved in PAR-degradation. Express PAR-degrading-inactive mutants in APTX-deficient cells and quantify PAR level kinetics after DNA damage with and without concurrent inhibition of PARGs to separate the putative HIT-ZnF activity from PARG. Identify the chemical nature of the *in vitro* HIT-ZnF-generated cleavage product and address its presence in APTX-proficient and APTX-deficient cells.

4.4 Closing remarks

As described in this dissertation, the HIT-ZnF domain of human APTX harbors a variety of nucleolytic activities and acts on different *ribo*nucleotide-containing substrates. Each of this substrates, namely adenylated DNA, capped RNA, ribonucleotides 5′-attached or embedded in DNA as well as poly(ADP-ribose), is involved in different cellular processes ranging from DNA repair to transcriptional regulation. It is of essential significance to separate each of the activity harbored by HIT-ZnF by means of specific mutants to have a tool for *in vivo* confirmation and characterization of each of the aforementioned activities. Even if only *one* of the observed activities can be successfully confirmed *in vivo*, my work has significantly contributed to the current molecular understanding of APTX's function and of the respective physiological process it is involved in. Therefore, we prefer to publish the data I presented here after the significance *in vivo* has been elucidated.

„Das Schönste, was wir entdecken können, ist das Geheimnisvolle.“ (Albert Einstein)
[The most beautiful thing we can discover is the mysterious.]

5 Summary

This dissertation addresses the structural and functional characterization of human Aprataxin, a DNA damage repair protein with a unique activity not found for any other protein involved in DNA repair. Single strand breaks (SSBs) threaten genomic stability, as more than 10.000 of such SSBs arise per cell and day. During the repair of SSBs, so-called "dirty" DNA breaks may arise, which cannot be ligated resulting in the formation of an abortive ligation intermediate: 5′-adenylated DNA. The 5′-adenylate cannot be processed by Ligase 3 itself but requires Aprataxin to be removed from the DNA. Aprataxin belongs to the superfamily of histidine triad (HIT) proteins, but its domain organization including a catalytically active HIT motif and a C_2H_2-type zinc finger (ZnF) is unique. Furthermore, only this HIT-ZnF domain is conserved among all APTX orthologs. This work is focused on the HIT-ZnF due to its functional significance. A structural characterization of HIT-ZnF was hampered by its high precipitation propensity already at moderate concentrations, precluding structure determination in solution *via* standard NMR methodology. By utilizing the specific individual *un*labeling of all 20 amino acids, 78% of the observable backbone resonances could be assigned. Based on the assigned chemical shifts a structural CS-ROSETTA-based model of the HIT-ZnF domain of human Aprataxin is presented, which reveals structural similarities to other members of the HIT superfamily. The resonance assignment also provides a basis for a detailed characterization of the interaction between HIT-ZnF, its substrate(s) and small chemical compounds by NMR spectroscopy in the future. Surprisingly, functional analysis revealed that HIT-ZnF harbors additional nucleolytic activities on *ribo*nucleotide-containing substrates different from the canonical adenylated DNA. I found that HIT-ZnF decaps m^7-GpppRNA, cleaves the phosphodiester bond 3′ of *ribo*nucleotides attached 5′ to or embedded in single-stranded DNA but not when present at the 3′ terminus of ssDNA. Furthermore, I provide strong *in vitro* evidence that HIT-ZnF is capable of degrading poly(ADP-ribose), one of the major protein modifications signalling DNA damage and control DNA damage response. Although the functional significance of these *in vitro* observations has yet to be evaluated *in vivo*, this work provides a solid basis for a comprehensive future assessment of Aprataxin's activities and a role beyond resolving an abortive ligation intermediate during single-strand break repair.

6 Zusammenfassung

Diese Dissertation behandelt die strukturelle und funktionelle Charakterisierung des humanen Aprataxin, welches eine enzymatische Aktivität besitzt, die bei keinem anderen Protein der DNA-Reparatur zu finden ist. DNA-Einzelstrangbrüche kompromittieren die genomische Stabiltät, da diese häufiger als 10.000-mal pro Zelle und Tag auftreten. Während der DNA-Einzelstrangbruch-Reparatur führen sogenannte “unsaubere” Bruchenden dazu, dass diese nicht ligiert werden können. An solchen Stellen ensteht ein abortives Ligations-Zwischenprodukt: ein rApp-DNA Terminus. Dieses sogenannte 5′-Adenylat kann von der DNA Ligase 3 nicht prozessiert werden. Hier ist Aprataxin notwendig um das Adenylat von der DNA zu entfernen. Aprataxin ist ein Mitglied der Histidin-Triade (HIT)-Protein Superfamilie. Seine Domänenorganisation ist einzigartig. Neben nicht konservierten akzessorischen Domänen enthalten alle Aprataxin-Orthologe eine konservierte katalytisch aktive Domäne mit einem HIT-Motif und einem Zinkfinger (ZnF) vom C_2H_2-Typ. Diese Arbeit konzentriert sich auf diese HIT-ZnF Domäne aufgrund deren funktionellen Bedeutung. Eine strukturelle Charakterisierung dieser Domäne wurde durch deren Eigenschaft schon bei moderaten Konzentrationen zu präzipitieren erheblich erschwert. Standardmethoden der NMR-Spektroskopie zur Strukturaufklärung konnten nicht angewandt werden. Das Einführen jeweils eines unmarkierten Aminosäuretyps in sonst isotopenmarkiertes Protein für alle 20 Aminosäuren erlaubte die Zuordnung von 78% der beobachtbaren Proteinrückgrat-Resonanzen. Deren chemischen Verschiebungen erlaubten die Berechnung eines Strukturmodells der HIT-ZnF Domäne mittels CS-ROSETTA. HIT-ZnF weist strukturelle Ähnlichkeiten zu weiteren Mitgliedern der HIT-Superfamilie auf. Die erfolgte Resonanzzuordnung erlaubt eine weitere detailierte Charakterisierung der Interaktionen zwischen HIT-ZnF und Substraten bzw. niedermolekularen Verbindungen. Überaschenderweise ergaben meine Aktivitätsuntersuchungen, dass HIT-ZnF noch andere nukleolytische Aktivitäten gegenüber *Ribo*nukleotid-enthaltenden Nukleinsäuresubstraten, die sich von 5’-adenlylierter DNA unterscheiden, entfaltet. Ich konnte *in vitro* zeigen, dass HIT-ZnF *in vitro* die Cap-Struktur von m^7-GpppRNA entfernt und dass HIT-ZnF die Phosphodiesterbindung 3′ von *Ribo*nukleotiden, die ihrerseits am 5′-Terminus einzelsträngiger DNA sitzen, spaltet jedoch keine am 3′-Terminus gebundenen *Ribo*nukleotide abspaltet.

Intern in Einzelstrang-DNA eingebettete Einzel*ribo*nukleotide werden ebenfalls als Substrat akzeptiert und 3′-seitig durch HIT-ZnF gespalten. Zusätzlich präsentiere ich experimentelle Evidenz dafür, das HIT-ZnF poly(ADP-ribose) *in vitro* spaltet. Poly(ADP-ribose) stellt eine der kovalenten Proteinmodifikationen dar, die DNA Schäden signalisieren und die die DNA-Schadensantwort steuern. Auch wenn die funktionelle Bedeutung dieser *in vitro* Beobachtungen noch *in vivo* validiert werden muss, bildet die vorgelegte Arbeit eine solide Grundlage für weiterführende Untersuchungen der biochemischen Aktivität des Aprataxin und dessen physiologische Rolle, die weit über das Entfernen eines abortiven Ligations-Zwischenprodukt bei der Reparatur von Einzelstrangbrüchen hinaus zu gehen scheint.

Literaturverzeichnis

[1] R T Abraham. Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes & Development*, 15(17):2177–2196, September 2001.

[2] Ivan Ahel, Ulrich Rass, Sherif F El-Khamisy, Sachin Katyal, Paula M Clements, Peter J McKinnon, Keith W Caldecott, and Stephen C West. The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates. *Nature*, 443(7112): 713–716, September 2006.

[3] Jean-Christophe Amé, Thomas Kalisch, Françoise Dantzer, and Valérie Schreiber. Purification of recombinant poly(ADP-ribose) polymerases. *Methods in molecular biology (Clifton, N.J.)*, 780:135–152, 2011.

[4] O T Avery, C M Macleod, and M McCarty. Studies on the chemcial natureof the substance including Transformation ofpneumiciccal types : Induction of transformation by a Desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of experimental medicine*, 79(2):137–158, February 1944.

[5] A K Banerjee. 5'-terminal cap structure in eucaryotic messenger ribonucleic acids. *Microbiological reviews*, 44(2):175–205, June 1980.

[6] C Barbot, P Coutinho, R Chorão, C Ferreira, J Barros, I Fineza, K Dias, J Monteiro, A Guimarães, P Mendonça, M do Céu Moreira, and J Sequeiros. Recessive ataxia with ocular apraxia: review of 22 Portuguese patients. *Archives of neurology*, 58(2): 201–205, February 2001.

[7] Olivier J Becherel, Nuri Gueven, Geoff W Birrell, Valérie Schreiber, Amila Suraweera, Burkhard Jakob, Gisela Taucher-Scholz, and Martin F Lavin. Nucleolar localization of aprataxin is dependent on interaction with nucleolin and on active ribosomal DNA transcription. *Human molecular genetics*, 15(14):2239–2249, July 2006.

[8] G Bertani. Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. *Journal of Bacteriology*, 62(3):293–300, September 1951.

[9] Philip C Bevilacqua and Rieko Yajima. Nucleobase catalysis in ribozyme mechanism. *Current opinion in chemical biology*, 10(5):455–464, October 2006.

[10] M E Bonicalzi, J-F Haince, A Droit, and G G Poirier. Poly-ADP-ribosylation in health and disease. *CMLS Cellular and Molecular Life Sciences*, 62(7-8):739–750, April 2005.

[11] C Brenner, P Garrison, J Gilmour, D Peisach, D Ringe, G A Petsko, and J M Lowenstein. Crystal structures of HINT demonstrate that histidine triad proteins are GalT-related nucleotide-binding proteins. *Nature structural biology*, 4(3):231–238, March 1997.

[12] Charles Brenner. *Histidine Triad (HIT) Superfamily*. John Wiley & Sons, Ltd, Chichester, UK, May 2001.

[13] Charles Brenner. Hint, Fhit, and GalT: function, structure, evolution, and mechanism of three branches of the histidine triad superfamily of nucleotide hydrolases and transferases. *Biochemistry*, 41(29):9003–9014, July 2002.

[14] Alexander Bürkle. Poly(ADP-ribose). The most elaborate metabolite of NAD+. *The FEBS journal*, 272(18):4576–4589, September 2005.

[15] Alexander Bürkle and László Virág. Poly(ADP-ribose): Paradigms and paradoxes. *Molecular Aspects of Medicine*, pages 1–20, January 2013.

[16] Keith W Caldecott. XRCC1 and DNA strand break repair. *DNA repair*, 2(9):955–969, September 2003.

[17] Keith W Caldecott. DNA single-strand breaks and neurodegeneration. *DNA repair*, 3 (8-9):875–882, 2004.

[18] Keith W Caldecott. Single-strand break repair and genetic disease. *Nature reviews Genetics*, 9(8):619–631, August 2008.

[19] Federica Castellani, Barth van Rossum, Annette Diehl, Mario Schubert, Kristina Rehbein, and Hartmut Oschkinat. Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature*, 420(6911):98–102, November 2002.

[20] J E Cleaver. Defective repair replication of DNA in xeroderma pigmentosum. *Nature*, 218(5142):652–656, May 1968.

[21] Paula M Clements, Claire Breslin, Emma D Deeks, Philip J Byrd, Limei Ju, Pawel Bieganowski, Charles Brenner, Maria-Céu Moreira, A Malcolm R Taylor, and Keith W Caldecott. The ataxia-oculomotor apraxia 1 gene product has a role distinct from ATM and interacts with the DNA strand break repair proteins XRCC1 and XRCC4. *DNA repair*, 3(11):1493–1502, November 2004.

[22] Roland Contreras, Hilde Cheroutre, Wim Degrave, and Walter Fiers. Simple, efficient in vitro synthesis of capped RNA useful for direct expression of cloned eukaryotic genes. *Nucleic Acids Research*, 10(20):6353–6362, 1982.

[23] F H Crick, L Barnett, C Brenner, and R J Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192:1227–1232, December 1961.

[24] H Date, O Onodera, H Tanaka, K Iwabuchi, K Uekawa, S Igarashi, R Koike, T Hiroi, T Yuasa, Y Awaya, T Sakai, T Takahashi, H Nagatomo, Y Sekijima, I Kawachi, Y Takiyama, M Nishizawa, N Fukuhara, K Saito, S Sugano, and S Tsuji. Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. *Nature genetics*, 29(2):184–188, October 2001.

[25] Hidetoshi Date, Shuichi Igarashi, Yasuteru Sano, Toshiaki Takahashi, Tetsuya Takahashi, Hiroki Takano, Shoji Tsuji, Masatoyo Nishizawa, and Osamu Onodera. The FHA domain of aprataxin interacts with the C-terminal region of XRCC1. *Biochemical and biophysical research communications*, 325(4):1279–1285, December 2004.

[26] G De Murcia and J Ménissier-de Murcia. Poly(ADP-ribose) polymerase: a molecular nick-sensor. *Trends in biochemical sciences*, 19(4):172–176, April 1994.

[27] G Dianov and J Parsons. Co-ordination of DNA single strand break repair. *DNA repair*, 6(4):454–460, April 2007.

[28] Higinio Dopeso, Silvia Mateo-Lozano, Elena Elez, Stefania Landolfi, Francisco Javier Ramos Pascual, Javier Hernández-Losa, Rocco Mazzolini, Paulo Rodrigues, Sarah Bazzocco, Maria Josep Carreras, Eloy Espín, Manel Armengol, Andrew J Wilson, John M Mariadason, Santiago Ramon Y Cajal, Josep Tabernero, Simo Schwartz, and Diego Arango. Aprataxin tumor levels predict response of colorectal cancer patients to irinotecan-based treatment. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 16(8):2375–2382, April 2010.

[29] D Durocher and S P Jackson. DNA-PK, ATM and ATR as sensors of DNA damage: variations on a theme? *Current opinion in cell biology*, 13(2):225–231, April 2001.

[30] Sergey V Dvinskikh, Vasco Castro, and Dick Sandström. Heating caused by radio-frequency irradiation and sample rotation in13C magic angle spinning NMR studies of lipid membranes. *Magnetic Resonance in Chemistry*, 42(10):875–881, September 2004.

[31] Sherif F El-Khamisy, Sachin Katyal, Poorvi Patel, Limei Ju, Peter J McKinnon, and Keith W Caldecott. Synergistic decrease of DNA single-strand break repair rates in mouse neural cells lacking both Tdp1 and aprataxin. *DNA repair*, 8(6):760, June 2009.

[32] Gary Felsenfeld and Mark Groudine. Controlling the double helix. *Nature*, 421(6921): 448–453, January 2003.

[33] W Trent Franks, Donghua H Zhou, Benjamin J Wylie, Brian G Money, Daniel T Graesser, Heather L Frericks, Gurmukh Sahota, and Chad M Rienstra. Magic-Angle Spinning Solid-State NMR Spectroscopy of the β1 Immunoglobulin Binding Domain of Protein G (GB1): 15N and 13C Chemical Shift Assignments and Conformational Analysis. *J. Am. Chem. Soc*, 127(35):12291–12305, September 2005.

[34] Heather L Frericks Schmidt, Lindsay J Sperling, Yi Gui Gao, Benjamin J Wylie, John M Boettcher, Scott R Wilson, and Chad M Rienstra. Crystal Polymorphism of Protein GB1 Examined by Solid-State NMR Spectroscopy and X-ray Diffraction. *The Journal of Physical Chemistry B*, 111(51):14362–14369, December 2007.

[35] Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, S'everine Duvaud, Marc R Wilkins, Ron D Appel, and Amos Bairoch. Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*, pages 571–607. Humana Press, Totowa, NJ, 2005.

[36] Bryan A Gibson and W Lee Kraus. New insights into the molecular and cellular functions of poly(ADP-ribose) and PARPs. *Nature Reviews Molecular Cell Biology*, 13(7): 411–424, June 2012.

[37] Yong Gong, Deyu Zhu, Jingjin Ding, Chuan-Na Dou, Xiaoming Ren, Lichuan Gu, Tao Jiang, and Da-Cheng Wang. Crystal structures of aprataxin ortholog Hnt3 reveal the mechanism for reversal of 5'-adenylated DNA. *Nature structural & molecular biology*, pages 1–3, October 2011.

[38] Ewa Grudzien-Nogalska, Janusz Stepinski, Jacek Jemielity, Joanna Zuberek, Ryszard Stolarski, Robert E Rhoads, and Edward Darzynkiewicz. Synthesis of anti-reverse cap

analogs (ARCAs) and their applications in mRNA translation and stability. *Methods in Enzymology*, 431:203–227, 2007.

[39] Nuri Gueven, Olivier J Becherel, Amanda W Kijas, Philip Chen, Orla Howe, Jeanette H Rudolph, Richard Gatti, Hidetoshi Date, Osamu Onodera, Gisela Taucher-Scholz, and Martin F Lavin. Aprataxin, a novel protein that protects against genotoxic stress. *Human molecular genetics*, 13(10):1081–1093, May 2004.

[40] Matthias Habeck, Christine Z hlke, Karl H P Bentele, Stephan Unkelbach, Wolfram Kre, Katrin B rk, Eberhard Schwinger, and Yorck Hellenbroich. Aprataxin mutations are a rare cause of early onset ataxia in Germany. *Journal of Neurology*, 251(5):591–594, May 2004.

[41] J L Hancock. *Biochemical Characterization of Aprataxin, the protein deficient in Ataxia with Oculomotor Apraxia type 1 (PhD thesis)*. Queensland University of Technology, Brisbane, 2008.

[42] Janelle L Harris, Burkhard Jakob, Gisela Taucher-Scholz, Grigory L Dianov, Olivier J Becherel, and Martin F Lavin. Aprataxin, poly-ADP ribose polymerase 1 (PARP-1) and apurinic endonuclease 1 (APE1) function together to protect the genome against oxidative damage. *Human molecular genetics*, 18(21):4102–4117, November 2009.

[43] Joshua Hersheson, Andrea Haworth, and Henry Houlden. The inherited ataxias: Genetic heterogeneity, mutation databases, and future directions in research and clinical diagnostics. *Human Mutation*, 33(9):1324–1332, July 2012.

[44] Makito Hirano, Tomohisa Nishiwaki, Shingo Kariya, Yoshiko Furiya, Makoto Kawahara, and Satoshi Ueno. Novel splice variants increase molecular diversity of aprataxin, the gene responsible for early-onset ataxia with ocular motor apraxia and hypoalbuminemia. *Neuroscience Letters*, 366(2):120–125, August 2004.

[45] Makito Hirano, Hirohide Asai, Takao Kiriyama, Yoshiko Furiya, Takaaki Iwamoto, Tomohisa Nishiwaki, Aya Yamamoto, Toshio Mori, and Satoshi Ueno. Short half-lives of ataxia-associated aprataxin proteins in neuronal cells. *Neuroscience Letters*, 419(2): 184–187, May 2007.

[46] Makito Hirano, Aya Yamamoto, Toshio Mori, Li Lan, Taka-Aki Iwamoto, Masashi Aoki, Keiji Shimada, Yoshiko Furiya, Shingo Kariya, Hirohide Asai, Akira Yasui, Tomohisa Nishiwaki, Kyoko Imoto, Nobuhiko Kobayashi, Takao Kiriyama, Tetsuya Nagata,

Noboru Konishi, Yasuto Itoyama, and Satoshi Ueno. DNA single-strand break repair is impaired in aprataxin-related ataxia. *Annals of Neurology*, 61(2):162–174, February 2007.

[47] M Itaya and K Kondo. Molecular cloning of a ribonuclease H (RNase HI) gene from an extreme thermophile Thermus thermophilus HB8: a thermostable RNase H can functionally replace the Escherichia coli enzyme in vivo. *Nucleic Acids Research*, 19(16): 4443–4449, August 1991.

[48] Masao Ito. Historical review of the significance of the cerebellum and the role of Purkinje cells in motor learning. *Annals of the New York Academy of Sciences*, 978:273–288, December 2002.

[49] J E Karp, T I Vener, M Raponi, E K Ritchie, B D Smith, S D Gore, L E Morris, E J Feldman, J M Greer, S Malek, H E Carraway, V Ironside, S Galkin, M J Levis, M A McDevitt, G R Roboz, C D Gocke, C Derecho, J Palma, Y Wang, S H Kaufmann, J J Wright, and E Garret-Mayer. Multi-institutional phase 2 clinical and pharmacogenomic trial of tipifarnib plus etoposide for elderly adults with newly diagnosed acute myelogenous leukemia. *Blood*, 119(1):55–63, January 2012.

[50] Amir Kheradmand and David S Zee. Cerebellum and ocular motor control. *Frontiers in neurology*, 2:53, 2011.

[51] Amanda W Kijas, Janelle L Harris, Jonathan M Harris, and Martin F Lavin. Aprataxin forms a discrete branch in the HIT (histidine triad) superfamily of proteins with both DNA/RNA binding and nucleotide hydrolase activities. *The Journal of biological chemistry*, 281(20):13939–13948, May 2006.

[52] N Kim, S y N Huang, J S Williams, Y C Li, A B Clark, J E Cho, T A Kunkel, Y Pommier, and S Jinks-Robertson. Mutagenic Processing of Ribonucleotides in DNA by Yeast Topoisomerase I. *Science*, 332(6037):1561–1564, June 2011.

[53] Raga Krishnakumar and W Lee Kraus. The PARP Side of the Nucleus: Molecular Actions, Physiological Outcomes, and Clinical Targets. *Molecular cell*, 39(1):8–24, July 2010.

[54] M Labuda, D Labuda, C Miranda, J Poirier, B W Soong, N E Barucha, and M Pandolfo. Unique origin and specific ethnic distribution of the Friedreich ataxia GAA expansion. *Neurology*, 54(12):2322–2324, June 2000.

[55] Adam Lange, Stefan Becker, Karsten Seidel, Karin Giller, Olaf Pongs, and Marc Baldus. A Concept for Rapid Protein-Structure Determination by Solid-State NMR Spectroscopy. *Angewandte Chemie International Edition*, 44(14):2089–2092, March 2005.

[56] Oliver F Lange, Paolo Rossi, Nikolaos G Sgourakis, Yifan Song, Hsiau-Wei Lee, James M Aramini, Asli Ertekin, Rong Xiao, Thomas B Acton, Gaetano T Montelione, and David Baker. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proceedings of the National Academy of Sciences of the United States of America*, 109(27):10873–10878, July 2012.

[57] Marie-France Langelier, Kristin M Servent, Elizabeth E Rogers, and John M Pascal. A third zinc-binding domain of human poly(ADP-ribose) polymerase-1 coordinates DNA-dependent enzyme activation. *The Journal of biological chemistry*, 283(7):4105–4114, February 2008.

[58] Martin F Lavin, Nuri Gueven, and Padraic Grattan-Smith. Defective responses to DNA single- and double-strand breaks in spinocerebellar ataxia. *DNA repair*, 7(7):1061–1076, July 2008.

[59] Isabelle Le Ber, Maria-Céu Moreira, Sophie Rivaud-Péchoux, Céline Chamayou, François Ochsner, Thierry Kuntzer, Marc Tardieu, Gérard Saïd, Marie-Odile Habert, Geneviève Demarquay, Christian Tannier, Jean-Marie Beis, Alexis Brice, Michel Koenig, and Alexandra Dürr. Cerebellar ataxia with oculomotor apraxia type 1: clinical and genetic studies. *Brain*, 126(Pt 12):2761–2772, December 2003.

[60] Isabelle Le Ber, Alexis Brice, and Alexandra Dürr. New autosomal recessive cerebellar ataxias with oculomotor apraxia. *Current neurology and neuroscience reports*, 5(5):411–417, September 2005.

[61] Jonathan G Lees, Andrew J Miles, Frank Wien, and B A Wallace. A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics*, 22(16):1955–1962, August 2006.

[62] Hudan Liu, Nancy D Rodgers, Xinfu Jiao, and Megerditch Kiledjian. The scavenger mRNA decapping enzyme DcpS is a member of the HIT family of pyrophosphatases. *The EMBO Journal*, 21(17):4699–4708, 2002.

[63] S W Liu, V Rajagopal, S S Patel, and M Kiledjian. Mechanistic and Kinetic Analysis of the DcpS Scavenger Decapping Enzyme. *Journal of Biological Chemistry*, 283(24): 16427–16436, March 2008.

[64] Antoine Loquet, Nikolaos G Sgourakis, Rashmi Gupta, Karin Giller, Dietmar Riedel, Christian Goosmann, Christian Griesinger, Michael Kolbe, David Baker, Stefan Becker, and Adam Lange. Atomic model of the type III secretion system needle. *Nature*, May 2012.

[65] R Martin. Preparation of protein nanocrystals and their characterization by solid state NMR. *Journal of Magnetic Resonance*, 165(1):162–174, November 2003.

[66] Heinrich J Matthaei, Oliver W Jones, Robert G Martin, and Marshall W Nirenberg. Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences of the United States of America*, 48:666–677, April 1962.

[67] A M Maxam and W Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, February 1977.

[68] Peter J McKinnon. DNA repair deficiency and neurological disease. *Nature Reviews Neuroscience*, 10(2):100–112, January 2009.

[69] K McQuillen, R B Roberts, and R J Britten. SYNTHESIS OF NASCENT PROTEIN BY RIBOSOMES IN ESCHERICHIA COLI. *Proceedings of the National Academy of Sciences of the United States of America*, 45(9):1437–1447, September 1959.

[70] John F Milligan, Duncan R Groebe, Gary W Witherell, and Olke C Uhlenbeck. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Research*, 15(21):8783–8798, 1987.

[71] P Mosesso, M Piane, F Palitti, G Pepe, S Penna, and L Chessa. The novel human gene aprataxin is directly involved in DNA single-strand-break repair. *CMLS Cellular and Molecular Life Sciences*, 62(4):485–491, February 2005.

[72] Thomas P Naidich, Henri M Duvernoy, Bradley N Delman, Gregory A Sorensen, Spyros S Kollias, and Mark E Haacke. Duvernoy's atlas of the human brain stem and cerebellum. *SpringerWienNewYork*, 2009.

[73] H Nakamura, Y Oda, S Iwai, H Inoue, E Ohtsuka, S Kanaya, S Kimura, C Katsuda, K Katayanagi, and K Morikawa. How does RNase H recognize a DNA.RNA hybrid? *Proceedings of the National Academy of Sciences of the United States of America*, 88 (24):11535–11539, December 1991.

[74] D A Nielsen and D J Shapiro. Preparation of capped RNA transcripts using T7 RNA polymerase. *Nucleic Acids Research*, 14(14):5936, July 1986.

[75] Osamu Onodera. Spinocerebellar ataxia with ocular motor apraxia and DNA repair. *Neuropathology*, 26(4):361–367, August 2006.

[76] Francesc Palau and Carmen Espinós. Autosomal recessive cerebellar ataxias. *Orphanet journal of rare diseases*, 1:47, 2006.

[77] J Pauli, M. Baldus, B van Rossum, H de Groot, and H Oschkinat. Backbone and side-chain 13C and 15N signal assignments of the alpha-spectrin SH3 domain by magic angle spinning solid-state NMR at 17.6 Tesla. *Chembiochem : a European journal of chemical biology*, 2(4):272–281, April 2001.

[78] J A Pleiss, M L Derrick, and O C Uhlenbeck. T7 RNA polymerase produces 5' end heterogeneity during in vitro transcription from certain templates. *RNA (New York, NY)*, 4(10):1313–1317, October 1998.

[79] Ulrich Rass, Ivan Ahel, and Stephen C West. Defective DNA repair and neurodegenerative disease. *Cell*, 130(6):991–1004, September 2007.

[80] Ulrich Rass, Ivan Ahel, and Stephen C West. Molecular mechanism of DNA deadenylation by the neurological disease protein aprataxin. *The Journal of biological chemistry*, 283(49):33994–34001, December 2008.

[81] Srebrenka Robic, James M Berger, and Susan Marqusee. Contributions of folding cores to the thermostabilities of two ribonucleases H. *Protein Science*, 11(2):381–389, April 2009.

[82] F Rottman, A J Shatkin, and R P Perry. Sequences containing methylated nucleotides at the 5' termini of messenger RNAs: possible implications for processing. *Cell*, 3(3): 197–199, November 1974.

[83] Amy M Ruschak, Algirdas Velyvis, and Lewis E Kay. A simple strategy for 13C,1H labeling at the Ile-γ2 methyl position in highly deuterated proteins. *Journal of Biomolecular NMR*, 48(3):129–135, October 2010.

[84] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977.

[85] Yasuteru Sano, Hidetoshi Date, Shuichi Igarashi, Osamu Onodera, Mutsuo Oyake, Toshiaki Takahashi, Shintaro Hayashi, Mitsunori Morimatsu, Hitoshi Takahashi, Takao Makifuchi, Nobuyoshi Fukuhara, and Shoji Tsuji. Aprataxin, the causative protein for EAOH is a nuclear protein with a potential role as a DNA repair protein. *Annals of Neurology*, 55(2):241–249, February 2004.

[86] Heather F Seidle, Pawel Bieganowski, and Charles Brenner. Disease-associated mutations inactivate AMP-lysine hydrolase activity of Aprataxin. *The Journal of biological chemistry*, 280(22):20927–20931, June 2005.

[87] J Sekiguchi and S Shuman. Site-specific ribonuclease activity of eukaryotic DNA topoisomerase I. *Molecular cell*, 1(1):89–97, December 1997.

[88] B Séraphin. The HIT protein family: a new family of proteins present in prokaryotes, yeast and mammals. *DNA sequence : the journal of DNA sequencing and mapping*, 3 (3):177–179, 1992.

[89] T P Shields, E Mollova, L Ste Marie, M R Hansen, and A Pardi. High-performance liquid chromatography purification of homogenous-length RNA produced by trans cleavage with a hammerhead ribozyme. *RNA (New York, NY)*, 5(9):1259–1267, September 1999.

[90] Nathalie Sibille, Xavier Hanoulle, Fanny Bonachera, Dries Verdegem, Isabelle Landrieu, Jean-Michel Wieruszeski, and Guy Lippens. Selective backbone labelling of ILV methyl labelled proteins. *Journal of Biomolecular NMR*, 43(4):219–227, March 2009.

[91] Dea Slade, Mark S Dunstan, Eva Barkauskaite, Westom Ria, Pierre Lafite, Neil Dixon, Marijan Ahel, David Leys, and Ivan Ahel. The structure and catalytic mechanism of a poly(ADP-ribose) glycohydrolase. *Nature*, 477(7366):616–620, October 2011.

[92] Justin L Sparks, Hyongi Chon, Susana M Cerritelli, Thomas A Kunkel, Erik Johansson, Robert J Crouch, and Peter M Burgers. RNase H2-Initiated Ribonucleotide Excision Repair. *Molecular cell*, 47(6):980–986, September 2012.

[93] N Sreerama and R W Woody. Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set. *Analytical Biochemistry*, 2000.

[94] M Stoldt, J Wöhnert, O Ohlenschläger, M Görlach, and L R Brown. The NMR structure of the 5S rRNA E-domain-protein L25 complex shows preformed and induced recognition. *The EMBO Journal*, 18(22):6508–6521, November 1999.

[95] John A Stringer, Charles E Bronnimann, Charles G Mullen, Donghua H Zhou, Sara A Stellfox, Ying Li, Evan H Williams, and Chad M Rienstra. Reduction of RF-induced sample heating with a scroll coil resonator structure for solid-state NMR probes. *Journal of Magnetic Resonance*, 173(1):40–48, March 2005.

[96] Masashiro Sugawara, Chizu Wada, Satoshi Okawa, Michio Kobayashi, Masato Sageshima, Tsuyoshi Imota, and Itaru Toyoshima. Purkinje Cell Loss in the Cerebellar Flocculus in Patients with Ataxia with Ocular Motor Apraxia Type 1/Early-Onset Ataxia with Ocular Motor Apraxia and Hypoalbuminemia. *European Neurology*, 59(1-2):18–23, 2008.

[97] Peter Sykora, Deborah L Croteau, Vilhelm A Bohr, and David M Wilson. Aprataxin localizes to mitochondria and preserves mitochondrial function. *Proceedings of the National Academy of Sciences of the United States of America*, 108(18):7437–7442, May 2011.

[98] Masayoshi Tada, Akio Yokoseki, Tatsuya Sato, Takao Makifuchi, and Osamu Onodera. Early-onset ataxia with ocular motor apraxia and hypoalbuminemia/ataxia with oculomotor apraxia 1. *Advances in experimental medicine and biology*, 685:21–33, 2010.

[99] Christopher Torchia, Yuko Takagi, and C Kiong Ho. Archaeal RNA ligase is a homodimeric protein that catalyzes intramolecular ligation of single-stranded RNA and DNA. *Nucleic Acids Research*, 36(19):6218–6227, November 2008.

[100] Percy Tumbale, C Denise Appel, Rolf Kraehenbuehl, Patrick D Robertson, Jessica S Williams, Joe Krahn, Ivan Ahel, and R Scott Williams. Structure of an aprataxin–DNA complex with insights into AOA1 neurodegenerative disease. *Nature structural & molecular biology*, October 2011.

[101] E L Ulrich, H Akutsu, J F Doreleijers, Y Harano, Y E Ioannidis, J Lin, M Livny, S Mading, D Maziuk, Z Miller, E Nakatani, C F Schulte, D E Tolmie, R Kent Wen-

ger, H Yao, and J L Markley. BioMagResBank. *Nucleic Acids Research*, 36(Database): D402–D408, December 2007.

[102] S Vermeer, B P C van de Warrenburg, M A A P Willemsen, M Cluitmans, H Scheffer, B P Kremer, and N V A M Knoers. Autosomal recessive cerebellar ataxias: the current state of affairs. *Journal of medical genetics*, 48(10):651–659, September 2011.

[103] B A Wallace, Lee Whitmore, and Robert W Janes. The Protein Circular Dichroism Data Bank (PCDDB): A bioinformatics and spectroscopic resource. *Proteins: Structure, Function, and Bioinformatics*, 62(1):1–3, October 2005.

[104] J D Watson. Involvement of RNA in the synthesis of proteins. *Science*, 1963.

[105] J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[106] O Yuce and S C West. Senataxin, Defective in the Neurodegenerative Disorder Ataxia with Oculomotor Apraxia 2, Lies at the Interface of Transcription and the DNA Damage Response. *Molecular and Cellular Biology*, 33(2):406–417, December 2012.

[107] Stephan G Zech, A Joshua Wand, and Ann E McDermott. Protein Structure Determination by High-Resolution Solid-State NMR Spectroscopy: Application to Microcrystalline Ubiquitin. *J. Am. Chem. Soc*, 127(24):8618–8626, June 2005.

[108] Roland Zell, Karim Sidigi, Enrico Bucci, Axel Stelzner, and Matthias Görlach. Determinants of the recognition of enteroviral cloverleaf RNA by coxsackievirus B3 proteinase 3C. *RNA (New York, NY)*, 8(2):188–201, February 2002.

[109] Alexander M Zhelkovsky and Larry A McReynolds. Simple and efficient synthesis of 5' pre-adenylated DNA using thermostable RNA ligase. *Nucleic Acids Research*, June 2011.

7 Danksagung

Diese Arbeit wäre ohne die Unterstützung zahlreicher Freunde und Kollegen nicht möglich gewesen. Vor allem möchte ich mich bei meinem Mentor Dr. Matthias Görlach für die intensive Betreuung insbesondere in den letzten kritischen Minuten bedanken. Dank gebührt auch Dr. Than und Prof. Wang für die Geduld während des Besprechend meiner Ergebnisse, sowie der Leibniz Graduate School on Ageing and Age-Related Diseases für die finanzielle Unterstützung während meiner Promotionsphase. Ein besonderes Dankeschön (!) geht an Sabine Häfner, die mich in schwierigen Situationen mit Schokolade, Kaffee und vor allem guter Laune unterstützte. Dank Sabines wachender Hand, ist unser Labor weiterhin RNAse-frei. Ohne die Unterstützung von Thomas Seiboth wäre diese Arbeit auch nicht zu dem geworden was sie ist. Also danke Thomas! Natürlich haben auch die übrigen Mitglieder unserer NMR-Arbeitsgruppe einen großen Dank verdient. Auch wenn sich die Welten einer Biochemikers und eines “NMR-Spektroskopikers” doch ab und zu stark unterscheiden, hat mich Ramadurai Ramachandran stets motiviert und mit neuen publikationsfähigen Ideen unterstützt. Danke auch an Christoph Wiedemann für seine oft sehr realistischen Ansichten. Die Arbeitsgruppe von Prof. Wang hat mich freundlich in Ihren “DNA-Reparatur”-Kreis aufgenommen und mit konstruktiver Kritik auch zur strategischen Ausrichtung meiner Arbeit beigetragen. Ich möchte Stephen West und Prof. Keith Caldecott für die zur Verfügung gestellten Zelllinien und Plasmide danken. Privat haben mich einige, für mich sehr wichtige Menschen begleitet, und auch wenn sich einige der Wege leider inzwischen getrennt haben, gebührt Ihnen dennoch ein großes Dankeschön. Insbesondere möchte ich Robert Wilke und Sascha Schneider für Ihre kontinuierliche Unterstüzung und Motivation danken. Meiner Familie, die mir beigebracht hat, Probleme und Konflikte stets freundlich zu begegnen, möchte ich sagen: Ja ich habe es tatsächlich geschafft meinen Zeitplan einzuhalten. Also danke für’s Daumen drücken.

8 Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst und keine anderen Hilfsmittel als angegeben verwendet habe. Insbesondere versichere ich, dass ich alle wörtlichen und sinngemäßen Übernahmen aus anderen Werken als solche kenntlich gemacht habe. Mir ist die geltende Promotionsordnung der Biol.-Pharmaz. Fakultät der Universität Jena bekannt. Bei der Auswahl und Auswertung des Materials, sowie bei der Herstellung des Manuskripts habe ich von Dr. Matthias Görlach (Mentor) Unterstützungsleistungen erhalten. Alle weiteren Personen, die mich bei experimentellen Arbeiten und/oder bei der Erstellung der Manuskripte unterstützt haben sind in der Autorenliste der Publikationen benannt. Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Die Hilfe eines Promotionsberaters wurde ebenfalls nicht in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorliegenden Arbeit stehen. Diese Dissertation wurde bisher, weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde oder Hochschule als Dissertation oder Prüfungsarbeit vorgelegt.

Jena, den 2. April 2014