Structural Studies on Inteins

Jesper S. Øemig

Research Program in Structural Biology and Biophysics
Institute of Biotechnology
University of Helsinki

Division of Biochemistry
Department of Biosciences
Faculty of Biological and Environmental Sciences
University of Helsinki

And

National Doctoral Programme in Informational and Structural Biology
Åbo Akademi University

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki, for public examination in the auditorium 2041 of Biocenter 2, Viikinkaari 5, Helsinki,
on December 5th 2013, at 12 noon.

Helsinki 2013

Supervisor
Docent Hideo Iwai
Institute of Biotechnology
University of Helsinki, Finland

Thesis advisory committee
Professor Kari Keinänen
Faculty of Biological and Environmental Sciences, Department of Bioscience,
University of Helsinki, Finland

Professor Adrian Goldman
Faculty of Biological and Environmental Sciences, Department of Bioscience,
University of Helsinki, Finland &,
Department of Biological Sciences,
University of Leeds, UK

Reviewers
Professor Ilkka Kilpeläinen
Department of Chemistry
Faculty of Science
University of Helsinki, Finland

Professor Rikkert W. Wierenga
Department of Biochemistry
Faculty of Science
University of Oulu, Finland

Opponent
Professor Christian Griesinger
Max Planck Institute for Biophysical Chemistry
Göttingen, Germany

Custos
Professor Kari Keinänen
Faculty of Biological and Environmental Sciences, Department of Bioscience,
University of Helsinki, Finland

ISBN 978-952-10-9334-0 (Paperback)
ISBN 978-952-10-9335-7 (PDF, http://ethesis.helsinki.fi/)
ISSN 1799-7372
Helsinki University Print
Helsinki 2013

Somewhere, something incredible is waiting to be known.
- Carl Sagan

Contents

Abbreviations

BIL	Bacterial Intein-Like
ChyRIR1 intein	Ribonucleoside-diphosphate reductase 1 intein from *Carboxydothermus hydrogenoformans*
*Civ*RIR1 intein	Ribonucleoside-diphosphate reductase 1 intein from *Chilo iridescent* virus
*Cne*PRP8 intein	PRP8 intein from *Cryptococcus neoformans*
COSY	Correlation spectroscopy
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EPL	Expressed protein ligation
ESRF	European Synchrotron Radiation Facility
Extein	External protein
HINT	Hedgehog/INTin
HSQC	Heteronuclear Single Quantum Correlation
Intein	Internal protein
IPTG	Isopropyl-β-D-1-thiogalactopyranoside
*Mja*KlbA intein	KlbA intein from *Methanococcus jannaschii*
MS	Mass spectrometry
*Mtu*RecA intein	RecA intein from *Mycobacterium tuberculosis*
*Mxe*GyrA intein	DNA gyrase subunit A intein from *Mycobacterium xenopi*
NMR	Nuclear magnetic resonance
NOE	Nuclear overhauser effect
*Npu*DnaE intein	DnaE intein form *Nostoc punctiforme*
*Pab*PolII intein	DNA polymerase II intein from *Pyrococcus abyssi*
PDB	Protein data bank
*Pfu*RIR1-1	Ribonucleoside-diphosphate reductase 1-1 intein from *Pyrococcus furiosus*
*Pho*RadA intein	RadA intein from *Pyrococcus horikoshii*
*Pho*RadA$_{min}$ intein	Minimized RadA intein from *Pyrococcus horikoshii*
PI	Protein insert
POI	Protein of interest
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
*Sce*VMA intein	Vacuolar ATPase intein from *Saccharomyces cerevisiae*
*Ssp*DnaB intein	DnaB intein form *Synechocystis sp.* strain PCC6803
*Ssp*DnaE intein	DnaE intein from *Synechocystis sp.* strain PCC6803
*Tko*PolII intein	DNA polymerase II intein from *Thermococcus kodakaraensis*
TOCSY	Total correlation spectroscopy
TROSY	Transverse relaxation optimized spectroscopy

Amino acids

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Argenine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

List of Original Publications

This thesis is based on the following publications:

I Heinämäki*, K., **Oeemig*, J.S.,** Djupsjöbacka, J., and Iwaï, H. (2009) NMR resonance assignment of DnaE intein from Nostoc punctiforme. *Biomol. NMR Assign*. **3**, 41-43.

II **Oeemig*, J.S.,** Aranko*, A.S., Djupsjöbacka, J., Heinämäki, K., and Iwaï, H. (2009) NMR solution structure of DnaE intein from *Nostoc punctiforme*: Structural basis for the design of a new split intein suitable for site-specific chemical modification. *FEBS lett*. **583**, 1451-1456.

III Lyskowski*, A., **Oeemig*, J.S.,** Jaakkonen, A., Rommi, K., DiMaio, K., Zhou, D., Kajander, T., Baker, D., Wlodawer, A., Goldman, A., and Iwaï, H. (2011) Cloning, expression, purification, crystallization and preliminary X-ray diffraction data of *Pyrococcus horikoshii* RadA intein. *Acta Cryst*. **F67**, 623–626.

IV **Oeemig, J.S.**, Zhou, D., Kajander, T., Wlodawer, A., and Iwaï, H. (2012) NMR and crystal structures of the *Pyrococcus horikoshii* RadA intein guide a strategy for engineering a highly efficient and promiscuous intein. *J. Mol. Biol.* **421**, 85-99.

* Equal contribution

Author contribution

I JSO completed the resonance assignment, and contributed in manuscript completion.

II JSO performed the NMR structure determination and NMR data analysis, and participated in manuscript preparation.

III JSO did the protein crystallization and participated in writing the manuscript.

IV JSO participated in planning experiments, conducted most experiments, performed data analysis, and participated in writing the manuscript.

The publications are referred to in the text by their roman numerals.

Additional related publication

Aranko, A.S.*, **Oeemig, J.S.***, and Iwaï, H. (2013a) Structural basis for Protein Trans-Splicing by a Bacterial Intein-Like domain: Protein Ligation without nucleophilic side-chains. *FEBS J*. **280**, 3256-3269.

Aranko, A.S., **Oeemig, J.S.**, Kajander, T. and Iwaï, H. (2013b) Increased protein diversity by intein-mediated protein alternative splicing. *Nat. Chem. Biol.* **9**, 616-622.

Abstract

Inteins are proteins that can mediate protein-splicing reactions. The reaction is a posttranslational modification where the intein sequence is excised from a protein precursor and the flanking sequences, termed exteins, are ligated together. Intein mediated protein splicing has many possible biotechnological applications, but the structural features of inteins influencing the protein splicing reaction remains unclear. A better understanding how inteins can be engineered for applications and how the extein sequences influence the protein splicing efficiency is needed to fully exploit the full potential of intein applications.

The aim of this thesis was to determine the structures of selected inteins in order to gain further insight in the structural-functional relationships of inteins. The goal was to engineer new inteins that can be used for biotechnological applications based on structural characterization. It was further aimed to investigate how the extein sequences influence protein splicing.

Structure determination of the *Npu*DnaE intein was performed using solution NMR spectroscopy and X-ray crystallography. The NMR structure and nuclear spin relaxation rates were found to be a powerful tool to identify novel split sites in inteins. A novel split *Npu*DnaE intein was engineered to have a split C-intein part consisting of six amino acid residues. The novel split intein has a big potential for site-specific modification of proteins.

The structure of *Pho*RadA intein was determined using NMR spectroscopy and X-ray crystallography. Both structures indicated *Pho*RadA intein residues 120-133 were flexible. Based on the structural data a minimized *Pho*RadA intein (*Pho*RadA$_{min}$ intein) was engineered without loss of protein splicing function by deletion of residues 121-130.

The crystal structures of *Npu*DnaE intein and *Pho*RadA$_{min}$ intein were determined as protein splicing precursors to 1.72 Å and 1.58 Å resolution, respectively. The structure of *Npu*DnaE intein resembled an “open” conformation that would require a conformational change for the protein splicing reaction to occur. By contrast, the structure of *Pho*RadA$_{min}$ intein resembled a previously unseen “closed” conformation. The structure conformation had the N- and C-protein splicing junctions in close proximity and a clear interaction between the -1 extein residue and the intein was observed.

The effect of the -1 extein residue on the protein splicing efficiency was systematically analysed by substitution of the -1 residue to all 20 natural amino acids. It was discovered that the negatively charged residues Asp and Glu caused reduced protein splicing efficiency. Based on the *Pho*RadA$_{min}$ intein structure an E71T mutation was introduced in *Pho*RadA intein that removed unfavourable electrostatic interactions. The E71T mutation of *Pho*RadA intein became protein splicing efficient (>90%) with a Glu-1 residue. Thus, *Pho*RadA$_{min}$ intein crystal structure provides the structural basis for *Pho*RadA intein -1 extein residue protein splicing dependency. This observation may have further application for engineering of protein splicing efficient inteins.

1 Introduction

Myoglobin and haemoglobin were the first protein crystal structures that were reported more than 50 years ago (Kendrew *et al.*, 1958; Perutz et al., 1960). The structures were determined at low resolution using X-ray crystallography but the method has since been developed for determination of high though-put high-resolution structures. Structural biology has progressed with other techniques such as NMR spectroscopy and cryo-electron microscopy to determine macromolecular structures at high resolution (Williamson *et al.*, 1985; Zhang *et al.*, 2008). Solved structures are deposited in the protein data bank where they are publically available. Currently more than 90,000 depositions are available and more than 1000 unique protein folds have been defined (Berman *et al.*, 2000; Murzin *et al.*, 1995).

Structural biology has been used to explain a function and mechanism of many macromolecules. Knowing the structure of macromolecules has made it possible to understand biological systems such as the DNA double helix (Watson and Crick, 1953), ribosomes (Ban *et al.*, 2000), and structural knowledge has been used for protein engineering with therapeutic purpose (Jones *et al.*, 1986; Brange *et al.*, 1988).

Inteins were reported more than two decades ago as an unusual "spacer protein" (Hirata *et al.*, 1990; Kane *et al.*, 1990). Inteins perform a posttranslational modification where an intein is excised from a precursor sequence. Upon removal of the intein sequence the flanking sequences are ligated together and form a mature protein. The underlying mechanism of intein-mediated protein splicing has been greatly investigated to understand the underlying function (Tori *et al.*, 2010). Despite much effort in understanding inteins mechanism and function no biological function of this elusive molecule has been proven. Inteins are merely considered to be parasitic genes that provide no benefit to the host organism. The number of discovered inteins and determined intein structures has been increasing since the first intein structure was solved in 1997 (Perler, 2002; Duan *et al.*, 1997). Intein structures have been useful for understanding the mechanism of intein reactions and for guiding their engineering for a large number of biotechnological applications.

This work focuses on the structural-functional features of selected inteins. Selection of an intein for structure determination is based upon literature data that have shown its efficient protein splicing ability (Iwai *et al.*, 2006; Ellilä *et al.*, 2011). In this work NMR spectroscopy and X-ray crystallography are used as supporting structural determination methods that provide further insight in how inteins can be engineered. At first the structure of DnaE intein from *Nostoc punctiforme* (*Npu*DnaE intein) was investigated using NMR spectroscopy. The NMR solution structure combined with nuclear spin relaxation studies showed to be a powerful tool in analysis of intein dynamic properties. This combination showed to be important for identification of novel split sites in inteins.

Further, the work focused on structural properties of RadA intein from *Pyrococcus horikoshii* (*Pho*RadA intein). *Pho*RadA intein structure was determined using NMR and X-ray crystallography, which was used as a guidance to further engineer a minimized intein (*Pho*RadA$_{min}$ intein). The structure of the *Pho*RadA$_{min}$ intein was then determined

as a protein splicing precursor using X-ray crystallography. The structure revealed an unseen conformation of a protein splicing precursor that provided insight in how the -1 N-extein residue influences the protein splicing efficiency. A structure of *Npu*DnaE intein was determined as a protein splicing precursor using X-ray crystallography. The conformation of *Npu*DnaE intein resembled an open conformation that has been seen in other inteins (Aranko *et al.*, 2013b).

A systematic analysis was performed on the protein splicing dependency of *Pho*RadA intein -1 residue position. The protein splicing efficiency of *Pho*RadA intein was reduced when the -1 residue was β-branched or a negatively charged amino acid type. The structure of *Pho*RadA$_{min}$ intein provided the structural basis for *Pho*RadA intein -1 extein residue dependency that guided in rational protein engineering for protein splicing of non-native residues at the -1 position.

2 Literature Review

2.1 Inteins and Exteins

Inteins are intervening protein sequences that are excised from a protein precursor. Upon excising of the intein sequence the flanking sequences are ligated together and forms the mature polypeptide chain. The process is an autocatalytic post-translational modification termed protein splicing and it is catalysed by the intein sequence. The term intein is derived form *in*ternal pro*tein* and the flanking sequences that are ligated together are termed exteins derived from *ex*ternal pro*tein* (Perler *et al.*, 1994). The extein sequences upstream and downstream of the intein sequence are referred to as N-extein and C-extein, respectively (see Figure 1).

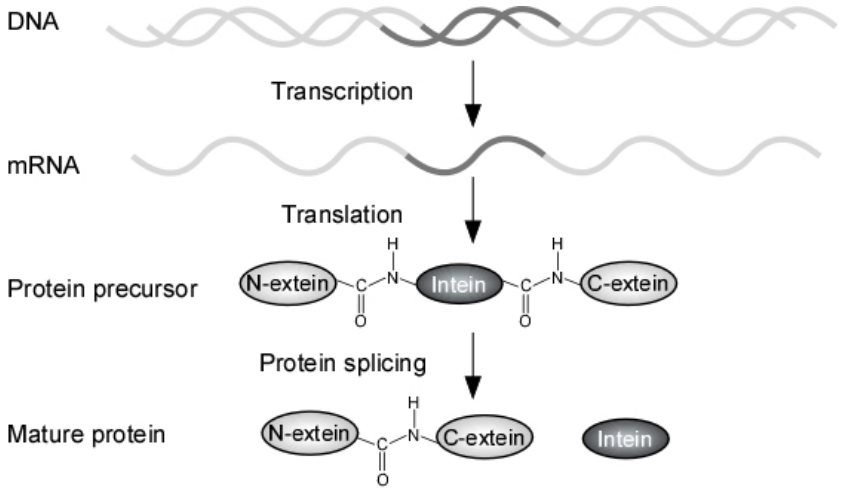

Figure 1 *Inteins are are transcribed from a DNA sequence and are translated from the mature mRNA into a polypeptide chain. After translation the intein sequence is excised from the polypeptide chain and the flanking extein sequences are ligated together forming a mature protein. Inteins and exteins are shown in dark and light grey, respectively.*

Inteins where initially reported in 1990 by two individual reports (Hirata *et al.*, 1990; Kane *et al.*, 1990). Both studies showed that the gene TFP1 from *Saccharomyces cerevisiae* produced a 69 kDa mature protein and a 50 kDa protein which was referred to as "spacer" protein. Kane *et al.* (1990) showed that the intein most likely was still present in the mature mRNA and that after the translation of the mRNA the mature protein product would be formed. The final proof of this process being a posttranslational modification came with the development of the first *in vitro* protein splicing system (Xu *et al.*, 1993). The system showed that the intein sequence is first excised from the gene after translation by a posttranslational protein splicing process. Intein mediated protein splicing is a self-catalytic and intra-molecular process that does not require any external energy source such as ATP or other macromolecules (Kawasaki *et al.*, 1997). However, the

understanding of reaction has progressed and a recent study has shown that besides an intra-molecular reaction inteins can also perform an intermolecular reactions (Aranko *et al.*, 2013b).

Since the discovery of the first intein in *S. cerevisiae* inteins have been also discovered in the three domains of life (Eukaryote, Bacteria, and Archaea), and in viruses, and bacterial phages (Pietrokovski, 1998a; Lazarevic *et al.*, 1998). But, inteins have only been identified in unicellular organisms. In order to keep track of the number of intein sequences discovered a database has been created (http://tools.neb.com/inbase/) in which more than 550 intein sequences from eucarya, eubacteria, and archaea are listed (Perler, 2002). The majority of the sequences are putative intein sequences that have been identified based on sequence homology and only few inteins have been experimentally tested. Identified inteins are named after organism of origin and the gene hosting the intein. Accordingly, intein naming contains a three letter genus/species designation followed by a host gene designation e.g. DnaE intein from *Nostoc punctiforme* is named *Npu*DnaE intein and GyrA intein from *Mycobacterium xenopi* is named *Mxe*GyrA intein. Discovered inteins often contain an endonuclease domain and studies that focused on the endonuclease domain have used a different naming system. Inteins containing endonuclease domain have been named with the prefix PI (for Protein Insert) followed by an organism name and a number related to the order of discovery, e.g. PI-*Sce*I (Perler *et al.*, 1994; Belfort and Roberts, 1997).

2.1.1 Intein Sequences, Naming, and Sequence Motifs

Intein sequences range from 134 residues for the smallest *cis* splicing intein to more than 1000 residues (Evans *et al.*, 1999b; Perler, 2002). Many discovered intein sequences are bifunctional because they contain an endonuclease domain inserted in the intein sequence. Homing endonucleases domains are DNA cutting proteins that have specific DNA recognition sequence from 12-40 base pairs (Roberts and Macelis, 1997) (the function of the endonuclease domain will be discussed in Section 2.1.2). The endonuclease domain is typically inserted more than 100 residues form the intein N-terminus and more than 30 residues from the intein C-terminus so the N- and C-terminus form the protein splicing domain (see Figure 2) (Duan *et al.*, 1997). Inteins lacking an endonuclease domain are in some cases referred to as mini-inteins and they are up to ~200-300 amino acid residues. It was earlier believed that the endonuclease domain was needed for the protein splicing. The first described natural mini-intein was *Mxe*GyrA intein (Telenti *et al.*, 1997) and it was shown that *Mxe*GyrA intein still was able to perform protein splicing despite lacking an endonuclease domain. Once when natural and engineered protein splicing mini-inteins were discovered it became evident that the endonuclase domain is dispensable for protein splicing. The two domains are able to fold independently and the endonuclease is not needed for stabilizing the intein domain (Chong and Xu, 1997; Derbyshire *et al.*, 1997; Shingledecker *et al.*, 1998; Wu *et al.*, 1998a).

Figure 2 *Protein splicing precursor where the intein (dark grey) contains an endonuclease domain (white). Numbering of intein, N- and C-exteins residues are indicated. The location of the conserved sequence blocks A, B, C-E, F, and G is indicated. Highly conserved residues are shown below in capital letters and less common residues are shown in lower case letters. X denotes a non-specified amino acid (Pietrokovski, 1994; Perler, 2002).*

Mini-inteins have been found as split inteins. Split inteins are inteins that are formed by association of two intein parts expressed by two separate genes (see Figure 3) (Wu *et al.*, 1998b; Gorbalenya, 1998). The split intein parts are unfolded separately but upon association of the split intein parts they fold together and form a functional intein (Zhen *et al.*, 2012). Split inteins can perform protein splicing in *trans*. Natural split inteins are split at the site where endonuclease domains normally are found inserted in intein sequences. The split parts of inteins are named after the N- and C-extein part they are ligated to and termed N- and C-inteins, respectively (see Figure 3). Split inteins are naturally occurring (Gorbalenya, 1998; Wu *et al.*, 1998b) but engineered split inteins were reported even before natural occurring ones were discovered (Southworth *et al.*, 1998; Mills *et al.*, 1998; Shingledecker *et al.*, 1998). Several natural split inteins have been discovered in cyanobacteria located in the DnaE gene (Caspi *et al.*, 2003) but only few natural split inteins have been described experimentally (Wu *et al.*, 1998b; Iwai *et al.*, 2006; Dassa *et al.*, 2007; Shah *et al.*, 2012). The discovery of split inteins and the ability to perform protein *trans* splicing has opened a big potential in biotechnological applications (see Section 2.4).

Intein residues are numbered in sequential order starting from 1 until the last residue of the intein. The N-extein residues upstream of the intein are named -1, -2, -3, where the -1 residue is the last N-extein residue. The C-extein residues downstream of the intein are named +1, +2, +3 where the +1 residue is the first C-extein residue (see Figure 2). Inteins have low sequence homology and only few residues are highly conserved among inteins. The most conserved residues are the first residue of the intein, a Cys or Ser, and the last residue, which is an Asn. Additionally protein splicing is dependent on the first residue of the C-extein (the +1 residue) being a Cys, Ser, or Thr (Xu and Perler, 1996). However, these residues are not fully conserved among inteins and exceptions are found. Intein sequences starting with an Ala or Pro and sequences ending with an Asp or Gln have been described (Southworth *et al.*, 2000; Tori *et al.*, 2010; Amitai *et al.*, 2004; Mills *et al.*, 2004). Because of the difference in the conserved residues it is believed that inteins have different mechanisms for protein splicing and consequently have been divided into three different intein classes (see Section 2.2). Inteins share some sequence similarity and

sequence motifs, which contain conserved residues. The sequence motif are referred to as block A, B, F, and G (or N1, N3, C2, and C1) (Pietrokovski, 1994; Perler *et al.*, 1997). Two additional sequence motifs have been described and are referred to as block N2 and N4 (Pietrokovski, 1998b). The sequence block A and B are located near the N-terminus of the intein whereas the blocks F and G are located near the C-terminus (see Figure 2). The sequence blocks C-E are located in the endonuclease domain and involved in endonuclease activity.

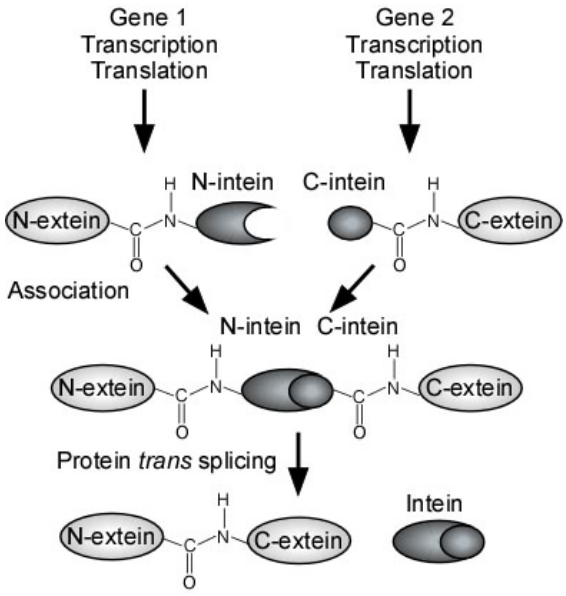

Figure 3 *Split inteins are expressed by two separate genes and upon association of the split intein parts protein* trans *splicing can occur. Inteins and exteins are shown in dark and light grey, respectively.*

Intein sequences have been found located in many different genes and inteins are inserted in the active site region of the host proteins (Perler, 2002). Intein insertion sites are not all conserved between different species and in some genes different insertion sites are found. Inteins inserted into the same gene at the same site are considered allelic and often have a higher sequence homology (Perler *et al.*, 1997). Genes have been found containing more than intein sequence. When more intein sequences are located in a single gene the intein sequencess are numbered in sequential order. The Vent DNA polymerase from *Thermococcus litoralis*, the DNA polymerase from *Thermococcus aggregans,* and the Ribonucleoside Triphosphate Reductase from *Trichodesmium erythraeum* contains two, three, and four inteins, respectively (Perler *et al.*, 1992; Hodges *et al.*, 1992; Niehaus *et al.*, 1997; Liu et al., 2003). The intein sequences can then make up a larger portion of the translated gene than the mature product e.g. ribonucleoside triphosphate reductase from *Trichodesmium erythraeum* contains four inteins and is translated to a 2,240 amino acid protein precursor that after protein splicing forms a 769 amino acid mature protein (Liu *et al.*, 2003).

2.1.2 Biological Function of Inteins

Intein are believed to be ancient because they are found in all three domains of life across diverse species (Shub and Goodrich-Blair, 1992). Inteins are related to Hedgehog proteins because they share a conserved protein fold with the Hedgehog protein C-terminal domain (see section 2.3). However, Hedgehog proteins are found in metazoa and are involved in cell signalling which have not been observed for inteins (Lee *et al.*, 1992). The Hedgehog protein C-terminal domain and inteins have a similar initial reaction step. The difference in the reaction is that the Hedgehog C-terminal protein coordinates a cholesterol modification of the N-terminal domain and a cleavage reaction between the N- and C-terminal domains (Porter *et al.*, 1995). The conserved fold and similar reaction mechanism indicates inteins and Hedgehog proteins likely originates from a common ancestor. The Hedgehog protein has appeared before metazoa evolved and at some time in evolution Hedgehog proteins have obtained a different function than inteins (Pietrokovski, 2001).

The biological function of inteins is not clear and inteins are considered "parasitic" element that provides no beneficial role for the host. Inteins are genetic elements found inserted in conserved functional regions of a host protein (Swithers *et al.*, 2009) e.g. ligand binding region, active site, etc. Insertions of inteins in conserved functional regions of a protein are believed to be important for the survival of inteins because there is no evidence inteins provide any benefit for the host organism, thus they are considered expendable. As long as a protein contains an intein sequence it is considered inactive (Paulus, 2003; Liu and Yang, 2004; Huet *et al.*, 2006) where the accurate removal of an intein sequence by protein splicing generates a functional protein. A genetic removal of the intein from a gene is difficult because the removal has to be very precise (Derbyshire and Belfort, 1998). Inaccurate removal could cause amino acid insertion/deletion or change in reading frame of the host gene, which leads to loss of function and could be lethal for the organism (Pietrokovski, 2001).

As mentioned many inteins contain an endonuclease domain that has been suggested to be important in spreading inteins by homing events by horizontal transfer. The result of homing is a replication of a parasitic element to an allele without the genetic element. Endonuclease domains are DNA cutting enzymes that have a recognition sequence between 12-40 base pairs which typically occur only once in the genome (Roberts and Macelis, 1997). A cut in the genome is lethal for the organism and a simple re-ligation would not be efficient because the endonuclease domain would cut the genome again. In order to repair the cut in the genome the gene of the endonuclease domain is used as template and thus is incorporated into the genome. Consequently, the homing endonuclease domain recognition sequence is removed. Insertion of an endonuclease domain into a protein active site likely disrupts the gene function, but insertion of an endonuclease domain in an intein would preserve the host gene function (Pietrokovski, 2001). There is evidence of this kind of phenomenon occurring in *S. cerevisiae* VMA gene during cell division (Gimble and Thorner, 1992). There are indications of horizontal intein transfer in GyrA gene between related mycobacterial species based upon the fact that the codon usage and dG/dC content of the intein differs from the host organism (Fsihi *et al.*, 1996). Similarly, it has been proposed with DnaB intein from two more distant species

where the inteins share much higher sequence identity than the extein sequence which indicates a more recent transfer (Liu and Hu, 1997). How the inteins and endonuclease domains came to co-exist is unknown but association or loss of intein and homing endonuclease domain could have happened several times (Barzel *et al.*, 2011). The endonuclease domain likely has been acquired by the intein and propagated the transfer of inteins but the endonuclease domain has lost its function by mutation or deleted over time (Telenti *et al.*, 1997; Liu, 2000).

There are few proposed biological functions of inteins. A significance of intein existence has been tested with recA gene that is involved in DNA recombinase. A recA deletion mutant of *Mycobacterium smegmatis* was tested for intein beneficial function. The recA gene was replaced with an allelic gene from *Mycobacterium tuberculosis* with and without its intein (Papavinasasundaram *et al.*, 1998; Frischkorn *et al.*, 1998). The *M. smegmatis* recA deletion mutant was sensitive to DNA damage but supplementing the deletion variant with *M. tuberculosis* recA gene the function of the wild type recA deletion was recovered. There was no difference if recA contained an intein or the intein was absent. Hence, there is no indication that the intein provides any advantage for the organism.

An obvious function would be intein meditated gene regulation since inteins are found inserted in active sites of genes. Thus, the host gene is inactive until the intein sequence has been removed (Paulus, 2003). Controlled protein splicing mechanism could regulate gene activation. In the presence of natural split inteins the function of inteins might be more obvious. Only if the split intein parts are present simultaneously and the split inteins associate *trans* protein splicing generate a functional protein (Wu *et al.*, 1998a).

In addition to *trans* protein splicing could regulate protein function redox potential could regulate protein splicing in some genes (Callahan *et al.,* 2011). A CxxC motif was identified in an intein and was introduced in a model system at the N-protein splicing junction sequence. The CxxC motif was introduced with the two Cys being the -3 and the 1 residue, respectively. The cysteines worked as a redox trap of the intein by disulfide bridge formation between Cys-3 and Cys1 (CxxC motif). Cys1 is essential for the protein splicing mechanism and protein splicing could only proceed when the disulfide bond was reduced (Callahan *et al.,* 2011). Thus, it indicates that inteins could produce a precursor protein that would be unable to perform protein splicing under oxidizing conditions. This could generate a high concentration of inactive precursor proteins and if cell conditions change, protein splicing could rapidly generate new functional proteins.

One more function of intein could be to increase protein diversity. As mentioned in Section 2.1 it has recently been shown that inteins can perform intermolecular reaction (Aranko *et al.*, 2013b). Inteins could increase the molecular diversity by an intein-mediated protein alternative splicing reaction where inteins containing different extein sequences could cross-react. Protein diversity can then be enhanced from a single genetic background. However, proof of this principle in a biological relevant context has not been provided.

2.2 Protein Splicing Mechanism

Protein splicing mechanism of inteins has been well studied and a generally accepted opinion on the process has been accepted (see Figure 4). However, with increasing knowledge on inteins a need to divide the mechanism into different classes has been recognised (Southworth *et al.*, 2000; Tori *et al.*, 2010). Studies have shown that inteins perform protein splicing by different mechanisms and consequently inteins have been divided into three different classes based upon sequence homology and what reaction mechanism they are likely to follow.

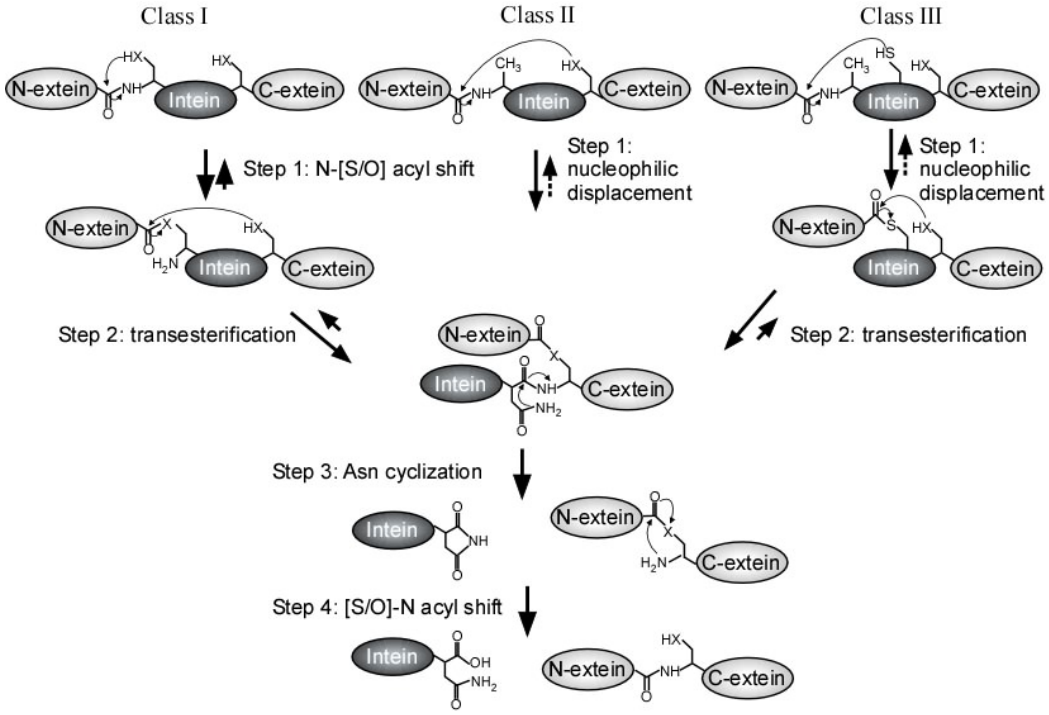

Figure 4 *Schematic representation of intein class I, II, and III protein splicing mechanism (Modified from Tori* et al.*, 2010).*

The majority of identified inteins are classified as class I (Perler, 2002) and this is the best-studied intein class. In the following sections the variation of the different classes will be described.

2.2.1 Standard Protein Splicing Mechanism

In this section the general protein splicing mechanism for class I inteins is described (see Figure 4). The protein splicing reaction consists of four steps where the first step is a S(O)-N acyl shift of the first intein residue (Cys or Ser) on the scissile bond of the last residue of the N-extein. This generates a linear (thio)ester intermediate (Xu and Perler, 1996; Shao *et al.*, 1996). The (thio)ester bond was identified in a model system of an

intein containing no C-extein and the system was treated with a weak nucleophilic reagents (hydroxylamine) (Xu and Perler, 1996). It was assumed that hydroxylamine would not break a peptide bond but would lead to increased cleavage reaction of a (thio)ester. The experiments showed that the presence of hydroxylamine increased the cleavage reaction at the N-terminal protein splicing junction. In step 2 a transesterification occurs where the (thio)ester formed in step 1 is cleaved by the +1 C-extein residue (Cys, Ser, or Thr). This forms a branched (thio)ester intermediate. The branched intermediate was initially identified in a model system where a slow migrating protein band was identified on a SDS-PAGE gel. N-terminal sequencing of the protein band showed the presence of two free N-termini (Xu *et al.*, 1993). Step 3 involves a cyclization of the last intein residue that most often is an Asn. This generates an aminosuccinimide and the extein part is released from the intein (Shao *et al.*, 1995; Xu *et al.*, 1994). The formation of the aminosuccinimide was shown by cyanogen bromide (CNBr) treatment of an intein mutant. Cyanogen bromide cleaved a part of a model peptide at the intein C-terminus and the cleaved part was identified with mass spectrometry (MS). The MS data indicated that aminosuccinimide formation was likely a mechanism for the peptide cleavage. In step 4 a rapid acyl shift generates an amide bond in the extein and the mature protein is formed. The aminosuccinimide in the intein is hydrolyzed and the Asn is re-generated (Xu and Perler, 1996; Shao and Paulus, 1997).

2.2.2 Variations of the Standard Protein Splicing Mechanism

In addition to the described standard protein splicing mechanism of class I inteins explained above different variations exist. Inteins contain conserved sequence motif (Block A, B, F, and G), but variations in the sequence motif have led to classification of inteins into three different classes (I, II, and III). Inteins which are lacking Cys or Ser as the first residue belong to class II and III and are described in following sections.

Class II Inteins

Inteins identified as class II contain an Ala as the first intein residue. Mutation of class I intein where Cys1 has been replaced by Ala prevents the formation of the initial (thio)ester bond, and thus prevents occurrence of the protein splicing reaction. Initially, it was thought that inteins encompassing an Ala1 residue were inactive intein elements (Gorbalenya, 1998). A study on KlbA intein from *Methanococcus jannaschii* (*Mja*KlbA intein) containing an initial Ala showed that the intein was proficient in protein splicing (Southworth *et al.*, 2000). These inteins have adopted a different mechanism to circumvent the first step of the standard protein splicing mechanism where it has been suggested that the +1 residue makes a direct attack on the -1 scissile bond and forms a branched intermediate (see Figure 4). The structure of *Mja*KlbA intein has been determined but the alternative mechanism is still unclear from the structure (Johnson *et al.*, 2007a; Johnson *et al.*, 2007b).

Class III Inteins

DnaB intein from *Mycobacteriophage Bethlehem* (*MP*-DnaB intein) was the first reported intein as class III (Tori *et al.*, 2010). This intein class lacks the initial Cys or Ser, and the protein splicing mechanism is suggested to involve a Trp-Cys-Thr (WCT) triplet. The Trp is located two residues downstream of the conserved block B His. Further the intein lacks the highly conserved block B Thr (see Figure 2). The Thr of the WCT triplet is the third last residue of the intein located in a central ß-hairpin and it was suggested that the Trp and Thr have a structural role and are not directly involved in the reaction mechanism. The Cys of the WCT triplet is located in block F where inteins often accommodate an Asp. The role of the Cys is postulated to be important in N- and C-terminal cleavage where the Cys initiates the first step of the protein splicing by attacking the N-terminal splice junction and forms a block F branched intermediate (see Figure 4). In an intein mutant the branched intermediate has been identified where N-terminal sequencing gave two amino acid signals (Brace *et al.*, 2010). By a transesterification the N-extein is then transferred to the +1 residue forming a branched intermediate as described in other inteins (Tori *et al.*, 2010; Brace *et al.*, 2010).

2.2.3 Cleavage Reaction during Protein Splicing

The protein splicing mechanism described above has been investigated using model systems. Initially it was believed that the native extein sequences were needed for the protein splicing reaction. But it was early realised that protein splicing could occur with foreign extein sequences and in a foreign organism (Cooper *et al.*, 1993). However, when inteins are used in foreign extein contexts protein splicing may become ineffective and side products formation can occur, which is referred to as cleavage reaction (see Figure 5).

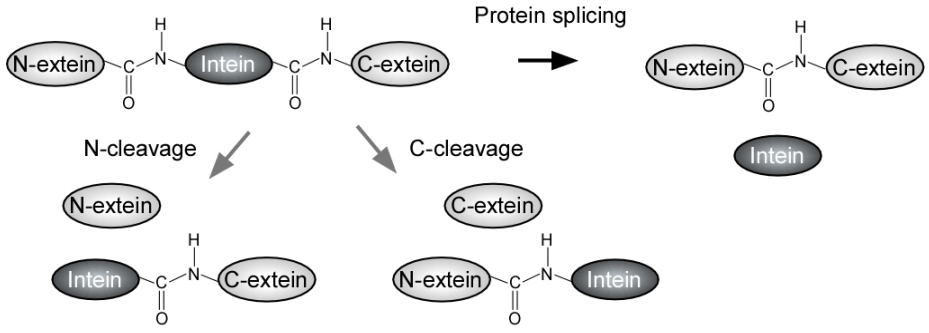

Figure 5 *Protein splicing and side product formation by intein reaction. The black arrow indicates the protein splicing product formation. The grey arrows indicate bi-products formation by N- or C-cleavage reaction.*

The cleavage reaction can occur at the peptide bond between the -1 or the +1 residue and the intein and is referred to as N- and C-cleavage, respectively. The cleavage reaction at the N-terminal splice junction can be caused by hydrolyses of the (thio)ester intermediate

that is formed in step 1 of the protein splicing mechanism or at the (thio)ester that occurs in the branched intermediate (Noren *et al.*, 2000). Cleavage reaction at the C-terminal protein splicing junction can occur by Asn cyclization happening before the transesterification has occurred in step 2. (Xu and Perler, 1996; Chong *et al.*, 1998; Mathys *et al.*, 1999; Shemella *et al.*, 2007)

2.2.4 Conserved Residues in the Protein Splicing Mechanism

In an attempt to understand the intein protein splicing mechanism many mutational characterizations have been performed. Initial studies identified the N- and C-protein splicing junction thereby identifying the sequence boundary between intein and extein (Davis *et al.*, 1992; Hodges *et al.*, 1992). Further studies have focused on how inteins work and which amino acid residues are involved in the protein splicing. The following sections provides a description of mutation studies performed on conserved residues located in the conserved sequence motifs A, B, F, and G (see Figure 6).

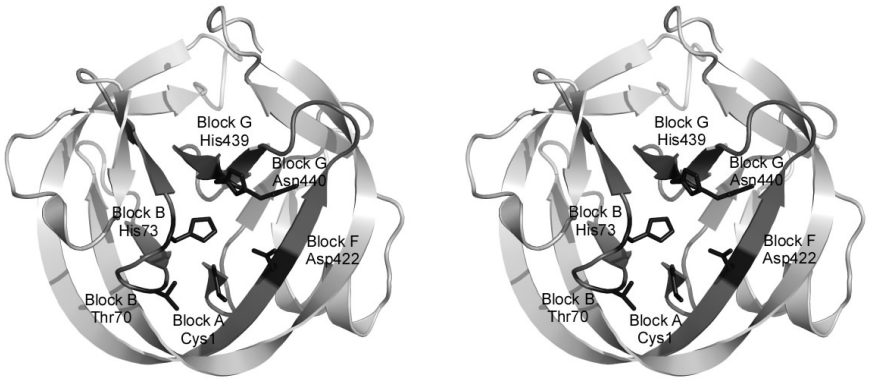

Figure 6 *A stereo view of post protein splicing structure of* Mtu*RecA intein (PDB: 2IN0, Van Roey et al., 2007). The conserved sequence blocks A, B, F, and G is highlighted in dark grey. The highly conserved residues block A Cys1, block B Thr70 and His73, block F Asp422, and block G Asn440 and penultimate His439 are shown in sticks and coloured black.*

A general conclusion on which residues catalyse the protein splicing reaction is not possible and the literature results indicate that residues are most likely intein specific in functionality. This indicates that inteins have different mechanisms and timing of the protein splicing reaction steps.

Block A Residue Mutation

In class I inteins the first residue of the block A sequence motif is directly involved in the protein splicing. In most inteins this is a Cys or Ser. Thr has not been found as the first intein residue yet but mutation of the first intein residue to Thr protein splicing can still occur for some inteins at a reduced efficiency (Hodges *et al.*, 1992). A Cys1 or Ser1 mutation to the opposite is not always compatible with the inteins (Chong *et al.*, 1996).

Mutation of residue one in inteins, often to an Ala, prevents protein splicing reaction from occurring but in some inteins cleavage at the C-junction occurs (Van Roay et al., 2007; Ding *et al.*, 2003; Du et al., 2011a). A C1A mutation can also blocks any reaction from occurring in other inteins (Pearl *et al.*, 2007b).

Obvious exceptions to this are inteins that belong to class II and III, which still are able to do protein splicing since they naturally exists without Cys or Ser as the first residue. Protein splicing was still proficient when Ala1 or Pro1 was mutated to Cys in *Mja*KlbA intein and *MP*-DnaB intein, respectively (Southworth *et al.*, 2000; Tori *et al.*, 2010).

Block B Residue Mutation

Block B TxxH residues are some of the most conserved intein residues that are not directly involved in the protein splicing reaction. The structures of inteins have shown that the residues are in close proximity to the first intein residue and can form hydrogen bond with the -1 N-scissile bond (Klabunde *et al.*, 1998; Poland *et al.*, 2000). Both block B Thr and His are believed to be important for the initial step in protein the protein splicing mechanism. The block B Thr is not as well studied as block B His but mutation of the Thr in some inteins have shown to completely blocks the protein splicing reaction (Pearl *et al.*, 2007b; O'Brien *et al.*, 2010).

The block B His is believed to be important for activation of the first step in the protein splicing mechanism and a mutation of the block B His abolished the protein splicing reaction and only N- or C-cleavage is observed in the mutants (Ghosh *et al.*, 2001; Du *et al.*, 2009; Johnson *et al.*, 2007a). Thus, it indicates block B His is essential for the protein splicing product formation. However, based upon sequence prediction CDC21-1 intein form *Thermococcus kodakaraensis* does not contain a block B His but it is still fully proficient in protein splicing (Tori *et al.*, 2012). An alternative mechanism was suggested that involved a non-conserved Lys residue. Similar was observed when block B His in PRP8 intein from *Cryptococcus neoformans* (*Cne*PRP8 intein) was mutated where minute amount of protein splicing seen (Pearl *et al.*, 2007b). It has been indicated that block B His is important in the initial reaction step by destabilising of the -1 scissile peptide bond (Binschik and Mootz, 2013). *Ssp*DnaB intein was able to perform protein splicing in a block B His to Ala mutant when the Cys1 was N-methylated. The N-methylation of the peptide bond destabilises the peptide bond. The block B His has been further characterised in RecA intein from *Mycobacterium tuberculosis* (*Mtu*RecA intein) where the pK_a was determined to 7.3 in a protein construct with a N-extein and C1A mutation (Du *et al.*, 2009). The pK_a of the same residue was determined to be below 3.5 in a post protein splicing construct with the wild type Cys1. It was suggested that His could act as base to initiate the reaction and by the lower pK_a act as acid that helps break the -1 scissile bond.

Block F Residue Mutation

A conserved Asp is commonly found in block F at position four, but other amino acids are also found at the same position (see Figure 2) (Perler, 2002). The block F Asp has been

suggested to have a dual role in catalysing and coordinating the protein splicing (van Roey *et al.*, 2007). Mutation of the block F Asp in *Mtu*RecA intein and *Npu*DnaE intein predominantly led to C-cleavage reaction (van Roey *et al.*, 2007; Ramirez *et al.*, 2013), while the same mutation in *Cne*PRP8 intein made the intein inactive and no cleavage product was observed (Pearl *et al.*, 2007b). The structure of *Mtu*RecA intein has been determined by solution NMR and the block F Asp (D422) pK_a was determined to ~6 in a post protein splicing construct and the same pK_a value was observed in a construct bearing a N-extein. The determined pK_a value is evaluated compared to typical Asp pK_a value that is around 4.1 (Du *et al.*, 2008; Du *et al.*, 2011b; Berg *et al.*, 2002). However, it must be noted that the studies are done on inteins without any C-extein. Above it was mentioned that mutation of block F Asp in *Mtu*RecA intein affects the protein splicing and mainly C-cleavage occur (van Roey *et al.*, 2007). Therefore, the presence of a C-extein must influence block F Asp and the block F Asp is likely to be involved in controlling Asn cyclization in *Mtu*RecA intein (van Roey *et al.*, 2007). Additionally, when no C-extein is present a free carboxyl-terminal is located in the central cavity changing surroundings of the active site.

Block G Residue Mutation

Block G sequence motif is the last part of the intein sequence (see Figure 2). The last residue of inteins is important for breaking the bond between the intein and C-extein and this residue is most commonly an Asn (step 3 in the protein splicing mechanism). Mutation of Asn to Ala produced mainly intact precursor protein in some inteins (Chong *et al.*, 1996; Pearl et al., 2007b). However, in other inteins the same mutation lead to N-cleavage (van Roey et al., 2007; Johnson *et al.*, 2007a).

DNA polymerase II intein from *Pyrococcus abyssi* (*Pab*PolII intein), which is ending on an atypical Gln, is able to facilitate protein splicing *in vitro* (Mills *et al.*, 2004). Mutating Gln to Asn protein splicing was facilitated at a slightly increased rate with improved efficiency. The NMR structure of *Pab*PolII intein was later determined but the structure did not give any insight into why Gln is functional as last intein residue (Du *et al.*, 2011a). Ribonucleoside-diphosphate reductase 1 intein from *Chilo iridescent* virus (*Civ*RIR1 intein) is also ending on an atypical Gln and has been shown to facilitate protein splicing *in vivo*. Mutating last Gln to Asn still facilitated protein splicing but with reduced efficiency as opposite with *Pab*PolII intein (Amitai *et al.*, 2004). Additionally, it has been discovered that some inteins have an Asp as the last intein residue. Ribonucleoside-diphosphate reductase 1 intein from *Carboxydothermus hydrogenoformans* (*Chy*RIR1) ends on an Asp and facilitated protein splicing *in vivo* but the protein splicing efficiency was poor and mainly yielded N-cleavage. Mutating the Asp to Asn, Gln, Glu, or Ala completely abolished the protein splicing product of *Chy*RIR1 intein (Amitai *et al.*, 2004). Interestingly, *Chy*RIR1 intein with an Asp to Ala mutation was able to undergo C-cleavage indicating the C-cleavage was independent of on Asp or Asn cyclization.

The +1 residue of block G is important for ligation of the N- and C-extein parts and it is the only extein residue involved the protein splicing reaction. The pK_a of *Mtu*RecA intein Cys+1 was estimated to 5.8 (Shingledecker *et al.*, 2000) and this low pK_a value

(normally at ~8.0) makes the Cys+1 residue more reactive. This is most likely reason for disulfide bond formation between Cys1 and Cys+1 under slight oxidizing conditions (Mills et al., 1998; Lew *et al.*, 1998; Chen *et al.*, 2012).

A penultimate block G His has been identified in many inteins. Clp protease intein from *Chlamydomonas eugametos* that contains a penultimate Gly was inactive but mutating the penultimate Gly to His protein splicing ability was recovered (Wang and Liu, 1997). Additionally, mutation of *Sce*VMA intein penultimate His decreased the protein splicing efficiency (Chong *et al.*, 1998). Based on this it was believed that the penultimate block G His catalyses Asn cyclization (step 3 in the protein splicing mechanism). However, later the opposite have been shown when the penultimate His has been mutated to Ala protein splicing was still proficient (Kerrigan *et al.*, 2009). The fact that many discovered inteins do not contain a penultimate His and are still proficient in protein splicing indicates that the presence of penultimate His is not essential and the role is not of the penultimate His is unclear (Wu et al., 1998a; Chen *et al.*, 2000; Perler, 2002).

It is obvious that mutations of the same conserved residues in different inteins have a different effect on protein splicing. Thus, no general conclusion can be made about how the intein protein splicing is catalysed but rather the studies provide an insight in a specific intein mechanism. Because of this diversity in the intein mechanism it is difficult to understand how the protein splicing reaction influenced when inteins are applied in foreign context. In next section work performed to understand how the extein sequence can influence protein splicing is described.

2.2.5 Extein Sequences and Effect on Protein Splicing Efficiency

The protein splicing mechanism of inteins is generally well established as discussed previously. It is clear the intein sequence catalyses and coordinates the protein splicing. The effect of extein sequence on protein splicing is not fully understood. The flanking extein residues -2, -1, +2, and +3, which are not directly involved in the reaction can influence the protein splicing (Chong *et al.*, 1998; Iwai *et al.*, 2006; Amitai *et al.*, 2009). The effect of foreign extein sequences at the protein splicing junction is observed by cleavage reaction or the presence of intact protein splicing precursors. Influence of extein residues on the protein splicing is unclear but understanding the effect of the flanking extein residues is particularly important when applying inteins in biotechnological applications (see Section 2.4). Few systematic studies on the extein residues influence on protein splicing have been performed and will be described further in this section but it still remains unclear how the extein influence the reaction.

*Mxe*GyrA intein -1 residue influence on protein splicing was investigated by systematically change the -1 residue to all 20 standard amino acids. *Mxe*GyrA intein showed highest protein splicing efficient with a Tyr or a Phe at the -1 residue position where Tyr is the wild type residue (Southworth *et al.*, 1999). Trp at the -1 residue position had more than 50% protein splicing efficiency while other residues showed significant lower protein splicing efficiency. In contrary *Sce*VMA intein seems less sensitive to mutation at residues at the -1 residue position. The wild type -1 residue is a Gly and by

changing the residue to all 20 common amino acids it was only with Val, Leu, Ile, Cys, Asn, or Pro that reduced protein splicing efficiency was observed (Chong *et al.*, 1998). *Cne*PRP8 intein is sensitive to mutation of the wild type Ala-1 residue to β-branched amino acid residue (Thr, Val, Ile) and hydrophobic residues (Leu) (Pearl *et al.*, 2007a). No further systematic analysis was performed on *Cne*PRP8 intein.

The natural split DnaE intein from *Synechocystis sp.* strain PCC6803 (*Ssp*DnaE intein) and *Nostoc punctiforme* (*Npu*DnaE intein) have been tested systematically for the +2 residue dependency (Iwai *et al.*, 2006). The C-intein part of *Ssp*DnaE intein was used to generate 20 mutants and *trans* protein splicing efficiency was tested with the N-intein part of *Ssp*DnaE intein and cross reactivity with *Npu*DnaE intein. The two inteins have 63% sequence identity but show a big difference in the protein splicing junction sequence tolerance. *Npu*DnaE intein has much higher sequence tolerance at the C-extein junction and more robust protein splicing activity (Iwai *et al.*, 2006; Zettler *et al.*, 2009; Shah *et al.*, 2012). The wild type +2 residue is a Tyr for both intein and *Ssp*DnaE intein only tolerated Tyr and Phe at the +2 residue position. Trp at the +2 residue position was tolerated but protein-splicing efficiency was reduced 50%. *Npu*DnaE intein has high protein splicing efficiency, which upon mutation of the +2 residue position was only reduced by a Pro, Gly, or Gln, or changed residue (Lys, Arg, Asp, Glu).

Directed evolution selection of inteins is a different approach used to make inteins function in foreign extein context. Different approaches have been used to generate new functional inteins in foreign context. The split *Npu*/*Ssp*DnaE intein chimera was developed to perform traceless protein splicing when the C-extein residues CFN (+1 to +3) was changed to a SGV sequence (Lockless and Muir, 2009). The split intein was inserted into aminoglycoside transferase that is used for kanamycin resistance. Through several round of selection a highly functional intein with a SGV C-extein had evolved and was isolated. The final isolated intein contained four mutations. None of the mutations were located in the vicinity of the activity site but all were scattered over the entire structure of *Ssp*DnaE intein (Sun *et al.*, 2005). A similar approach has been used to make direct developed *Ssp*DnaB intein by error prone PCR. Amplified intein was inserted into aminoglycoside transferase and a library of random intein mutants was created (Appleby-Tagoe *et al.*, 2011). Selected mutants were continuously evolved by several rounds of PCR. As in previously described study a similar result was obtained. The isolated inteins contained mutations scattered all over the structure and none of the mutations were at the vicinity of the active site (Ding *et al.*, 2003). A similar approach has been later used to generate a selection system of *Npu*DnaE intein that was inserted into aminoglycoside transferase where only the -3, -2, -1, +2, and +3 residue were selected by random mutation. It was shown that *Npu*DnaE intein has a wide sequence tolerance at the N-extein protein splicing junction while there was a high preference toward Trp and at the +2 residue position (Cheriyan *et al.*, 2013). The methods showed *Npu*DnaE intein have a wide variability at the N-extein sequence protein splicing junction.

From this section it is clear that different inteins have a different tolerance to amino acids located at the -2, -1, +2, and +3 residue position. Closely related inteins can have high difference in the protein splicing junction sequence tolerance and the difference is not

obvious from the primary structure. To understand the difference in intein mechanism the tertiary structure of inteins have been studied and is described in following section.

2.3 The Structure of HINT Domain

Structural studies on inteins have shown that the tertiary structure of inteins belong to a structural family referred to as Hedgehog/INTein (HINT) fold (see Figure 7) (Hall *et al.*, 1997; Dalgaard *et al*., 1997). As mentioned in Section 2.1.2 inteins and Hedgehog C-terminal domains are related by sharing some sequence similarity, a common fold, and similarity in function. Inteins and Hedgehog structures likely share the same mechanism for catalysing the first reaction step of a S(O)-N acyl shift generating a thioester, but the Hedgehog protein then coordinates an intermolecular reaction with cholesterol.

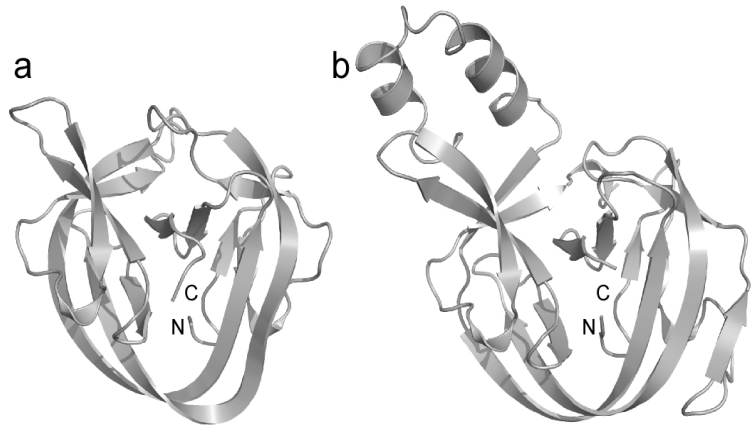

Figure 7 *The HINT protein fold. a) Crystal structure of* Drosophila melanogaster *Hedgehog C-terminal domain (Hall et al., 1997). b) Crystal structure of* Mxe*GyrA intein (Klabunde et al., 1998). N and C indicate the N- and C-termini, respectively.*

To understand the mechanism of inteins several structure studies have been performed using NMR spectroscopy and X-ray crystallography. Only one structure of a Hedgehog C-terminal domain has been determined which is derived from *Drosophila melanogaster* (Hall *et al.*, 1997). Intein structures have been determined deriving from several different organisms and different host genes. Including the intein structures determined in this thesis there have been determined 11 different inteins structures originating from different organism and host genes. An overview of intein structures available in the RCSB protein data bank is listed in Table 1.

Table 1. *Intein structures available in RCSB PDB. The structures are listed in order of publication date. Endonuclease containing inteins have the names for the endonuclease domain indicated in brackets. PSP: Protein Splicing Precursor; NEP: N-Extein Present; CEP: C-Extein Present; PPS: Post Protein Splicing domain.*

Intein	Res.	RCSB	Remarks	Reference
*Sce*VMA intein	2.4 Å	1VDE	PPS, contain endonuclease domain (PI-*Sce*I)	Duan *et al*., 1997
*Mxe*GyrA intein	2.2 Å	1AM2	NEP, C1S	Klabunde *et al.,* 1998
*Sce*VMA intein	2.0 Å	1DFA	PPS, Contain endonuclease domain (PI-SceI)	Hu *et al.*, 2000
*Sce*VMA intein	2.1 Å	1EF0	PSP, contain endonuclease domain (PI-SceI), C1A, N454A, C+1 is coordinated by Zn-ion	Poland *et al*., 2000
*Pfu*RIR1-1	2.1 Å	1DQ3	PPS, contain endonuclease domain (PI-*PfuI*)	Ichiyanagi *et al.*, 2000
*Sce*VMA intein	2.1 Å	1JVA	PSP (PI-SceI intein), C1S, H79N, N454S, C+1S	Mizutani *et al*., 2002
*Sce*VMA intein	1.35 Å	1GPP	NEP, deleted residue 184-410, R44S, V67M, I132V, L183G, N454 is deleted (C-terminal residue is H453) (PI-*Sce*I intein)	Werner *et al*., 2002
*Sce*VMA intein	3.5 Å	1LWS	PPS, contain endonuclease domain (PI-*Sce*I intein) (added Ca^{2+}), protein-DNA complex	Moure *et al.*, 2002
*Sce*VMA intein	3.2 Å	1LWT	PPS, contain endonuclease domain (PI-*Sce*I intein) (Ca^{2+} free), protein-DNA complex	Moure *et al.*, 2002
*Ssp*DnaB intein	2.0 Å	1MI8	PSP, C1A, N429A, deleted endonuclease domain	Ding *et al*., 2003
*Sce*VMA intein	2.9 Å	1UM2	PPS, contain endonuclease domain (PI-*Sce*I intein), C1S, H79N, C+1S	Mizutani *et al*., 2004
*Ssp*DnaE intein	1.95 Å	1ZDE	PSP, C1A, N158A, C+1 is coordinated by Zn-ion	Sun *et al*., 2005
*Ssp*DnaE intein	1.7 Å	1ZD7	PPS, single chain variant of natural split intein	Sun *et al*., 2005
*Tko*PolII intein	2.7 Å	2CW7	PPS, contain endonuclease domain (PI-*Pko*II)	Matsumura *et al.*, 2006
*Tko*PolII intein	2.5 Å	2CW8	PPS, contain endonuclease domain (PI-*Pko*II)	Matsumura *et al.*, 2006
*Mtu*RecA intein	1.8 Å	2IN9	PPS, removed endonuclease domain	Van Roey *et al*., 2007
*Mtu*RecA intein	1.7 Å	2IMZ	PPS, C1 is a cysteine-S-dioxide, V67L, removed endonuclease domain, cyclic Asn440	Van Roey *et al*., 2007
*Mtu*RecA intein	1.6 Å	2IN0	PPS, V67L, removed endonuclease domain	Van Roey *et al*., 2007
*Mtu*RecA intein	1.7 Å	2IN8	PPS, V67L, D422G, removed endonuclease domain	Van Roey *et al*., 2007
*Mja*KlbA intein	NMR	2JMZ	PSP, N168A, C+1S	Johnson *et al*., 2007a
*Mja*KlbA intein	NMR	2JNQ	Mean structure of 2JMZ	Johnson *et al*., 2007a
*Npu*DnaE intein	NMR	2KEQ	NEP, C1A, single chain variant of natural split intein	Study II
*Mtu*RecA intein	2.4 Å	3IGD	PPS, cyclic Asn, D24Y, removed endonuclease domain	Hiraga *et al*., 2009
*Mtu*RecA intein	1.9 Å	3IFJ	PPS, cyclic Asn, F421Y, removed endonuclease domain	Hiraga *et al*., 2009
*Ssp*DnaE intein	1.55 Å	3NZM	NEP, redox trapped intein (C-3 & C1 are SS-bound), single chain variant of natural split intein	Callahan *et al*., 2011
*Mtu*RecA intein	NMR	2L8L	PPS	Du *et al*., 2011b
*Pab*PolII intein	NMR	2LCJ	PPS	Du *et al*., 2011a
*Pho*RadA intein	NMR	2LQM	CEP, C1A, T+1A	Study IV
*Pho*RadA intein	1.75 Å	4E2T	CEP, C1A, T+1A	Study IV
*Pho*RadA intein	1.58 Å	4E2U	PSP, C1A, T+1A, deleted residue 121-130	Study IV
*Ssp*DnaE intein	1.8 Å	4GIG	NEP, T69A, single chain variant of natural split intein	Dearden *et al.*, 2013
*Npu*DnaE intein	1.72 Å	4KL5	PSP, C1A, C+1A, single chain variant of natural split intein	Aranko *et al.*, 2013b
*Npu*DnaE intein	2.2 Å	4KL6	PSP, dimeric form, C1A, V97N, N102G, C+1A, deleted residue T76 & 98-100, single chain variant of natural split intein	Aranko *et al.*, 2013b

Several of the determined intein structures contain an endonuclease domain and these studies typically have focused on the endonuclease activity, DNA binding, or recognition instead of intein function (Duan *et al.*, 1997; Ichiyanagi *et al.*, 2000; Hu *et al.*, 2000; Moure *et al.*, 2002; Werner *et al.*, 2002; Matsumura *et al.*, 2006). Other intein structures determined are natural mini-inteins (naturally does not contain endonuclease domain) or inteins whose endonuclease domain has been deleted before structural studies (e.g. *Mtu*RecA intein) (Hiraga *et al.*, 2005; van Roey *et al*., 2007). The structure of some inteins has been determined several times as different mutants or by different techniques. For example the structure of *Sce*VMA intein and *Mtu*RecA intein that have been determined seven and eight times, respectively (see Table 1). In most cases the intein structures are determined as post protein splicing domains (PPS, Table 1), which contain no extein residues or in some cases only a short N- or C-extein sequence (NEP/CEP, Table 1). Few structures have been determined as protein splicing precursor that contains both a N- and C-extein sequences and these will be discussed separately in Section 2.3.2 (PSP, Table 1). Most commonly intein structures have been determined by introducing mutations at conserved residues that was described in Section 2.2.4. Those mutations prevent any reaction (protein splicing or cleavage reaction) thus making the construct stable for a period of time needed to form protein crystals or perform NMR measurements.

2.3.1 Structure of Inteins

It has been important to know the structure of inteins to understand the mechanism and how inteins could be used for biotechnological applications (see Section 2.4). Since the structure of the first intein domain (Duan *et al.*, 1997) it became evident that intein protein splicing domain has a complex horseshoe disk-shaped like fold. The intein N- and C-termini are located in the central cavity bringing the protein splicing junction in close contact (see Figure 7b). Thus, the structures provide the basic for the excitation of intein sequence from the protein precursor, but no conclusion on intein catalyzation can be made based on their structure. Inteins are divided into different class and the majority of intein structures available are class I. To support the variation of different mechanism only one class II intein has been determined (*Mja*KlbA intein) (Johnson *et al.*, 2007a) and currently there is no structural data available for a class III intein.

In general, the structures of the different inteins are similar and the structural differences are seen in the length of loop regions. Some inteins are made up of more β-strands that are found inserted between sequences block A and B. *Sce*VMA intein structure has a different structural feature compared to other known intein structures. There is a sequence insertion after block B at residue 86-153 that is involved in DNA binding (He *et al.*, 1998; Grindl *et al.*, 1998). DNA binding domains are commonly not found as part of inteins. However, a DNA binding domain is also found in RIR1-1 intein from *Pyrococcus furiosus* (*Pfu*RIR1-1 intein) but it is inserted between the endonuclease domain and the C-terminus part of the intein (Ichiyanagi *et al.*, 2000).

The majority of intein structures have been resolved as post protein splicing domains where the extein sequence is completely missing. As a consequence the influence of the extein sequences on the protein splicing described in Section 2.2.5 is not understood.

2.3.2 Structure of Protein Splicing Precursor

It is evident that the extein sequence influence the protein reaction but it is unknown how this happens (see Section 2.2.5). Studies of post protein splicing intein provides no insight how the extein might be important. Few structures have been determined as protein splicing precursors where both the N- and C-extein sequences are present in the structure model. The structure of *Ssp*DnaE intein (Sun *et al.*, 2005), *Ssp*DnaB intein (Ding *et al*., 2003), *Sce*VMA intein (Poland *et al.*, 2000; Mizutani *et al.*, 2002), and *Mja*KlbA intein (Johnson *et al.*, 2007a) has been determined as a protein splicing precursors (see Table 1). For this purpose several mutations where introduced into the inteins in order to make the constructs inactive. These mutations have been introduced into the first intein residue and the first C-extein residue, but not always both residues were mutated at the same time. In all the structures the last intein residue (Asn) was mutated to Ala or Ser. In a structure of *Sce*VMA intein (PBD: 1JVA) the conserved block B His was mutated to Asn. Consequently, there is no structural information of an unmodified protein splicing precursor. The determined protein splicing precursor structures differ in their structural conformations. In the structures of *Ssp*DnaE intein (Sun *et al.*, 2005), *Ssp*DnaB intein (Ding *et al*., 2003), *Sce*VMA intein (Mizutani *et al.*, 2002), and *Mja*KlbA intein (Johnson et al., 2007a) there is a large cavity space between the N- and C-protein splicing junction. The distance between the -1 C' and the +1 S(O) is between 8.0-9.3 Å and a large conformation exchange is needed to bring the exteins close enough together for ligation of the extein sequences. This long distance is considered an "open" conformation that exists before the initial reaction step. The "open" state conformation has a long distance between the N- and C-protein splicing junction. For the second reaction step of the protein splicing reaction to occur the protein splicing junction need to be in close proximity. Thus, a larger conformational change is needed to bring the splicing junction close together. In one structure determined of *Sce*VMA intein the distance between -1 C' and the +1 O is 3.8 Å and could indicate a "closed" conformation that occurs during the protein splicing reaction (Mizutani *et al.*, 2002).

The crystal structure of *Mxe*GyrA intein contains one N-extein residue, Ala-1, that is in an unusual *cis* conformation (Klabunde *et al.*, 1998). The nature of this unusual scissile bond conformation would be to destabilize the peptide bond. A destabilised bond would be easier to break. However, *Mxe*GyrA intein containing an Ala-1 residue was reported to have protein splicing efficiency decreased to ~10%. Since cleavage efficiency by DTT was reduced an Ala-1 residue does not seem to be a driving force for destabilization of the peptide bond and initiating the first reaction step because (Southworth *et al.*, 1999). It is questionable if the observed conformation of Ala-1 is caused by the C1S mutation, absence of a C-extein in the structure, or the N-extein consists of only one residue which gives additional conformation freedom. An NMR spectroscopic study has provided

supporting results for the unusual scissile bond conformation in *Mxe*GyrA intein (Romanelli *et al.*, 2004). *Mxe*GyrA intein construct bearing an AAMRF N-extein with no C-extein and a N198A mutation had an unusual $^{1}J_{NC'}$ coupling of 12.3 ± 0.3 Hz for the -1 scissile amide bond. Other reported $^{1}J_{NC'}$ value about 13.5-17.5 Hz (Delaglio *et al.*, 1991; Juranić *et al.*, 1995; Liu and Prestegard, 2009). The difference in J-coupling is not big but the smaller J-coupling was interpreted as lengthening of the amide bond that would make the breaking of the peptide bond easier (Romanelli *et al.*, 2004).

Opposite to the previously described unusual *cis* conformation a normal *trans* conformation of the -1 scissile bond is seen in the other protein splicing precursors (Poland *et al.*, 2000; Mizutani *et al.*, 2002; Sun *et al.*, 2005; Ding *et al.*, 2003). However, the Gly-1 τ angle (N-Cα-C') of *Sce*VMA intein precursor was reported distorted from the ideal 110° angle (105°, molecule A) (Poland *et al.*, 2000), which could facilitate the initial step in the protein splicing mechanism. In the protein splicing precursor of *Ssp*DnaB intein Ser+1 τ angle was reported distorted (122°) and this could favour the Asn cyclication step (Ding *et al.*, 2003). The two different inteins have a different angle distorted. This could be interpreted by the inteins slight difference in protein splicing mechanism or the two inteins being in a different conformation.

As described before inteins are divided into three different classes based on their protein splicing mechanism, but it is possible that class I intein have slight different timing in the occurrence of the protein splicing. The structures of precursor proteins provide a further understanding of intein mechanism but most of the structures are in an "open" conformation that needs to change to a "closed" conformation. How the extein sequence can influence the protein splicing is still not evident from the precursor structures.

2.3.3 Bacterial Intein-Like Domain

A different protein class that belongs to the HINT fold family is the Bacterial Intein-Like (BIL) protein domains (Aranko *et al.*, 2013a). BIL domains are interesting because they are able to perform protein splicing despite they are not considered inteins. Identified BIL domains are between 130-156 amino residues and they share common sequence motifs with other HINT domains (Amitai *et al.*, 2003; Dassa *et al.*, 2004b). However, they contain additional unique sequence motifs, which are not found in other HINT protein. BIL domains differ from inteins by being inserted between domains instead of in conserved active sites regions and BIL domains do not contain any endonuclease domains. BIL domains differ from Hedgehog C-terminal domains by being flanked by different domains, they also contain different conserved sequence motifs, and they can occur in bacteria while Hedgehog protein are only found in metazoan (Amitai *et al.*, 2003).

BIL domains are divided into A, B, and C-type based upon sequence features (Amitai *et al.*, 2003; Dori-Bachash *et al.*, 2009). A striking difference between inteins and BIL domains is the amino acid residue at +1 position, where inteins have a conserved Cys/Ser/Thr residue and BIL domains have higher amino acid diversity at the same position (e.g. Ala, Val, Lys, Tyr, etc). In spite of the presence of unusual +1 residue BIL domains are able to perform protein splicing albeit at low efficiency (20-25%) (Dassa *et*

al., 2004a). However, *Magnetospirillum magnetotacticum* strain MS-1 type A BIL-domain was reported to only produce cleavage reaction but protein splicing was recovered at low efficiency by mutating the wild type Tyr+1 to Cys (Southworth *et al.*, 2004). Opposite, an A+1C mutation in BIL4 domain from *Clostridium thermocellum* made protein splicing very efficient as seen in other inteins (Aranko *et al.*, 2013a).

The function of BIL domains is unclear but experimental results have shown this type of HINT domains can perform protein splicing and cleavage reaction. This might be a mechanism to diversify gene population (Dassa *et al.*, 2004b; Dori-Bachash *et al.*, 2009).

2.4 Applications of Intein Technology

Intein technology has opened various possible biotechnological applications. The first important discovery was that inteins could perform protein splicing in non-native context, which allows modifications in foreign organisms and with foreign extein sequences. Secondly, the discovery of split inteins, either naturally split inteins or engineered split inteins, has been used for many protein *trans* splicing applications.

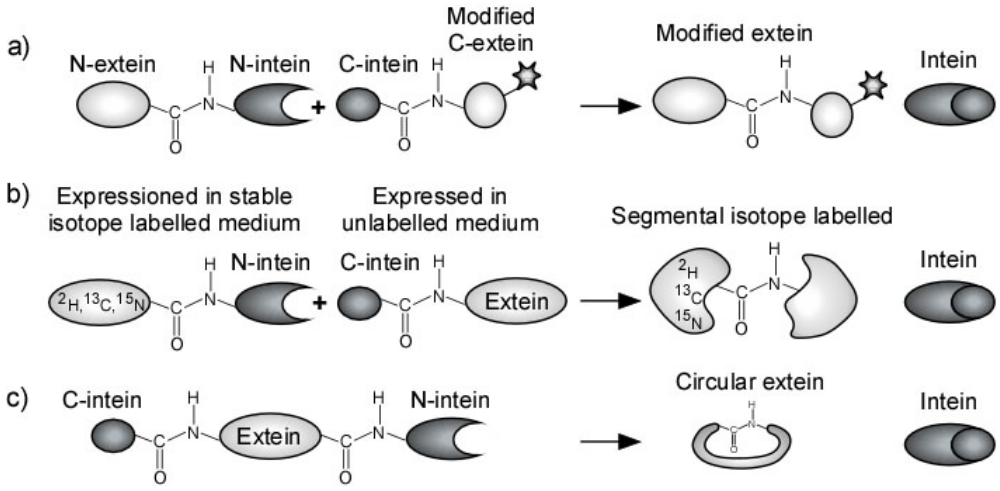

Figure 8 *Biotechnical applications of inteins. a) Site specific chemical modification. b) Segmental isotope labelling. c) Protein cyclization.*

Intein mutants have been used for protein purification because of their ability to induce cleavage of the peptide bond in controlled conditions. This has been achieved by mutating the last Asn of the intein to Ala, which prevents C-cleavage from occuring. The protein of interest (POI) is then expressed as fusion protein with a C-terminal modified intein and an additional protein purification tag. The modified intein is unable to perform the full protein splicing reaction but is able to form initial thio(ester) that can be cleaved at controlled condition by addition of a thiol like DTT. This is was to perform protease free cleavage. Similarly intein mutants have been modified for cleavage by Asn cyclization

where POI is fused at the intein C-terminus (Xu *et al.*, 2000). Cleavage reaction has also been controlled with pH sensitive inteins or temperature controlled conditions (Wood *et al.*, 1999; Southworth *et al.*, 1999; Mills *et al.*, 2006). This makes it possible to express target gene in *E. coli* at 37 °C and during protein purification steps change the temperature or pH to induce the cleavage.

2.4.1 Split Inteins and Site Specific Modification

Since the first intein structure was determined (Duan *et al.*, 1997) it was obvious that many inteins were composed of an intein protein splicing domain and an endonuclease domain. It is possible to reduce the intein size by genetically removing the endonuclease domain and, thus, creating a mini-intein (Derbyshire *et al.*, 1997; Wu *et al.*, 1998b; Hiraga *et al.*, 2005). Natural split inteins are split where the endonuclease domain is found inserted in inteins. Split inteins have been engineered by splitting the intein at the same site where the endonuclease domain is found inserted and where natural split inteins contain a split site. Actually was the first split intein genetically split in two parts before the first natural split intein was reported (Derbyshire *et al.*, 1997). Different split inteins have been engineered by changing position of the split site in primary structure. The split sites have been moved closer to either terminus to make the one split intein part smaller. An engineered *Ssp*DnaB mini-intein has been split with the N-intein part consisting of 11 amino acids though the protein splicing efficiency was lower than when the split site was located at the endonuclease domain (Sun *et al.*, 2004). Split sites located near the intein C-terminus have been created of natural split inteins *Ssp*DnaE intein and *Npu*DnaE intein (Aranko *et al.*, 2009). The shorter C-split intein consisted of 15 and 14 residues for *Ssp*DnaE intein and *Npu*DnaE intein, respectively. The shorter split inteins are of high interest for site-specific chemical modification of protein (see Figure 8). Short peptides are cheaper for chemical synthesis making them more attractive. A short peptide can be modified as desired and such obtained a modified peptide can be attached to POI using inteins. Split inteins have been used to generate proteins with different modifications e.g. labelling of fluorescence tags at N- or C-termimus (Ludwig *et al.*, 2006; Kurpiers and Mootz, 2007; Volkamnn and Liu, 2009), lipidation (glycosylphosphatidylinositol-anchor) (Olschewski et al., 2007), and *in cell* labelling (Giriat and Muir, 2003; Borra *et al.*, 2012). Split intein technology has also been utilised for segmental isotope labelling and protein cyclization (see Section 2.4.2 and 2.4.3).

2.4.2 Segmental Isotope Labelling

Inteins have a big potential in biological NMR to overcome problems of signal overlap when working with larger proteins (>25 kDa). An approach to reduce signal overlap is to incorporate single amino acid residues labelled thereby less signals are observed. This method may not be practically applicable because of the cost and time needed for sample preparation. Intein mediated protein *trans* splicing makes it possible to incorporate a

specific part of the peptide chain stable isotopic labelled rather than only single amino acids (Wider and Wüthrich, 1999). This also reduces the amount of signal overlaps but the approach is applicable for conventional NMR schemes when performing resonance assignment and structure determination (Vitali *et al.*, 2006). Segmental isotope labelling makes NMR studies on non-truncated proteins feasible which may be important for understanding the biological function of proteins.

The principle of segmental isotope labelling was first demonstrated by labelling the C-terminal domain of the RNA polymerase α subunit using artificially engineered split PI-*Pfu*I inteins (Yamazaki *et al.*, 1998). The different split parts were expressed separately and protein splicing was performed *in vitro*. Later segmental isotope labelling of maltose binding protein (Otomo *et al.*, 1999) and the F_0F_1 ATPase β subunit (Yagi et al., 2004) has been demonstrated. However, wide utilization of segmental isotope labelling using artificial split inteins has been limited due to the precursor segments may become insoluble. Refolding from denaturizing conditions and optimization of reaction conditions hamper the application.

A significant improvement in segmental isotope labelling technique came with the idea to perform it *in vivo* (Züger and Iwai, 2005). The principle is to make a segmental isotope labelled protein from a single culture. The technique requires co-transformation of cells with two plasmids containing the different precursor segments. Each vectors contains a different promoters that allow controlled protein expression of the different precursor segments. Initially one of the precursor segments is expressed in an isotope labelled or unlabelled form. The cell culture is then collected by centrifugation and the media is exchanged and the other precursor part is expressed. This technique has been used to segmental isotopic label multiple domain proteins with a molecular weight up to 140 kDa (Muona *et al.*, 2008; Minato *et al.*, 2012).

The described segmental isotope labelling technique allows labelling of the N- or C-terminal part of POI (see Figure 8). However, segmental isotope labelling of a middle part posses an additional technical challenge. The principle of labelling the middle part has been demonstrated using two orthologous engineered split PI-*Pfu*I and PI-*Pfu*II inteins (Otomo *et al.,* 1999). However, the yield from the protein splicing reaction was limited by the need of protein refolding and reaction condition optimization. Three piece ligation has been improved by the use of one natural robust split *Npu*DnaE intein. The reaction was performed in combination of *in vivo* and *in vitro* ligation but did not require refolding of precursor fragments (Busche *et al.*, 2010). A later approach for three piece ligation used electrostatic controlled reaction by engineered *Npu*DnaE intein for in vitro ligation (Shah *et al.*, 2011).

2.4.3 Protein Cyclization

Cyclization of proteins or small peptides is an interesting modification that makes polypeptide chains chemically and thermally more stable and resistant to enzymatic degradation from exopeptidases. Cyclic polypeptide chain would be of particular interest for therapeutic applications where increased half-life of polypeptide as therapeutic agents

would be of importance. Natural cyclic peptides exist derived from plants, mammals, and microbes where they have been found to have various biological properties e.g. antimicrobial, trypsin inhibiter, and pesticide (Saether *et al.*, 1995; Tang *et al.*, 1999; Luckett *et al.*, 1999). The exact mechanism for synthesis of cyclic peptides is complex and it is not well understood (Craik, 2006). Protein cyclization could be induced *in vivo* and *in vitro* using inteins.

Split intein arranged on a single extein sequence can be used for protein cyclization (see Figure 8). The C-intein is fused to the N-terminal of the extein sequence and the N-intein is fused to C-terminal of the extein sequence. This method has been used to produce recombinant cyclic polypeptide chain of dihydrofolate reductase (Scott *et al.*, 1999), maltose binding protein (Evans *et al.*, 2000), green fluorence protein (Iwai et al., 2001), bacterial acyl carrier protein (Volkmann *et al.*, 2010), and *Momordica cochinchinensi*s trypsin inhibitor II (Jagadish *et al.*, 2013).

A different intein-mediated approach for protein cyclization is to use inteins to generate a thioester that is used in native chemical ligation or expressed protein ligation (EPL) (Muir *et al.*, 1998; Evans *et al.*, 1998). EPL is based on the reaction between a polypeptide containing a C-terminal thioester and a polypeptide containing an N-terminal cysteine (Dawson *et al.*, 1994). Inteins have been used in EPL to generate a C-terminal thioester that can undergo intra or inter molecular reaction to generate cyclic polypeptide or ligated product (Evans et al., 1999b). This method has been used to generate cyclic β-lactamase (Iwai & Plückthun, 1999), Src homology 3 domain (Camarero and Muir, 1999), and *Momordica cochinchinensi*s trypsin inhibitor II (Camarero *et al.*, 2007). A similar approach is a two-intein system (TWIN) where an intein is located at both N- and C-terminal end of POI (Evans *et al.*, 1999a). The inteins are modified in order to perform cleavage reaction at the junction sequences at POI. A problem with this approach is that inter-molecular reactions can occur which can lead to polymerization rather the intra-molecular reaction that would form a cyclic peptide.

Inteins have many possible biotechnological applications. However, these applications are only interesting if the protein splicing reaction is very efficient. This highlights the requirement of a better understanding on how foreign extein sequences influence protein splicing. A way to explain and understand the function of macromolecules is by structure determination.

2.5 Structure Determination Methods

The main methods used in structural biology include NMR spectroscopy, X-ray crystallography, and cryo-electron microscopy. In this thesis NMR spectroscopy and X-ray crystallography have been used for structure determination of proteins at high resolution. In the following section the general procedure for structure determination using the two methods is described.

2.5.1 Protein Sample Preparation

Structural studies with NMR spectroscopy and X-ray crystallography require similar initial steps of molecular cloning to generate a protein construct that can be used for recombinant protein production and purification. However, X-ray crystallography is not dependent on recombinant proteins production. Proteins that are highly abundant in a natural source can be purified directly and used for structural studies. NMR spectroscopy structure determination requires stable isotope labelling (^{2}H, ^{13}C, ^{15}N) of target protein. NMR studies of protein are limited by the ability to obtain a stable isotopic labelled sample. Protein expression is then typically limited to expression of recombinant proteins in *E. coli* or cell free protein expression systems (Ohki and Kainosho, 2008). The protein should be stable (no degradation or precipitation) at experimental conditions for the time needed to complete measurements. If the protein is unstable or unfolded re-cloning of a new protein constructs might be needed (see Figure 9a). For structure determination standard NMR spectroscopy techniques is limited to molecules at ~25-30 kDa (Goto and Kay, 2000). In higher molecular weight systems signal overlap becomes a problem and too fast relaxation of the NMR signal becomes a problem because of slow molecular tumbling properties. Transverse relaxation optimized spectroscopy (TROSY) is a technique that is used when working on higher molecular systems (Pervushin *et al.*, 1997). TROSY is used to select slowly relaxing NMR signals that make it possible to work with large proteins. Only in special cases with special isotopic labelling structure determination of higher molecular weight systems is possible (Tugarinov *et al.*, 2005; Kainosho *et al.*, 2006; Gautier et al., 2010).

X-ray crystallography has been used for structure determinations at atomic details of large macromolecules (Wynne *et al.*, 1999; Ban *et al.*, 2000). The limiting factor in X-ray crystallography is the ability to obtain protein crystals that diffract X-rays to high resolution. The formation of protein crystals requires a soluble protein that is able to self-associate (nucleation process). Self-association is a random process that depends on weak molecular forces. Thus, a very pure protein sample (>99%) is needed to improve the chances for the association to occur. The process of crystallization has been highly automated for high throughput (Chayen and Saridakis, 2008). Crystallization setup is used to screen a large sampling space for different crystallization conditions.

2D NMR spectra are typically assessed to judge quality the NMR sample. From a 2D NMR spectrum the quality of the protein sample can be assessed by determining if the number of expected signals is present and the signals well resolved or if the protein is unfolded. This gives useful information whether a new protein construct would be required (see Figure 9). NMR spectroscopy has been suggested as useful tool for screening of protein constructs that could be selected for crystallization based on 1D NMR spectra quality (signal dispersion/line width/folded state) (Page *et al.*, 2005). However, studies have indicated there is no correlation between the 2D NMR spectra quality that are suitable for NMR structure determination and the success rate in crystallization studies (Snyder *et al.*, 2005; Yee *et al.*, 2005). It is not always possible to obtain protein crystals for structure determination and protein constructs might need optimization. Protein crystallization relies on weak inter-molecular interactions to form crystals. Flexible and

unstructured part is often a disadvantage and re-cloning of construct is needed before new screening trials for new crystallization conditions are performed (see Figure 9b) (Oksanen and Goldman, 2010; Rupp, 2010).

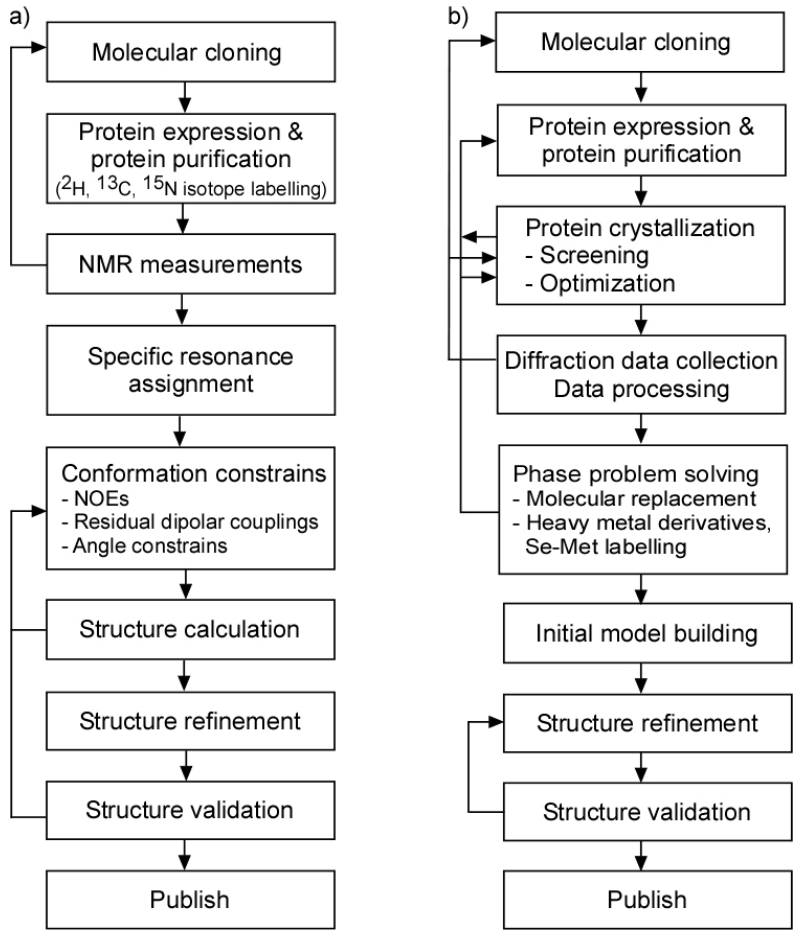

Figure 9 *Schematic flow diagram for a) NMR protein structure determination and b) X-ray crystallography protein crystal structure determination.*

2.5.2 Structure Determination

Structure determination using NMR spectroscopy is based on data collection of a series of different spectra that contain different information. In a combined approach the information is used for the data analysis. Typically NMR structure determination is based on a series of 2D, 3D, or higher dimensional spectra to make the (near) complete resonance assignment, and structure determination (Sattler *et al.*, 1999; Güntert, 2009). For structure determination (almost) complete specific resonance assignment is needed together with distance constrains derived form NOESY spectra. Protein NMR structure determination is commonly based upon the measurements of Nuclear Overhauser Effect (NOE). NOEs are observation of nuclear dipole cross-relaxation that is transferred though space and are measured in NOESY spectra. Distance information is obtained from the

experiment of atoms up to ~6 Å apart. NOEs are divided in strong/medium/weak NOE classes based upon signal intensity that are used in the structure calculation. A lower distance constraint is typically set to 2 Å reflecting van der Walls interaction. The three classes then have different upper limit restrain (Güntert, 2009). NOEs are divided into different distance classes because the intensity of a NOE signal can be difficult to relate accurately to a distance (Cavanagh *et al.*, 2007). Signal intensity can be affected by spin diffusion, signal overlap, or internal dynamics. Signal information from NOESY spectra and the specific resonance assignment are used for input in programs like CYANA (Güntert, 2009), or ARIA (Habeck *et al.*, 2004). The softwares have slightly different approaches for assignment of NOEs. CYANA that is used in this thesis applies a network-anchoring approach (Hermann *et al.*, 2002). Network anchoring is based on the observation of a NOE between two residues are not manifested by only a single NOE but several NOEs are observed between the two residues (Güntert, 2009). The method relies on calculation of an initial structure model that is used to filter (reject) NOE assignments. A nearly complete resonance assignment is needed for accurate structure determination (Jee and Güntert, 2003) because calculation of correct initial structure is very important. Many missing or wrong resonance assignments make the NOE assignment unreliable. After an initial structure calculation several cycles of structure calculation follows which does not change the overall structure but gradually improved the NOE assignment and structure root-mean-square deviation (RMSD). Besides NOE information additional data can be used as input for the structure determination (Güntert, 1998) including secondary structure information based upon chemical shifts (Wishart and Sykes, 1994; Cornilescu *et al.*, 1999), hydrogen bonds (Cordier and Grzesiek, 1999), scalar couplings (Torda *et al.*, 1993), or residual dipolar couplings (Tolman *et al.*, 1995). The different restrains can be applied in structure calculation where torsion angle dynamics typically is used and this is followed by an energy refinement in Cartesian space (Güntert *et al.*, 1991; Güntert *et al.*, 1997) (see Figure 9a). Molecules in solution are dynamic and conformational exchange occurs, thus the NMR data describes an average of possible conformations. Consequently, the final NMR model is an assembly of typically 20 NMR structures models that best describes the NMR data. This can be seen as less precise for the structure determination but this is perhaps one of the advantages by NMR spectroscopy. NMR spectroscopy provides the possibility to measure dynamic properties of proteins at different time scale (Palmer, 2004; Boehr *et al.*, 2006).

As mentioned before the limiting factor in X-ray crystallography is the ability to obtain crystals that diffract X-ray waves to high resolution. A protein crystal is macroscopic object where the molecule atoms are arranged with similar orientation (Oksanen and Goldman, 2010). Protein crystals are formed by a phase transition of the protein from liquid state to crystalline state. This is achieved in experimental setup by varying different parameters like protein concentration, temperature, pH, and precipitant concentration. The protein is subjected to a state transition from a soluble phase to a metastable zone and a nucleation zone. In the nucleation zone spontaneous association can occur and protein crystals start forming (Chayen and Saridakis, 2008). In the metastable phase new protein crystals are not formed but existing crystals continue to grow. Ideally, protein crystal consists of repeating identical units, referred to as the unit cell. The unit cell is defined as

the smallest repeating unit where unit cell can be translated in all directions to make up identical units. Protein crystals are used for diffraction measurements by applying high energy X-rays on the protein crystal. The X-ray beam is scattered by the electrons located in the crystal lattice that are arranged in an ordered manner (Rupp, 2010). The scattered X-rays generate a diffraction pattern and the signal intensities of the diffractions are measured.

Initially the measured reflections are indexed, scaled, and space group is assigned based on symmetry. From diffraction experiments only the intensity can be measured and the phase information is lost. The intensity of the diffraction is proportional to the square of the structure factor. The structure factor is complex number that describes the amplitude and phase of wave but the phase information is lost in the experiment and is referred to as "the phase problem" and is the central problem in X-ray crystallography (Rupp, 2010). Solving the phase problem is needed to obtain an electron density map that can be interpreted. The phase problem is solved by additional experiments that include incorporation of heavy-metal derivatives, Selenomethionine labelling, or molecular replacement (use existing model to obtain phase information). Heavy atom derivatives are used to solve the phase problem by measurements of a native data set and a data with the incorporated heavy atoms. The data sets can then be used together to determine the phases. Solving the phase problem can only be performed if the crystals are isomorphous, meaning the incorporation of a heavy atom does not change the unit cell or the macromolecular conformation. The change in diffraction should only be the presence of the heavy atom. Selenomethionine labelling is a different approach to solve the phase problem by incorporation of heavy atoms (Hendrickson *et al.*, 1990). Alternatively if an existing or similar structure model is available of the target macromolecule the phase problem can be solved by molecular replacement (Rossmann, 1990). After obtaining initial phases the electron density maps can be calculated and by iterative refinement and model building cycles a refined model is obtained (Rupp, 2010) (see Figure 9b).

3 Aims of the Study

The aim of this thesis is to structurally characterize selected inteins using NMR spectroscopy and X-ray crystallography. From the structure determination the second aim is to use the structural information to guide in protein-engineering approaches. The third aim is to elucidate how the extein-intein parts interact at a structural level and apply this knowledge for rational engineering of an intein for efficient protein splicing in a non-natural, foreign extein context.

Specific aims:

- To gain high-resolution structural information on *Npu*DnaE intein and *Pho*RadA intein.
- To investigate molecular dynamics of *Npu*DnaE intein with NMR spectroscopy.
- To use structural data to guide in protein engineering of *Npu*DnaE intein and *Pho*RadA intein.
- To characterize extein sequence residue influence on protein splicing efficiency and further understand extein-intein interactions in protein splicing.

4 Materials and Methods

This section summarizes the methods used in the described research and is meant to assist in understanding the following sections. For more detailed information the reader is referred to the individual publications.

4.1 Molecular cloning

4.1.1 Single Chain Variant of *Npu*DnaE Intein

A single chain variant of the natural split *Npu*DnaE intein was constructed by genetically assembling the N- and C-intein fragments derived from genomic DNA of *Nostoc punctiforme* (Iwai *et al.,* 2006). A C1A mutation was introduced in *Npu*DnaE intein to prevent any cleavage reaction from occurring. The PCR product was inserted in pHYRSF53 vector resulting in pJDJRSF05 that expresses an N-terminal hexahistidine and yeast Smt3 purification tag followed by *Npu*DnaE intein single chain variant (Muona *et al.*, 2008). The final primary structure consists of a 137 residue single chain variant of *Np*uDnaE intein with two N-extein glycine residues.

For crystallization of *Npu*DnaE intein an inactive protein splicing precursor was generated amplifying *Npu*DnaE intein using PCR and insert the product into pHYRSF53 vector generating pALBRSF12 (Aranko *et al.*, 2013b). The final primary structure consists of a 137 residue single chain variant of *Np*uDnaE intein with three and four N- and C-extein residues, respectively. The construct contains a C1A and C+1A mutation to prevent any reaction from occurring.

4.1.2 *Pho*RadA Intein

The gene of *Pho*RadA intein was cloned from genomic DNA of *Pyrococcus horikoshii* (ATCC 700860). *Pho*RadA intein was amplified using PCR introducing a C1A mutation in the intein and a T+1A mutation in a two residue C-extein sequence (AQ). The PCR product was inserted into pRSET-A vector (Invitrogen) that expresses a 172 residue inactive intein with a two residue C-extein.

4.1.3 *Pho*RadA$_{min}$ Intein

*Pho*RadA intein was minimized genetically by removal of residues 121-130 and introducing a K131N mutation to accommodate a turn. *Pho*RadA$_{min}$ intein was amplified and inserted into pHYRSF53 vector generating pCARSF15 (Muona *et al.*, 2008). The

construct expresses an N-terminal hexahistidine and yeast Smt3 fusion purification tag followed by *Pho*RadA$_{min}$ intein. The final primary sequence consists of a 162 *Pho*RadA$_{min}$ intein with a four and a two residue N- and C-extein, respectively.

4.1.4 Mutagenesis of *Pho*RadA Intein

Variants of *Pho*RadA intein were generated by mutating the -1 amino acid residue to all 20 standard residues. The 20 mutant constructs were made using pHYDuet183 as template (Ellilä *et al.*, 2011). The construct expressing a hexahistidine-GB1-X-*Pho*RadA intein-GB1 protein splicing precursor where X and GB1 denotes residue subjected for 20 residue mutation and the B1 domain of IgG binding protein G, respectively.

The mutation was introduced into pHYDuet183 by cassette mutagenesis using *Bse*RI and *Hind*III restriction sites or QuikChange Protocol (Stratagene) with synthetic oligonucleotides containing the mutations. An E71T mutation was introduced into constructs of *Pho*RadA intein bearing a D-1, E-1, and K-1 N-junction residue.

4.2 Evaluation of *Cis*-Splicing

In vivo cis-protein splicing efficiency of *Pho*RadA intein variants were evaluated by expression of individual constructs in *E. coli* ER2566 strain. The cells were grown in 5 mL LB-media supplied with 25 µg/mL kanamycin at 37 °C. At OD_{600} ~0.5-0.6 protein expression was induced by addition of IPTG to a final concentration of 1 mM and protein expression continued for four hours. Cells were harvested by centrifugation at 4,500xg at 4 °C for 10 minutes and cell pellet was kept for further purification.

Cell pellets were resuspended and lysed in 100 µL B-PER® Bacterial Protein Extraction Reagent (Thermo Scientific) and suspensions were incubated at 25 °C for 10 minutes. Cell lysates were cleared by centrifugation at 14.100xg for 5 minutes and the supernatant was loaded on Ni-NTA spin column (Qiagen). Unbound protein was removed by washing column with 50 mM sodium phosphate buffer, 300 mM NaCl (pH 8.0). Bound protein was eluded from spin column by applying 150 µL 50 mM sodium phosphate buffer, 300 mM NaCl, and 250 mM imidazole (pH 8.0). Elution fractions were analysed by loading samples on 18% SDS-PAGE and protein bands were visualized with PhastGelTM Blue R (GE Healthcare) stain. Protein band intensities were quantified using ImageJ 1.45 and protein splicing efficiency was determined from protein band intensity assuming the staining dye equally binds to all proteins. Protein splicing efficiency was determined from three independent protein expressions.

4.3 Protein Expression

4.3.1 Unlabelled and $^{15}N,^{13}C$ Labelled Sample of *Npu*DnaE Intein

*Npu*DnaE intein was expressed in *E. coli* ER2566 strain grown at 37 °C. For unlabelled and stable isotope labelled sample expressions were conducted in LB-medium and M9 media supplied with 25 µg/mL kanamycin, respectively. For stable isotope labelled protein expression the M9 medium was supplied with $^{15}NH_4Cl$ and ^{13}C-D-glucose as sole nitrogen and carbon sources, respectively. Cells were grown to OD_{600} 0.6 and protein expression was induced with 0.5 mM IPTG and expression continued for three hours. The cells were harvested by centrifugation at 8,900xg at 4 °C for 10 minutes. The cell pellet was respuspended in 50 mM sodium phosphate buffer (pH 8.0), and 300 mM NaCl and stored until protein purification at -80 °C.

4.3.2 Unlabelled Sample of *Pho*RadA Intein

*Pho*RadA intein was expressed in *E. coli* ER2566 strain in LB-medium supplied with 100 µg/mL ampicillin. Cells were grown at 37 °C and at OD_{600} ~0.6 protein expression was induced with 0.1 mM IPTG. Protein expression continued for three hours and cells were harvested by centrifugation at 8,900xg at 4 °C for 10 minutes. The cell pellet was respuspended in 50 mM Tris-HCl, 10 mM NaCl, 1mM EDTA (pH 7.9) and flash frozen for storage at -80 °C until protein purification.

4.3.3 $^{15}N,^{13}C$ Labelled Sample of *Pho*RadA intein

*Pho*RadA intein was expressed in *E. coli* ER2566 strain co-transformed with pLysRare plasmid to compensate for rare codons and suppress leaky expression. Cells were initially grown in LB-media supplied with 0.05% (w/v) D-glucose, 100 µg/mL ampicillin, and 5 µg/mL chloramphenicol until OD_{600} reached ~0.5. Cells were collected by centrifugation at 900xg at 25 °C for 10 min and cells were respuspended in M9-media supplied with 1.3 g/L $^{15}NH_4Cl$ and 2 g/L ^{13}C-D-glucose as sole nitrogen and carbon source, respectively. Cells were then induced with a final IPTG concentration of 1 mM IPTG and protein expression continued for four hours. Cells were collected by centrifugation at 8,900xg at 4 °C for 10 minutes. Cell pellet was resuspended in 50 mM Tris-HCl, 10 mM NaCl, and 1 mM EDTA (pH 8.0) and subsequently flash frozen in liquid nitrogen and stored at -80 °C until protein purification.

4.3.4 Unlabelled Sample of *Pho*RadA$_{min}$ Intein

*Pho*RadA$_{min}$ intein was expressed in *E. coli* ER2566 strain in LB-medium supplied with 25 µg/mL kanamycin. Cells were grown at 37 °C and at OD_{600} ~0.6 protein expression was induced with a final IPTG concentration of 1 mM. Protein expression continued for 3 hours and cells were harvested by centrifugation at 6,700xg at 4 °C for 10 minutes. The cell pellet was respuspended in 50 mM sodium phosphate buffer (pH 8.0), and 300 mM NaCl and flash frozen and store at -75 °C until protein purification.

4.4 Protein Purification

4.4.1 *Npu*DnaE Intein

Cells were disrupted by ultrasonication and cell lysate was cleared by centrifugation. The supernatant was loaded on a HisTrapFF column (GE Healthcare) and unbound protein was removed by washing the column with buffer (50 mM sodium phosphate buffer and 300 mM NaCl, pH 8.0). Bound protein was eluted from column by applying 50 mM sodium phosphate buffer, 300 mM NaCl, and 250 mM imidazole (pH 8.0). Hexahistidine-tagged Smt3-*Npu*DnaE intein fusion protein was dialysed against PBS buffer and after dialysis the protein was digested with N-terminal hexahistidine tagged ubiquitin-like protease-1 (Ulp1) (Mossessova and Lima, 2000). The digested sample was loaded onto a HisTrapFF column to remove hexahistidine tagged Smt3 and Ulp1. *Npu*DnaE intein was collected in the flow-through and dialysed against 10 mM sodium phosphate buffer (pH 8.0). *Npu*DnaE intein was concentrated using centrifugal concentrator (Amicon Ultra 4). The buffer for unlabelled and the stable isotope labelled sample was exchanged with MQ-grade water and 10 mM $NaPO_i$ buffer (pH 6.0), respectively.

4.4.2 *Pho*RadA Intein

Cell suspension was lysed by heating the suspension at 75 °C for 20 minutes. The solution was cleared by centrifugation at 34,000xg at 4 °C for 40 minutes. DNase I was added to the supernatant and suspension was incubated at 25 °C for 2.5 hours. The supernatant was loaded on DEAE Sephearose FF 5 mL column (GE Healthcare). Bound protein was eluted with a gradient of 2.5 M NaCl and fractions containing *Pho*RadA intein were dialysed against 10 mM sodium phosphate buffer (pH 8.2). The dialysed sample was further purified on a MonoQ 5/50 GL ion exchange column and protein was eluted with a gradient of 2.5 M NaCl. Fractions containing *Pho*RadA intein were concentrated using a centrifugal concentrator (Vivaspin, GE Healthcare). For protein crystallization of unlabelled protein the buffer was exchanged with Milli-Q-grade water. Stable isotope labelled protein buffer was exchanged with 20 mM NaP_i buffer (pH 6.0).

4.4.3 *Pho*RadA$_{min}$ Intein

Thawed cells were disrupted by ultrasonication and cell lysate was cleared by centrifugation at 42,000xg at 4 °C for 45 minutes. The supernatant was loaded on a HisTrapFF column and bound protein was eluted with 50 mM sodium phosphate, 300 mM NaCl, and 250 mM imidazole (pH 8.0). Fractions containing *Pho*RadA$_{min}$ intein fusion was dialysed against PBS buffer. The dialysed fraction was digested with N-terminal hexahistidine tagged Ulp1 and loaded on HisTrapFF column and *Pho*RadA$_{min}$ intein was collected in the flow-though. *Pho*RadA$_{min}$ intein was concentrated using centrifugal concentrator (Amicon Ultra 3000-molecular-weight cutoff) and buffer was exchanged with Milli-Q-grade water

4.5 NMR Studies

4.5.1 NMR Measurements

All NMR measurements were performed on a Varian INOVA 600 MHz or 800 MHz both equipped with a cold-probe. Measurements on *Npu*DnaE intein were performed on a 2 mM protein concentration in 10 mM NaP_i buffer (pH 6.0) at 298 K. Measurements on *Pho*RadA intein was performed on a 0.4 mM protein sample in 20 mM NaPi buffer (pH 6.0) at 308 K.

For resonance assignment of *Npu*DnaE intein and *Pho*RadA intein a series of standard NMR spectra were recorded: [^{1}H,^{15}N]-HSQC, [^{1}H,^{13}C]-HSQC, HNCA, HN(CO)CA, HNCACB, CBCA(CO)NH, HNCO, HN(CA)CO, and CC(CO)NH, HBHA(CO)NH, HNHB, CC(CO)NH, HCC(CO)NH, H(C)CH-TOCSY with 50 ms mixing time, HCCH-COSY, ^{15}N-edited TOCSY with 50 ms mixing time. The assignments of aromatic residue side-chains were based on (HB)CB(CGCD)HD, (HB)CB(CGCDCE)HE, and CT-[^{1}H,^{13}C]-HSQC spectra.

All relaxation measurements were performed at a proton frequency of 600 MHz. *Npu*DnaE intein relaxation data was based upon following measurements. T_1(^{15}N)-relaxation rates were determined using following T_1 delays: 0, 50, 100, 150, 200, 300, and 500 ms. T_2(^{15}N)-relaxation rates were determined using a CMPG-type sequence with the interval of 1.3 ms between ^{15}N 180° pulse in the CPMG cycles with the following T_2 relaxation delays: 10, 30, 50, 70, 90, 110, 150, and 190 ms (Farrow *et al.*, 1994). Heteronuclear ^{15}N{^{1}H}NOEs were determined by comparing peak intensity of [^{1}H,^{15}N]-HSQC spectra with and without 2.5 seconds proton saturation. *Pho*RadA intein relaxation data was based upon following measurements. T_1(^{15}N) relaxation rates were determined using the following T_1 delays: 10, 30, 50, 70, 90, 110, 130, and 150 ms. For T_2(^{15}N) relaxation rates, the following T_2 delays were used: 10, 30, 50, 70, 90, and 110 ms. Heteronuclear ^{15}N{^{1}H}NOEs were determined by comparing peak intensity of [^{1}H,^{15}N]-HSQC spectra with and without 4.0 s of ^{1}H saturation.

Spectra were recorded using VNMRJ and processed using NMRPipe (Delaglio *et al.*, 1995). Specific resonance assignment and data analysis were performed using Sparky (SPARKY 3, University of California, San Francisco) and CcpNmr analysis (Vranken *et al.*, 2005). Peak intensities were used for determination of relaxation rates.

4.5.2 NMR Solution Structure Determination

For structure determination of *Npu*DnaE intein a ^{15}N- and ^{13}C-edited NOESY-HSQC were recorded with 80 and 70 ms mixing time, respectively. For structure determination of *Pho*RadA intein a ^{15}N-edited NOESY-HSQC spectrum and a sensitivity enhanced ^{13}C-edited NOESY-HSQC spectrum were recorded with 80 ms and 100 ms mixing time, respectively.

Structure calculation was performed using CYANA v. 3.0 (Güntert *et al.*, 2009). Chemical shift and unassigned NOE-peak lists were used as input for automatic NOE assignment incorporated in CYANA (Hermann *et al.*, 2002). TALOS+ protein backbone angle were used as additional input data for the structure determination of *Pho*RadA intein (Shen *et al.*, 2009). Calculations were performed using torsion angle dynamics calculation with 100 structure calculations in each cycle. The 20 structures with lowest CYANA target function in the final calculation cycle were energy refined using AMBER (Bertini *et al.*, 2011). Protein structure quality was validated using PROCHECK-NMR (Laskowski *et al.*, 1996) and NMR-CING software (Doreleijers *et al.*, 2012).

4.6 X-ray Crystallography

4.6.1 Protein Crystallization

For protein crystallization *Npu*DnaE intein, *Pho*RadA intein and *Pho*RadA$_{min}$ intein were concentrated to 29 mg/mL, 7.1 mg/mL and 31 mg/mL, respectively. Protein crystallization experiments were performed using sitting drop vapour diffusion method at 293 K. Crystallization setups were performed using Cartesian MicroSys pipetting robot mixing 100 nL protein drops and 100 nL reservoir solution in an Innovadyne SD2 96-well plate with 80 µL well solution. Initial crystallization condition was obtained using Helsinki Random I and II screen (www.biocenter.helsinki.fi/bi/xray/automation/index.html). From initial crystallization hits, crystallization conditions were optimized using grid design varying pH, protein concentration, salt, and PEG concentration. *Npu*DnaE intein was collected from a drop containing 1.4 M tri-ammonium citrate/citric acid (pH 6.5). *Pho*RadA intein was collected from a crystal grown over several months in 0.1 M HEPES (pH 7.5) and 3 M NaCl. *Pho*RadA$_{min}$ intein was collected from a drop containing 1.6 M tri-sodium citrate. Crystals were picked and cryo-protected using Paratone-N before freezing in liquid nitrogen.

4.6.2 Diffraction Data Collection, Processing, and Structure Determination

Protein diffraction data were collected at European Synchrotron Radiation Facility (ESRF) in Grenoble, France. Diffraction data of *Npu*DnaE intein, *Pho*RadA intein and *Pho*RadA$_{min}$ intein were collected at 100 K on Beam line ID23-1 and ID14-1. Diffraction data were processed using the XDS (Kabsch, 1993) or HKL2000 (Otwinowski and Minor, 1997) software packages. The phases of the *Npu*DnaE intein crystal structure determination protocol were solved using the NMR model of *Npu*DnaE intein as input using Phaser (PDB: 2KEQ, Oeemig *et al.*, 2009 (Study II)) (McCoy *et al.*, 2007). The phase of the *Pho*RadA intein structure determination protocol were solved using a preliminary NMR model of *Pho*RadA intein as search model using a Rosetta force-field (DiMaio *et al.*, 2011). The phases of *Pho*RadA$_{min}$ were solved using Phaser with the crystal structure of *Pho*RadA intein as search model (McCoy *et al.*, 2007). Refinement was performed using Refmac5 (Murshudov *et al.*, 1997) and PHENIX (Adams *et al.*, 2002). Coot was used to manually model the protein structures in the electron density map (Emsley and Cowtan, 2004).

5 Results and Discussion

5.1 Protein Crystallization

Protein crystallization was utilized to investigate the structure of *Npu*DnaE intein and *Pho*RadA intein. In order to study the protein structures with X-ray crystallization conditions of the proteins were screened using sitting drop vapour diffusion. Promising crystallization conditions were optimized with grid designed around conditions where initial crystal hits were observed. Crystal hits from screens and grid design are shown in Figure 10. An engineered minimized version of *Pho*RadA intein (*Pho*RadA$_{min}$) (see Section 4.1.3 and Section 5.4) was crystallized using similar approach as with *Npu*DnaE intein and *Pho*RadA intein.

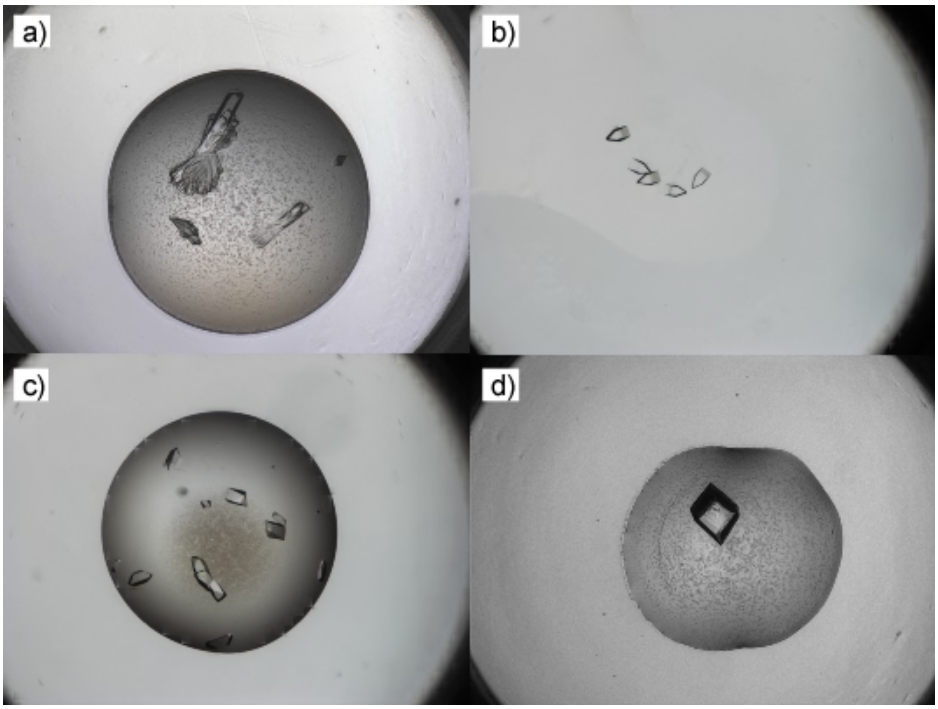

Figure 10 *Protein crystals. a)* Npu*DnaE intein grown in 1.4 M tri-ammonium citrate/citric acid (pH 6.5). b)* Pho*RadA intein grown in 0.1 M HEPES pH 7.5 and 3 M NaCl. c)* Pho*RadA intein grown in 0.1 M HEPES pH 7.0, 10.4% Tacsimate pH 7.0, and 3.0% PEG 3350. d)* Pho*RadA$_{min}$ crystal grown in 1.6 M tri-sodium citrate.*

Crystals were picked and cryo-protected before data collection at ERSF Grenoble. The best diffracting data sets of *Npu*DnaE intein, *Pho*RadA intein, and *Pho*RadA$_{min}$ intein were collected and refined to 1.72 Å, 1.75 Å, and 1.58 Å resolution, respectively (see section 5.2 and section 5.4).

5.2 *Npu*DnaE Intein Structure

*Npu*DnaE intein was selected for structural studies because studies have shown it has fast, efficient, and robust protein splicing activity compared to other inteins (Iwai *et al.*, 2006; Zettler *et al.*, 2009; Ellilä *et al.*, 2011). Thus, *Npu*DnaE intein is a good target for protein engineering and it is of interest to understand the structural features of *Npu*DnaE intein and therefore structural studies were performed. At the time when the studies I and II were performed one NMR intein structure had been determined (Johnson *et al.*, 2007b; Johnson *et al.*, 2007a) and the NMR structure of *Npu*DnaE intein could give a further insight into the structural features and dynamics of intein. For the studies a single chain variant of *Npu*DnaE intein was constructed by assembling the split intein parts genetically. This was performed to make the structure determination easier by having the molecule consisting of one polypeptide chain instead of two individual polypeptide chains. The single chain variant of *Npu*DnaE intein shows high protein splicing efficiency as observed with the natural split intein (Iwai *et al.*, 2006; Ellilä *et al.*, 2011).

The single chain variant of *Npu*DnaE intein was made inactive by a C1A mutation to prevent possible reaction occur during the experiments. The construct was expressed stable isotopic labelled to see if the construct was suitable for structural studies. From initial analysis of a [^{1}H-^{15}N]-HSQC spectra of *Npu*DnaE intein it was observed the amide proton signals are well dispersed between 6-11 ppm (see study I), which indicates a folded protein. Most residues of *Npu*DnaE intein could be assigned from the [^{1}H-^{15}N]-HSQC spectrum with the exception of residues Gly-1, 122-124, and Asn131. In total 96.4% of H^{N}, Hα, N, CA, and CO atoms were assigned, and 96.2% of expected side chain atoms were assigned. The structure of *Npu*DnaE intein was determined based on NOE distance constrains derived from ^{15}N- and ^{13}C-edited NOESY-HSQC spectra. The statistics from the structure determination and refinement are listed in Table 2.

The structure of *Npu*DnaE intein resembles a complex horseshoe-disk shaped fold as seen in the HINT fold (Hall *et al.*, 1997) (see Figure 11). The backbone RMSD is 0.45 ± 0.07 Å indicating an overall well defined structure. The most disordered region is located at the loop between β-strands 12 and 13 where amide signals were unassigned at residues 122-124, consequently no NOEs could be assigned to fix the backbone conformer.

The crystal structure of *Npu*DnaE intein was determined at 1.72 Å resolution with two models in the asymmetric unit (see Table 3) (Aranko *et al.*, 2013b) (see Figure 12). The two models are similar with a backbone RMSD of 0.404 Å for residue 1-137. The backbone RMSD of a mean model of the NMR structure and the crystal structure chain A and B is 0.823 Å and 0.658 Å, respectively.

A mean structure of *Npu*DnaE intein NMR structure was subjected to a Dali server search (Holm and Rosenström, 2010). The search showed *Ssp*DnaE intein (PDB: 1ZD7) (Sun *et al.*, 2005) has the most similar structure fold with a backbone RMSD of 1.1 Å covering 137 residues. Other similar structures include *Ssp*DnaB intein (Ding *et al.*, 2003) and *Mtu*RecA intein (Van Roy *et al.*, 2007) both with a backbone RMSD of 2.0 Å covering 135 and 133 residues, respectively. Thus, the structure of *Npu*DnaE intein resembles the structures of other known inteins and *Npu*DnaE intein is most similar to *Ssp*DnaE intein that has 63% sequence identity.

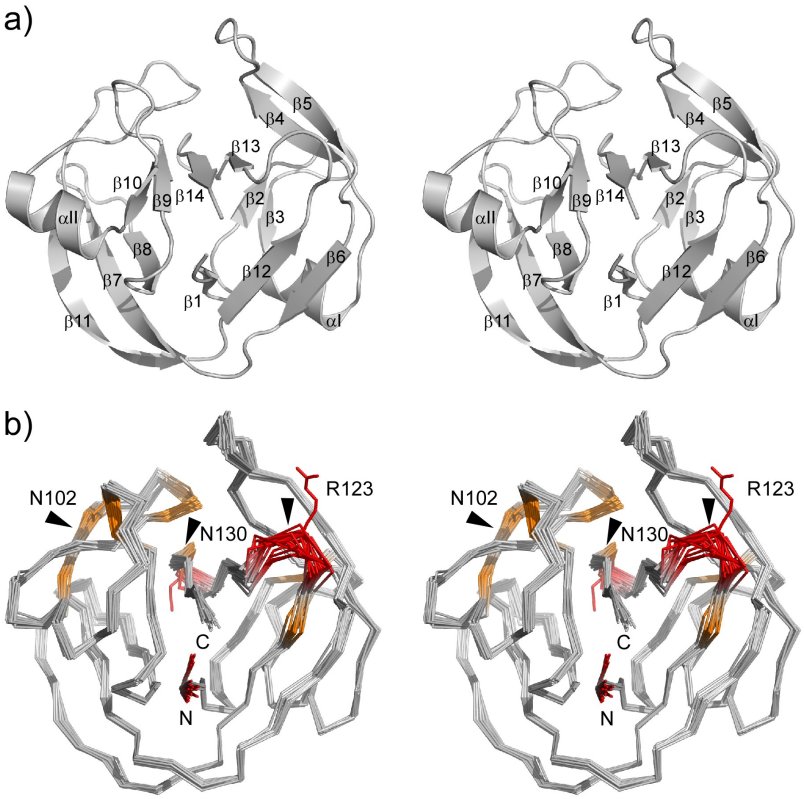

Figure 11 *Stereo view of* Npu*DnaE intein solution NMR structure. a) Schematic structure representation with labelling of secondary structure. b) Ribbon representation of an assembly of 20 NMR models. Residues that show internal flexibility are highlighted with orange colour. Residues with unassigned amide signal are coloured red.* Npu*DnaE intein split sites are indicated with triangles, presiding residue is labelled, and side chain heavy atoms are shown for the model closest to mean. N and C indicate the N- and C-terminus, respectively.*

The structure of *Npu*DnaE intein was the second intein structure determined using NMR spectroscopy. An advantage by NMR spectroscopy is the ability to investigate dynamic properties of protein structures. Longitudinal (T_1) and transverse (T_2) relaxation rates, and heteronuclear NOE cross-relaxation were determined for *Npu*DnaE intein. Data analysis comparing T_1 and T_2 relaxation rates identified following residues to experience internal conformation exchange: Thr10, Thr76, Asp78, 96-104, G120, and Gly132 (See study II for plot of relaxation rates). The conformational exchange is seen by shortened T_2-relaxation rates that could be caused by internal dynamics.

The residues 96-104 are located at the site where the natural split site of *Npu*DnaE intein is located which indicates there is some flexibility at the natural split site of *Npu*DnaE intein. In the crystal structure the residue 98-100 (chain A) showed poor electron density and could indicates some structural flexibility. In another study it was shown that the wild type split site of *Npu*DnaE intein and *Ssp*DnaE intein could be moved to a position corresponding to residue 123 and 122, respectively (Aranko *et al.*, 2009).

Table 2. *Experimental data and statistic for the NMR structure calculation of* Npu*DnaE intein and* Pho*RadA intein.*

Quantity	*Npu*DnaE intein
NOE upper distance limits	3154
Short range NOE (i-j≤1)	1498
Medium-range NOE (1<i-j<5)	433
Long-range NOE (i-j≥5)	1223
Residual CYANA target function	0.37±0.17
Residual NOE violation	
Number ≥ 0.1 Å	4±2
Maximum [Å]	0.24±0.07
Amber energies [kcal/mol]	
Total	-5,711±27
van der Waals	-1,173±9
Electrostatic	-9,503±240
RMSD from ideal geometry	
Bond length [Å]	0.0098±0.0001
Bond angels [°]	1.979±0.016
RMSD to mean coordinate	
Backbone 1-137 [Å]	0.45±0.07
Heavy atoms 1-137 [Å]	0.94±0.07
Ramachandran plot statistics [%][a]	
Most favored regions	87.5
Additional allowed regions	12.4
Generously allowed regions	0.1
Disallowed regions	0.0
BMRB entry	16009
PDB code	2KEQ

[a]Derived by PROCHECK-NMR (Laskowski *et al.*, 1996).

The residues 122-124 are missing the amide resonance assignments, which are likely to be caused by conformational exchange that causes broading of the NMR signals. The *Npu*DnaE intein crystal structure model supports that this loop has some flexibility. In the crystal structure it was not possible to model Glu122 (chain A) side chain where the electron density was absent. But it was possible to model two conformations of Arg123 (chain A) backbone in the high-resolution electron density map. The NMR relaxation data indicated that residue Gly120 experience conformational exchange and Gly120 is located just before the flexible loop, which has been shown suitable for introducing new split sites. The conformational exchange at residue 10 is comparable with the split site identified in *Ssp*DnaB intein located after residue 11 (Sun *et al.*, 2004). The connection between conformational exchange and internal flexibility observed from the nuclear spin relaxation data and the location of split sites in inteins indicate that NMR spectroscopy could be a useful tool to identify novel split sites.

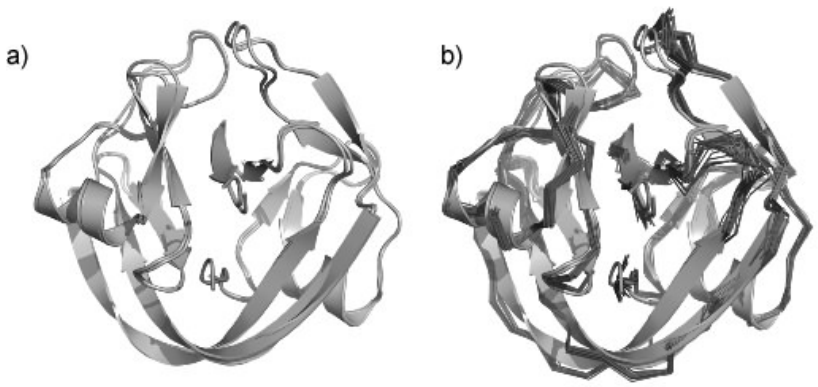

Figure 12 *Crystal structure of* Npu*DnaE intein. a) Two models from the asymmetric unit superimposed. b) Superimposed crystal structure models with the 20 NMR model coloured in light and dark, respectively.*

It was additionally observed that Gly132 showed conformational exchange and the amide signal of Asn131 was unobserved in the [^{1}H-^{15}N]-HSQC spectrum. This observation guided to engineer a novel split intein by moving the split site to after residue Asn131. Thus, the newly engineered C-intein consists of only six amino acid residues and this is the shortest split intein there has been described in the literature. The new split intein was tested *in vivo* and *in vitro* and it showed good protein splicing ability. Though, in the *in vitro* experiment C-cleavage was observed, which indicates experimental conditions might require further optimization to fully harness the potential of this short split intein.

At the same time a split site was engineered in gyrase B subunit intein from *Synechocystis sp.* strain PCC6803 intein where the C-intein had a similar size of six residues as in *Npu*DnaE intein (Appleby *et al.*, 2009). An engineered split intein that only consists of six amino acids would be very applicable for site-specific modification of proteins because of the short residue length. The requirement of maximum six residues for the C-intein makes it suitable for chemical synthesis at a reasonable cost compared to the wild type split intein where the C-intein is 35 amino acid residues or the previous engineered intein with a C-intein of 14 residues (Aranko *et al.*, 2009).

Table 3. *Crystallographic data collection and structure refinement of* Npu*DnaE intein.*

Data collection	*Npu*DnaE intein
Beamline	ESRF ID23-1
Space group	$P2_12_12_1$
Molecules per asymmetric unit	2
Cell dimensions	
a, *b*, *c* (Å)	57.55, 66.70, 67.48
α, β, γ (°)	90, 90, 90
Resolution (Å)[a]	50-1.72 (1.82-1.72)
R_{merge}(%)[b]	6.5 (59.1)
$I/\sigma I$	21.13 (3.67)
Completeness (%)	99.2 (96.3)
Redundancy	7.04 (7.06)
Refinement	
Resolution (Å)	1.72
No. reflections	28,252
R_{work}/ R_{free}[c]	0.169/0.209
Wilson B-factor	20.7
No. atoms	
Protein	2310
Ligand/ion	32
Water	217
R.M.S. deviations	
Bond lengths (Å)	0.010
Bond angles (º)	1.252
Ramachandran plot statistics [%]	
Residues in favoured region	97.8
Residues in allowed region	1.8
Residues in outlier region	0.4
PDB code	4KL5

[a] Values are from the highest resolution shell. [b] $R_{merge}= \Sigma_h\Sigma_i |I_i-\langle I\rangle| / \Sigma_h\Sigma_i I_i$ where I_i is the observed intensity of the *i*th measurement of reflection h, and $\langle I\rangle$ is the average intensity of that reflection obtained from multiple observations. [c] $R = \Sigma||F_0| - |F_c|| / \Sigma|F_0|$ where F_o and F_c are the observed and calculated structure factors, respectively, calculated for all data.

5.3 *Pho*RadA Intein Solution NMR Structure

Studies have shown that *Pho*RadA intein in comparison with other inteins has high protein splicing efficiency, which makes *Pho*RadA intein an attractive candidate for biotechnological applications (Ellilä *et al.*, 2011). *Pho*RadA intein was selected as

candidate for structural studies to investigate the structural features of the intein because of the high protein splicing efficiency. To utilize intein technologies it is important to understand how the extein sequences near the protein splicing junction influence protein splicing efficiency. In this study it was focused on how the extein sequences flanking the protein splicing junction can influence the protein splicing efficiency. This was performed by producing a construct of *Pho*RadA intein with a C1A and a T+1A mutation of the two residues C-extein to prevent any reaction from occurring.

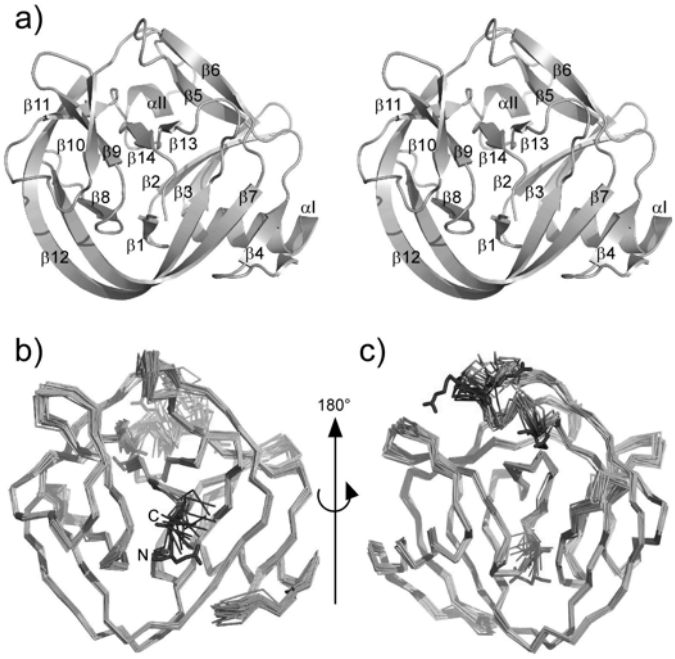

Figure 13 *Solution NMR structure of* Pho*RadA. a) Stereo view of schematic representation of* Pho*RadA structure closest to mean. Secondary structure elements are labelled. b) Assembly of 20 NMR structures. N- and C-terminus are indicated. Side chain heavy atoms are shown for the model closest to mean in dark grey for residue lacking resonance assignment. c) Same as in b) turned 180° along axes as indicated.*

The structure of *Pho*RadA intein was initially investigated using NMR spectroscopy and a [^{1}H-^{15}N]-HSQC spectrum of *Pho*RadA intein showed that the signals were well dispersed between 5.8-10 ppm, which indicated the protein was folded. Specific resonance assignment of *Pho*RadA intein was achieved using a series of standard triple resonance spectra where 94% of H^N, Hα, N, CA, and CO atoms and 97% of side chain atoms were assigned. The NMR structure was determined based upon NOE assignment from ^{15}N- and ^{13}C-NOESY-HSQC spectra and dihedral backbone angle prediction from TALOS+. Statistics from the structure determination are listed in Table 4.

The structure of *Pho*RadA intein resembles a horseshoe disk-shape as seen by other intein structures (see Figure 13). The backbone RMSD is 0.61 ± 0.24 Å indicating less defined structure than the structure determined of *Npu*DnaE intein (Study II). However, a flexible region at residues 121-133 give rise to increased RMSD. Excluding the flexible region *Pho*RadA intein backbone RMSD is 0.44 ± 0.06 Å. The flexibility in the model

originates from lacking resonance assignment that consequently have less NOE constrains assigned. Missing resonance assignment is likely due to broadening of signals caused by conformational exchange. The C-extein shows high conformational flexibility and no clear interaction can be seen between the intein and extein. A similar flexibility is observed in the NMR structure of *Mja*KlbA intein (Johnson *et al.*, 2007) where no NOEs could be assigned between the extein sequences and the intein.

Table 4. *Experimental data and statistic for the NMR structure calculation of* Pho*RadA intein.*

Quantity	*Pho*RadA intein
NOE upper distance limits	3515
Short range NOE (i-j≤1)	1654
Medium-range NOE (1<i-j<5)	425
Long-range NOE (i-j≥5)	1436
Dihedral angels [a]	278
Residual CYANA target function	1.14±0.08
Residual NOE violation	
Number ≥ 0.3 Å	3±2
Maximum [Å]	0.67±0.24
Residual dihedral angle violation	
Number≥5°	6±2
Maximum (°)	12.28±2.96
Amber energies [kcal/mol]	
Total	-6,518±24
van der Waals	-1,402±15
Electrostatic	-11,520±330
RMSD from ideal geometry	
Bond length [Å]	0.0229±0.0001
Bond angels [°]	2.120±0.015
RMSD to mean coordinate	
Backbone 1-172 [Å]	0.61±0.10
Heavy atoms 1-172 [Å]	1.09±0.16
Ramachandran plot statistics [%][b]	
Most favored regions	91.3
Additional allowed regions	8.2
Generously allowed regions	0.5
Disallowed regions	0.0
BMRB entry	18320
PDB code	2LMQ

[a] TALOS+ derived angle predictions (Shen *et al.*, 2009); [b] Derived by PROCHECK-NMR (Laskowski *et al.*, 1996).

5.4 *Pho*RadA Intein and *Pho*RadA$_{min}$ Intein Crystal Structures

The crystal structure of *Pho*RadA intein was determined to 1.75 Å resolution with two polypeptide chains in the asymmetric unit. Statistics from the structure refinement are listed in Table 5. Chain A was modelled with 172 amino acids where loops at residues 13-14, 55-57, and 125-129 showed weak electron density. Chain B was modelled with 169 residues where the loop at residue 12-15 showed weak electron density and the loop at residue 127-129 was not included in the model because of completely missing electron density. The molecules in the asymmetric unit are very similar with a backbone RMSD of 0.254 Å superimposed over 138 residues (see Figure 14). Comparison of the crystal structure chain A and B with a mean NMR structure gives a RMSD of 0.905 Å for 142 superimposed residues and 0.808 Å for 135 superimposed residues, respectively.

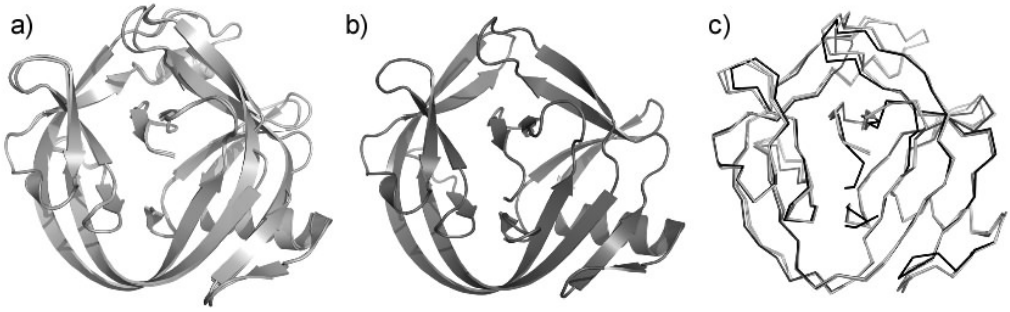

Figure 14 Crystal structure of Pho*RadA intein and* Pho*RadA$_{min}$ intein a) Schematic representation of superimposed* Pho*RadA intein crystal structure chain A and B from the asymmetric unit. b)* Pho*RadA$_{min}$ intein crystal structure. c) Superimposed crystal structure models of* Pho*RadA intein (light grey) and* Pho*RadA$_{min}$ intein (dark grey).*

The electron density of the C-extein was very poor in both chains, indicating flexibility, and was not modelled in the structure. The largest deviation between the NMR and crystal structure is in the loop at residue 120-133. The NMR structure and X-ray structure supports that the loop is flexible. The flexible region could be suitable for introducing split sites in *Pho*RadA intein considering the previous analysis of *Npu*DnaE intein structure, but this was not the aim of this study.

The flexible region at residue 120-133 led to the design of minimized *Pho*RadA intein by genetically removing residues 121-130 to generate *Pho*RadA$_{min}$ intein. The minimized version of *Pho*RadA intein was tested for *cis* protein splicing activity and it was shown that *Pho*RadA$_{min}$ intein has similar protein splicing efficiency as the wild type sequence. The crystal structure of *Pho*RadA$_{min}$ intein was determined to 1.58 Å resolution with one molecule in the asymmetric unit (see Figure 14) (see Table 5 for structure refinement statistics). The flanking residues of the deleted loop (residue 120, 131-133) became ordered in the crystal structure and all where well defined. For the minimization of *Pho*RadA intein a K131N mutation was introduced to accommodate a turn and the residue showed clear electron density. The RMSD between *Pho*RadA$_{min}$ intein and *Pho*RadA intein chain A and B is 0.493 Å and 0.541 Å for 130 and 133 superimposed residues, respectively. The major difference between *Pho*RadA intein and *Pho*RadA$_{min}$ intein is

both the N- and C-extein sequences are well defined in *Pho*RadA$_{min}$ intein. The four N-extein and two C-extein residues have a clear electron density and the orientations of the residues are clearly defined in the structure.

Table 5. *Crystallographic data collection and structure refinement of* Pho*RadA intein and* Pho*RadA$_{min}$ intein.*

Data collection	*Pho*RadA intein	*Pho*RadA$_{min}$ intein
Beamline	ESRF ID23-1	ERSF ID14-1
Space group	$P2_12_12_1$	$P2_12_12_1$
Molecules per asymmetric unit	2	1
Cell dimensions		
a, *b*, *c* (Å)	58.1, 67.4, 82.9	46.9, 63.6, 66.6
α, β, γ (°)	90, 90, 90	90, 90, 90
Resolution (Å)[a]	33.7-1.75 (1.80-1.75)	46.0-1.58 (1.64-1-58)
R_{merge}(%)[b]	7.0 (95.3)	8.6 (45.4)
$I/\sigma I$	17.3 (2.5)	12.2 (2.4)
Completeness (%)	99.3 (99.8)	98.9 (89.6)
Redundancy	7.1 (7.2)	8.6 (7.9)
Refinement		
Resolution (Å)	33.7-1.75	46.0-1.58
No. reflections	33,296	27,680
R_{work}/ R_{free}[c]	0.190/0.234	0.167/0.195
Wilson B-factor	25.5	17.2
No. atoms		
Protein	2788	1366
Ligand/ion	39	0
Water	121	251
R.M.S. deviations		
Bond lengths (Å)	0.012	0.011
Bond angles (°)	1.65	1.4
Ramachandran plot statistics [%]		
Residues in favoured region	96.7	97.6
Residues in allowed region	3.3	2.4
Residues in outlier region	0.0	0.0
PDB code	4E2T	4E2U

[a] Values are from the highest resolution shell. [b] $R_{merge}= \Sigma_h\Sigma_i |I_i-\langle I\rangle| / \Sigma_h\Sigma_i I_i$ where I_i is the observed intensity of the *i*th measurement of reflection h, and $\langle I\rangle$ is the average intensity of that reflection obtained from multiple observations. [c] $R = \Sigma||F_0| - |F_c|| / \Sigma|F_0|$ where F_o and F_c are the observed and calculated structure factors, respectively, calculated for all data.

A Dali Server search of *Pho*RadA intein showed the most similar inteins are two archaeal inteins, *Pfu*RIR1-1 intein (Ichiyanagi *et al.*, 2000) and DNA polymerase II intein from *Thermococcus kodakaraensis* (*Tko*PolII intein) (Matsumura *et al.*, 2006) with a RMSD of

1.6 and 2.0, respectively (Holm and Rosenström, 2010). *Mja*KlbA intein (Johnson *et al.*, 2007a), *Pab*PolII intein (Du *et al.*, 2011a), and *Mtu*RecA (Van Roay *et al.*, 2007) have the next three highest Z-scores but all have a RMSD above 2.0. Thus, the structure of *Pho*RadA intein resembles other intein structures.

5.5 Active Site of *Npu*DnaE Intein and *Pho*RadA Intein

The structure of *Pho*RadA$_{min}$ intein was determined as an inactive protein splicing precursor where the structure contains four N-extein and two C-extein residues. The residues conformation is well defined with excellent electron density for both extein sequences. The -1 scissile bond between Met-1 and Ala1 is a *trans* conformation rather than the unusual *cis* conformation seen in *Mxe*GyrA intein (Klabunde et al., 1998). The usual *trans* conformation is similar to what have been observed in *Sce*VMA intein, *Ssp*DnaE intein, and *Ssp*DnaB intein (Sun *et al.*, 2005; Ding *et al.*, 2003; Poland *et al.*, 2000; Mizutani *et al.*, 2002).

The crystal structure of *Npu*DnaE intein was determined as inactive protein splicing precursor where the structure contains three N-extein and four C-extein residues. The conformation of the extein residues is poorly defined and the electron density is missing for part of the extein sequences. It was not possible to model the first and the two first N-extein residue in chain A and B, respectively. The first C-extein residue could be modelled in chain A while it was possible to model the two first C-extein residues in chain B. A *trans* conformation was observed in both chain A and B of the -1 scissile bond between Gly-1 and Ala1 as observed in other protein splicing precursors.

In the structure of *Sce*VMA intein (PDB: 1JVA) the C1S Oγ atom and the Gly-1 carbonyl oxygen is 3.2 Å apart. The close distance is important for the initial step in protein splicing because it has been suggested that by a nucleophile attack by the first intein residue forms a five-membered ring thiazolidine intermideate during the N-S(O)-acyl shift (Mizutani et al., 2002). The presence of a thiazolidine intermediate was identified in a mutant of *Ssp*DnaB intein (Ludwig et al., 2008). The same distance in *Pho*RadA$_{min}$ intein cannot be measured because of the C1A mutation, but when modelling a Cys1 in the structure model the distance between Cys1 Sγ and the -1 carbonyl oxygen could be as small as 3.0 Å. Thus, *Pho*RadA$_{min}$ intein could be in a conformation needed just before the initial acyl shift occurs. Additionally, the τ (N-Cα-C) angel of Ala1 is 116.5°, which indicates a slightly distorted angle compared to ideal 110°. In other protein splicing precursor structures the -1 or +1 τ angel has been reported as distorted which has been suggested to be important for breaking the peptide bonds at the protein splicing junction. The disordered bond of Ala1 would be unfavourable and could be a driving force for the initial N-S acyl shift and thio(ester) formation.

The distance between the Ala+1 Cβ and the Met-1 C' is 3.6 Å in *Pho*RadA$_{min}$ intein (see Figure 15). The close proximity of the atoms would be needed for the second step in the protein splicing reaction to occur without the need for large conformational change. The distance is very different from the structures of *Ssp*DnaE intein (PDB: 1ZDE), *Ssp*DnaB intein (PDB: 1MI8), and *Sce*VMA intein (1EF0) where the distance between the

+1 Cβ and the -1 C' atoms are 7.8 Å. In *Npu*DnaE intein the distance between Ala+1 Cβ and Gly-1 C' is 12.1±0.3 Å and *Npu*DnaE intein resembles an open conformation as seen in *Ssp*DnaE intein, *Ssp*DnaB intein, and *Sce*VMA intein. It is possible that *Pho*RadA$_{min}$ intein has been trapped in a closed conformation where the three previously mentioned structures are in an open conformation. The structures of *Ssp*DnaB intein (PDB: 1ZDE) and *Sce*VMA intein (PDB: 1EF0) have the Cys+1 residue coordinated by a zinc ion. It is known form the literature that zinc-ions inhibits protein splicing of some inteins (Mills and Paulus, 2001) and by trapping the intein in an open conformation could be the mechanism of zinc inhibition. The difference in the open and closed conformation is seen with a very different Ψ angle of the first intein residue. In *Pho*RadA$_{min}$ intein Ala1 Ψ angle is ÷35° while the same angle is 159° in *Ssp*DnaE (PDB: 1ZDE), 151° in *Ssp*DnaB (PDB: 1MI8), 127±7° in *Sce*VMA (PDB: 1JVA), and 173.6 (Chain A) and 124.7 (Chain B) in *Sce*VMA (PDB: 1EF0).

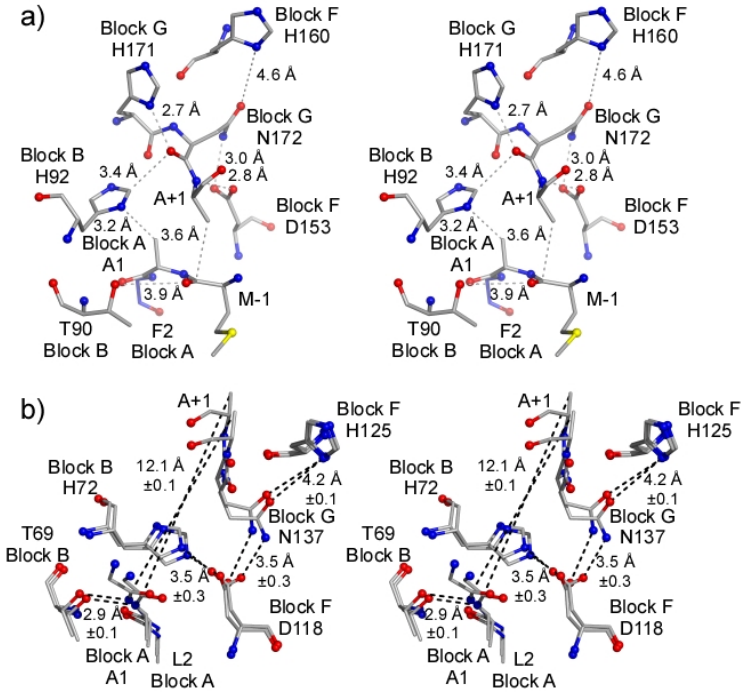

Figure 15 *Stereo view of residues in the vicinity of the active site of a)* Pho*RadA$_{min}$ intein and b)* Npu*DnaE intein. Distances between atoms of conserved residues are indicated with broken lines. For* Npu*DnaE a mean value and deviation of the distance in the two molecules in the asymmetric unit are shown. N, O, and S atoms are indicated with blue, red, and yellow spheres, respectively. Residue numbers are labelled with the locations in the conserved intein motifs (blocks A, B, F, and G). Only backbone atoms of Phe2 in* Pho*RadA$_{min}$ intein and Leu2 in* Npu*DnaE intein are shown.*

In the structure of *Pho*RadA$_{min}$ intein the block F Asp153 side chain is in close interaction with Asn172 and Ala+1, where hydrogen bonding between Asp153 and Asn172 prevents Asn cyclization from occurring. The conformation of Asp153 indicates why a mutation of the residue in some intein yields C-cleavage product formation (van Roey *et al.*, 2007; Ramirez *et al.*, 2013). The conformation of Asp153 may be important for controlling the protein splicing steps and would require some structural rearrangements for Asn

cyclization to proceed. In the structure of *Npu*DnaE intein block F Asp118 is also interacting with Asn137 but not with the +1 residue. A similar interaction has not been observed in other protein-splicing precursors because the last Asn has been mutated to an Ala or Ser (see Table 1). The intein structures likely represent a snapshot of one of several conformations that an intein go through in the protein splicing mechanism.

5.6 *Pho*RadA Intein -1 Residue Mutation Analysis

Inteins used in foreign context can become insufficient. It is an important to understand how the flanking extein residues influence the protein splicing efficiency. The wild type sequence of the *Pho*RadA intein contains a Lys at the -1 position and *Pho*RadA intein -1 residue dependency was tested using a model system (see Figure 16). The -1 residue position was changed to 19 different residues each with one of the 20 common amino acids.

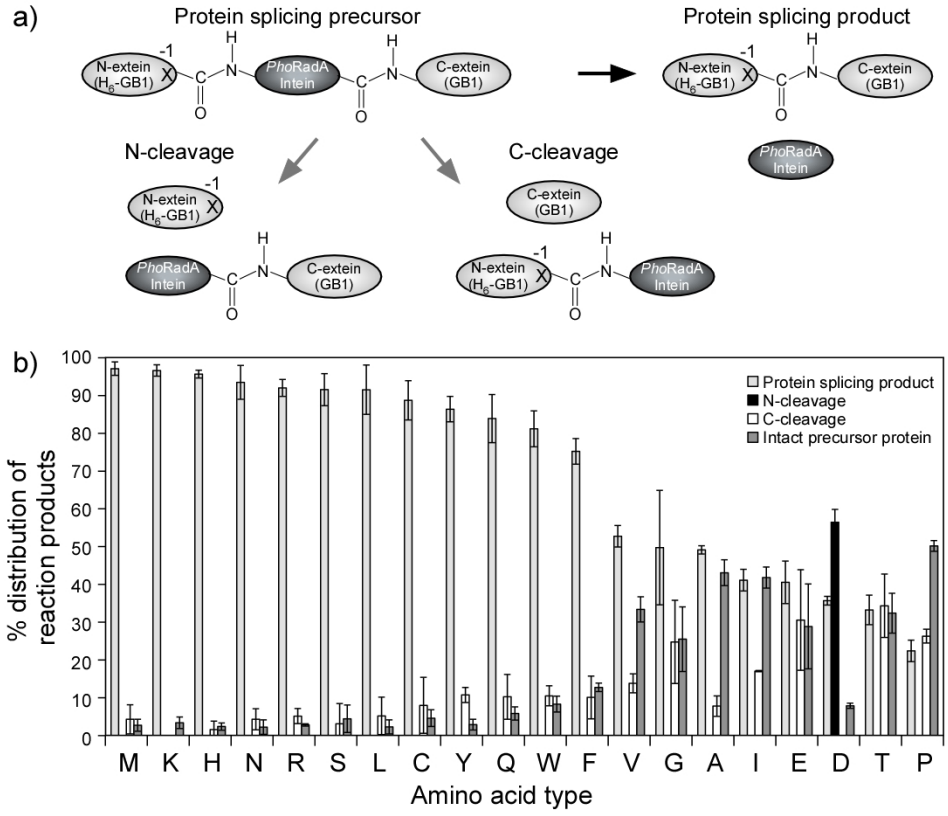

Figure 16 *Protein splicing* by Pho*RadA intein. a) Schematic drawing of the used model system for testing* Pho*RadA intein protein splicing efficiency. Black and grey arrows indicates product and side-product, respectively. X: -1 residue position tested for 20 amino acid substitution. GB1: the B1 domain of IgG binding protein G b) Protein splicing efficiency of* Pho*RadA intein -1 residue mutants. The -1 residues are sorted after decreasing protein* cis*-splicing efficiency of* Pho*RadA intein. The amount of product, N- and C-cleavage and intact precursor are shown with colour representation in upper right corner.*

The protein splicing efficiency of the 20 variants were characterized by protein expression of a model system containing H_6-GB1-*Pho*RadA intein-GB1 in *E. coli*. The protein splicing efficiency was quantified by the amount of protein spliced product (H_6-GB1-GB1), N- (H_6-GB1) and C-cleavage (H_6-GB1-*Pho*RadA intein), and intact protein splicing precursor protein (H_6-GB1-*Pho*RadA intein-GB1) present after protein expression (see Figure 16a).

The mutation studies showed *Pho*RadA intein has a high sequence tolerance at the -1 residue position (see Figure 16b). *Pho*RadA intein retained high protein splicing efficiency (>90%) for Met, Lys, His, Asn, Arg, Ser, and Leu at the -1 residue position. The protein spicing efficiency for Cys, Tyr, Gln, Trp, and Phe are less proficient in protein splicing (75-90% protein splicing efficiency), but are more substantial in protein splicing than Val, Gly, Ala, Ile, Glu, Asp, Thr, and Pro which show between 20-55% protein splicing efficiency.

Pro at -1 position has the lowest protein splicing efficiency, which is not surprising because of different structure of the scissile peptide bond compared to other amino acids. Gly and Ala has low protein splicing efficiency which could be explained by that Gly and Ala are the two smallest amino acids and they might be unable to form favourable van der Walls interaction needed for the protein splicing reaction to precede. A significant difference in protein splicing efficiency is seen with Leu that has a high protein splicing efficiency (>90%), which is significantly better than Ile and Val that have less than 55% protein splicing efficiency. All three residues are hydrophobic and have similar side chain structure. Similar is observed with a Ser at the -1 residue position which has high protein splicing efficiency but Thr at the -1 residue position has poor protein splicing efficiency despite both residues contain a hydroxyl group and have similar chemical properties. However, the amino acid residues Thr, Ile, and Val have a similarity in structure by being ß-branched. The ß-branched structure could perhaps form a steric hindrance that could prevent a conformational change from occurring. The ß-branched residues have C-cleavage as bi-product, which indicates Asn cyclization has occurred while no N-cleavage is observed. This could indicate the initial thio(ester) was unable to form or N-cleavage products should have been observed. Additionally, it is seen Gly, Ala, Ile, Thr, and Val at -1 residue position have 25-41% intact precursor left after ended experiment and the *Pho*RadA intein is perhaps unable to form a closed conformation seen in the crystal structure. Asp and Glu at the -1 position have poor protein splicing efficiency. Both side chains are negatively charge, which is the opposite of the wild type residue in *Pho*RadA intein. It is likely that the charge of the side chain is important to form proper interactions.

5.6.1 *Pho*RadA Intein N-extein Interaction

The structure of *Pho*$RadA_{min}$ intein resembles an unseen conformation of an intein protein splicing precursor. A clear interaction is seen between Met-1 and the intein structure. Intein residues that are within 4 Å of Met-1 include Ala1, Tyr69, Glu71, Val73, Thr90, and Val151. Additionally, the extein residues Gln-3, His-2, and Ala+1 are within 4 Å of Met-1. Thr90 is a highly conserved block B residue that is believed to be involved in

catalysing the initial protein splicing step and coordinating the protein splicing reaction. Val151 is not a conserved residue but is located in the block F sequence motif. The residues Tyr69, Glu71, and Val73 are located between the sequence motifs of block A and B and are non-conserved residues. The wild type sequence of *Pho*RadA intein contains a Lys-1 that could be expected to have a similar orientation as Met-1 in *Pho*RadA$_{min}$ intein because Met and Lys both have long hydrophobic side chains. Lys differs from Met by containing an ε-amino group, which in physiological conditions is positively charge.

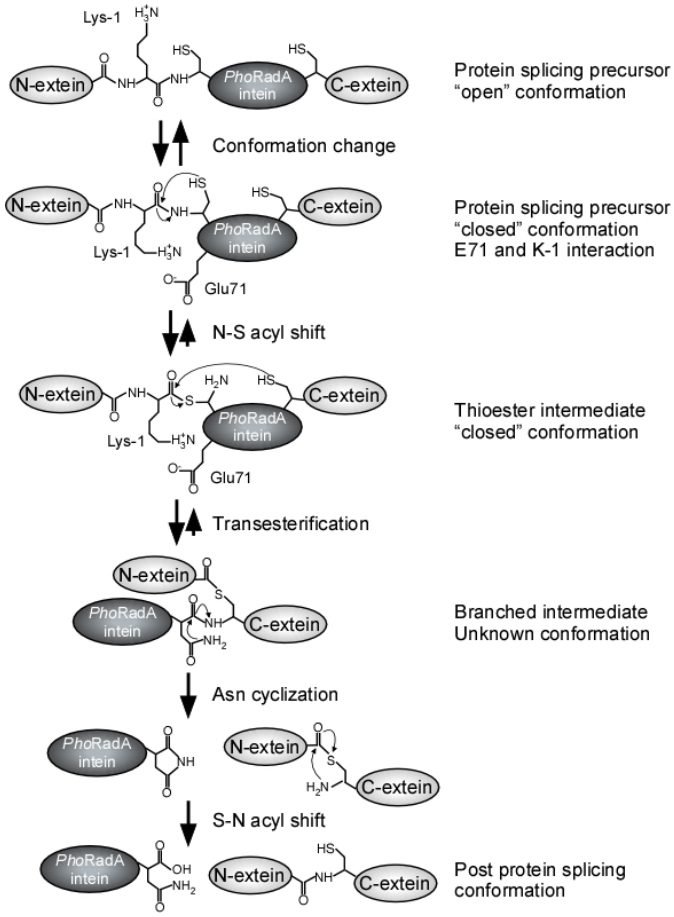

Figure 17 *Possible reaction mechanism scheme of protein splicing reaction* by Pho*RadA intein.*

In the previously described analysis where the -1 residue position is varied showed that *Pho*RadA intein has poor protein splicing efficiency with Asp and Glu at the -1 residue position. The structure of *Pho*RadA$_{min}$ intein shows why the negatively charged residues at -1 position have poor protein splicing efficiency. Met at the -1 residue position is interacting with Glu71 and a Glu or Asp at the -1 residue position would introduce unfavourable electrostatic interaction. Consequently, protein splicing by *Pho*RadA intein becomes insufficient which likely is because the intein is unable to make the conformer seen in the structure of *Pho*RadA$_{min}$ intein. To verify that the observed interaction of *Pho*RadA intein Glu71 is true and the interaction is needed for protein splicing reaction the negatively charged Glu71 was mutated to a Thr.

The protein splicing efficiency of *Pho*RadA intein with an E71T mutation was analysed with an Asp, Glu, and Lys at the -1 residue position. The same model system as previous was used for the analysis. The protein splicing efficiency was recovered with Glu-1 to more than 90% with the *Pho*RadA intein E71T mutation. Lys at the -1 residue position in *Pho*RadA intein E71T mutant still showed high protein splicing efficiency. Thus, it seems likely Glu71 would be interacting with the -1 residue during the reaction step and the interaction occurs in a "closed" conformation (see Figure 17). However, the protein splicing efficiency did not increase with an Asp at the -1 residue position whereas the amount of N-cleavage reaction increased. N-cleavage reaction with an Asp at the -1 position has previously been reported for *Mxe*GyrA intein and *Ssp*DnaE intein (Southwood *et al.*, 1999; Amitai et al., 2009). It is known that an aspartic residue can hydrolyze a peptide chain at the C-terminal side under acidic conditions (Inglis, 1983). The hydrolysis reaction occurs by a nucleophilic attack by the carboxyl group on the side chain carbonyl group. A similar mechanism could occur with an Asp at the -1 residue position which during the initial N-S(O) acyl shift would produce a cleavage reaction. There are several intein with an Asp at the -1 residue position in the InBase database (Perler, 2002) but they might have a different mechanism to prevent the cleavage reaction.

5.6.2 N-Extein Interaction in Inteins

Inteins have different mechanisms to perform protein splicing and inteins have been divided into three different classes (Tori *et al.*, 2010) (see section 2.2). It is plausible that even class I inteins have more than one way to coordinate the protein splicing mechanism. Mutational studies of inteins have indicated that residues involved in the protein splicing are intein specific and it can therefore not be generalised how the reaction is catalysed. Thus, the N-extein-intein interaction identified in this study could be a unique interaction that only occurring in *Pho*RadA intein and other closely related inteins. Sequence identity of inteins is low which makes structure prediction difficult. However, several intein structures derived from different organism and host genes have been determined (see Table 1) and structural comparison of these and *Pho*RadA intein is possible. There has been determined 11 different intein structures including the intein structures determined in this thesis. If a similar extein-intein interaction, as seen in *Pho*RadA intein, would occur in other inteins it should be possible identify interactions between the extein and intein.

In the structure of *Pho*RadA$_{min}$ intein the residues that were located within 4 Å of Met-1 included Gln-2, His-2, Ala1, Tyr69, Glu71, Val73, Thr90, Val151, and Ala+1. Residues located at the same position in other inteins are identified by superimposition the intein structures with *Pho*RadA$_{min}$ intein and are listed in Table 6. Not all structures contain the wild type sequence or extein residues but the structure comparison is based upon the wild type sequences. The inteins listed in Table 6 are listed according to the amino acid at the -1 position.

*Pho*RadA intein and *Mtu*RecA intein have a Lys-1 residue and *Pho*RadA intein Glu71 is involved in intein-extein interaction (see Figure 17). There is a clear favourable electrostatic interaction between Lys-1 and Glu71 in *Pho*RadA intein but the same

interaction is not visible in *Mtu*RecA intein. *Mtu*RecA intein has Gln51 located at the corresponding position of Glu71 in *Pho*RadA intein. However, similar interaction could be present because *Pho*RadA intein with an E71T mutation and Lys at the -1 residue position showed high protein splicing efficiency, which is comparable with the wild type sequence. The side chain of *Mtu*RecA intein Gln51 could make similar interaction as the E71T mutation in *Pho*RadA intein. However, Arg54 is located at the active site of *Mtu*RecA intein and would introduce unfavourable interactions but the long flexible side chain of an Arg has conformational flexibility that makes different conformations possible.

*Npu*DnaE intein, *Ssp*DnaE intein, and *Mxe*GyrA intein contains a Tyr at the -1 residue position. *Npu*DnaE intein and *Mxe*GyrA intein structures have been determined with a Gly-1 and Ala-1 residue, respectively (Study II; Klabunde *et al.*, 1998), but no clear interaction is observed in the structures. The structure of *Ssp*DnaE intein has been solved as a protein splicing precursor, but the structure resembled an open conformation with a long distance between the N- and C-extein sequences (Sun *et al.*, 2005). Thus, the interaction of Tyr-1 in *Ssp*DnaE intein is not similar to Met-1 in *Pho*RadA$_{min}$ intein. However, protein splicing in *Ssp*DnaE intein requires a conformational change. If *Mxe*GyrA intein, *Ssp*DnaE intein, and *Npu*DnaE intein would have a closed conformation similar to the conformation seen in *Pho*RadA$_{min}$ intein all three inteins have aromatic side chain among the non-conserved residues that could interact with the Tyr-1 residue. Additionally, it is seen *Npu*DnaE intein and *Ssp*DnaE intein has a Glu-2 residue and both inteins have an Arg residue where Glu71 is located in *Pho*RadA intein. The positively charged Arg could interact with the -2 extein residue.

Table 6. *Residues within 4 Å of Met-1 in* Pho*RadA$_{min}$ intein structure. Residues located at the same position in other known intein structures have been identified by superposition of the structure models. The wild type sequence numbering and amino acid type is listed.*

Intein	N-extein			Block A	Non-conserved residues			Block B	Block F	C-extein
*Pho*RadA	S-3	G-2	K-1	C1	Y69	E71	V73	T90	V151	T+1
*Mtu*RecA	K-3	N-2	K-1	C1	F49	Q51	R54	T70	T420	C+1
*Npu*DnaE	A-3	E-2	Y-1	C1	H48	R50	Q53	T69	V116	C+1
*Ssp*DnaE	A-3	E-2	Y-1	C1	H48	R50	Q53	T69	I138	C+1
*Mxe*GyrA	M-3	R-2	Y-1	C1	F51	S53	H56	T72	V177	T+1
*Ssp*DnaB	E-3	S-2	G-1	C1	F49	T51	K54	T70	V409	S+1
*Sce*VMA	Y-3	V-2	G-1	C1	P41	G43	E45	N76	Y431	C+1
*Mja*KlbA	H-3	D-2	G-1	A1	W71	K73	Y75	T93	I145	C+1
*Pfu*RIR1-1	G-3	G-2	G-1	C1	W72	Y74	L76	S96	F431	T+1
*Pab*PolII	R-2	R-2	N-1	C1	I68	A70	T73	T90	V164	C+1
*Tko*PolII	L-3	A-2	N-1	S1	I71	H73	Y75	T93	V515	S+1

Four determined intein structures have a Gly-1 extein residue in the wild type sequence. *Pho*RadA intein with a Gly at the -1 residue position showed inefficient protein splicing, which likely is because of too weak van der Walls interactions. A structural comparison of *Pho*RadA intein with the structures of *Sce*VMA intein, *Mja*KlbA intein, *Ssp*DnaB intein, and *Pfu*RIR1-1 intein should show residues that would compensate for the weak or

missing van der Walls interaction. *Pho*RadA intein contains a Block F Val151 whereas *Sce*VMA intein, *Mja*KlbA intein, and *Pfu*RIR1-1 intein contain a Tyr, Ile, and Phe, respectively. Additionally, *Sce*VMA intein contains a Block B Asn instead of the more commonly found Block B Thr. All these larger residues could compensate for Gly not having any side chain by making the active site smaller. *Sce*VMA intein has been characterized by changing the -1 residue into the 20 common amino acids (Chong *et al.*, 1998) and it was shown that the *Sce*VMA intein has high sequence tolerance, even to Asp and Glu. *Sce*VMA intein contains a Glu45 and it would have been expected that *Sce*VMA intein would have less efficient protein splicing with an Asp or Glu at the -1 residue position. The active site may be flexible in size since *Sce*VMA intein can compensate more bulky residues (Tyr, Phe, Trp, Lys, Arg, Met, Gln) (Chong *et al.*, 1998). If residues could adopt a different orientation of side chains the active site could accommodate larger amino acids at the -1 residue position. *Ssp*DnaB intein has the same Block F Val residue as *Pho*RadA intein and the non-conserved residues in *Ssp*DnaB do not seem to provide the proper suggested interaction. Only by the Glu-3 residue and Lys54 favourable interactions seems visible.

*Pab*PolII intein and *Tko*PolII intein both have Asn at the -1 residue position. Asn could fulfil hydrogen bonding with the hydroxyl group of Tyr75 and Thr73 found in *Pab*PolII intein and *Tko*PolII intein, respectively. In general, it seems that the interactions seen in *Pho*RadA intein could occur in other inteins where no clear unfavourable interaction can be seen. However, it is obvious from the literature that inteins have different reaction mechanisms and mutation studies have shown that residues involved in the protein splicing mechanisms seem to be intein specific. Therefore the extein-intein interactions seen in *Pho*RadA intein can be intein specific. It would be of interest to explore this interaction further for engineering of inteins to accommodate foreign extein sequences. Inteins could then be designed with great care to function in desired context.

6 Conclusion and Future Perspective

Inteins are interesting macromolecules with an unusual function that does not seem to provide any benefit for its host. Intein mediated protein splicing is a complex reaction coordinated by the intein molecule. From the literature it is obvious that inteins are highly diverse molecules that have evolved several different mechanisms to perform the same chemical reaction. Inteins have been found useful by exploiting the intein mechanism for many biotechnological applications. An important aspect of intein technology was the discovery that they could be used in foreign context *in vivo* and *in vitro* and inteins could be split to perform *trans* protein splicing. However, intein applications are hampered by poor knowledge of how extein sequences influence the protein splicing efficiency and it is not clear how the inteins could be optimised for their application usage.

In this thesis the structure of *Npu*DnaE intein was determined using NMR spectroscopy and X-ray crystallography and it was shown that the structure has a HINT fold as other inteins. The dynamic features of *Npu*DnaE intein were analysed using nuclear spin relaxation measurements. Measurements of longitudinal and transverse relaxation rates were shown to be a very powerful tool together with the NMR structure to identify and design sites suitable for new split site in *Npu*DnaE intein. The NMR data identified possible split sites where local conformational exchange occurs. From this data a novel split site was identified with split C-intein part consisting of six residues. This is among the smallest split intein fragments described in literature.

The structure of *Pho*RadA intein was determined using NMR spectroscopy and X-ray crystallization. The two techniques gave similar results where the RMSD between the crystal structure and a mean NMR structure was less than 1 Å. The structural fold of *Pho*RadA intein is similar to other inteins with a backbone conformation closest to the endonuclease containing *Pfu*RIR1-1 intein (Ichiyanagi *et al.*, 2000) and *Tko*PolII intein (Matsumura *et al.*, 2006). The NMR and crystal structures confirmed that the region near residues 120-133 is flexible. The residues 121-130 could be deleted without loss of protein splicing function to generate a minimized *Pho*RadA intein. The deleted loop is likely a remnant of an endonuclease domain that has been lost during evolution.

The structure of *Pho*RadA$_{min}$ intein was determined as an inactive protein splicing precursor using X-ray crystallography to 1.58 Å resolution. *Pho*RadA$_{min}$ intein crystal structure resembles an unseen conformation of a protein splicing precursor. The conformation of *Pho*RadA$_{min}$ intein is a “closed” conformation with the N- and C-protein splicing junction in close proximity. The crystal structure of *Npu*DnaE intein was determined to 1.72 Å resolution and resembled an “open” conformation. In the structure model of *Pho*RadA$_{min}$ intein the Met-1 residue interacts with the intein residues Tyr69, Glu71, Val73, Thr90, and Val153. The interaction was confirmed in an E71T mutation of *Pho*RadA intein. The mutation removed unfavourable electrostatic interactions between Glu -1 residue and Glu71 in *Pho*RadA intein. The *Pho*RadA intein E71T mutant became efficient in protein splicing with a Glu-1 residue as opposite to the wild type intein sequence. The conformation observed with *Pho*RadA$_{min}$ intein likely resembles one of many conformations inteins might go through in the protein splicing reaction steps.

The structure of the *Pho*RadA intein was compared to other inteins with known structures. The structures were compared to identify if the intein-extein interactions observed in *Pho*RadA$_{min}$ intein exist in other inteins. The structural comparison was inconclusive and the intein-extein seen in *Pho*RadA$_{min}$ intein cannot be concluded as general interaction for other inteins. However, similar intein-extein interactions cannot be excluded either. It would require further experiments to verify a similar extein-intein interaction in other inteins. A systematic analysis of the -1 residues position on the protein splicing dependency in other inteins could provide a further insight. The intein-extein interaction seen in *Pho*RadA intein is likely similar among closely related inteins. The comparison of intein structures provides a possible insight how *Pho*RadA intein could be engineered to compensate a Gly or Ala at the -1 residue position. Inteins with a Gly-1 seem to have more bulky residues in the active site and *Pho*RadA intein active site could be engineered smaller by introducing a T90N, V73L, V73Y, V151I, or V151Y as these residues are present in inteins that contains a Gly-1 residue (see Table 6).

Unlike the -1 residue in the *Pho*RadA$_{min}$ intein no interactions of the +2 residue was observed. Though, it is known the +2 and +3 residues significantly can influence protein splicing efficiency (Iwai et al., 2006; Amital *et al.*, 2009). A further systematically analyses of residues in the protein junction would provide further insight how the C-extein influences the protein splicing efficiency. There exist several studies on the protein splicing junction sequence dependency but only few inteins have been analysed systematically. The next step would be to investigate *Pho*RadA intein +2 residue dependency and currently there is no data available in the literature providing a systematic analysis of residues at both the N- and C-extein protein splicing junction sequence.

From the structural data currently available it is not obvious how the C-extein can influence the protein splicing efficiency. The C-extein could interact with the intein in a different conformation that would occur during the protein splicing reaction steps. The structure of the branched intermediate is unknown and little is known about this conformation (see Figure 17). Structural data of this intermediate could assist in understanding how the protein splicing reaction occur and how the extein might influence at this step. Studies have indicated that a local structural change around the +1 residue happens at the formation of branched intermediate (Frutos *et al.*, 2010). It is though a challenge to obtain a stable construct of a branched intermediate because the thio(ester) intermediate likely would be hydrolysed. The cleavage reaction at the C-extein junction should be possible to prevent by mutation of the last intein residue (Asn). This intein construct would require the first intein residue to be Cys or Ser and the +1 C-extein residue to be Cys, Ser, or Thr for formation of the branched intermediate (step 1 and 2 in the reaction mechanism). Residues flanking the protein splicing junction have been shown to influence the formation of the branched intermediate and could be optimised for structural studies (O'Brien et al., 2010).

A possible biological role of inteins is as of yet unidentified. Despite recent publications that have indicated possible gene regulating function and the possibility to increase genetic diversity there has not been provided any proof of such biological function (Aranko *et al.*, 2013b). It has been postulated that inteins might have had a function in early stages of life but that it has been lost during evolution (Pietrokovski,

2001). An unknown question is also what happens to the intein after the protein splicing has occurred. Studies have shown that the Hedgehog C-terminal domain is degraded by endoplasmic-reticulum-associated protein degradation after cholesterol modification of the N-terminal domain (Chen *et al.*, 2011). It is unknown if the same happens for the intein. In some inteins a DNA binding domain has been identified which might have some function after protein splicing. Inteins could have a function by interactions with other molecules in the cell that might have some regulatory function. To identify possible interaction it would require further studies on native host organisms to identify intein interactions.

Acknowledgements

This work was carried out in the period 2008-2013 in the Structural Biology and Biophysics research program at Institute of Biotechnology, University of Helsinki. I would like to acknowledge the director Tommi Mäkelä and former director Mart Saarma of the institute for providing the excellent facilities for conducting research. Additionally Biocenter Finland is acknowledged for supporting the NMR center and protein crystallization facility at Institute of biotechnology.

First and foremost I would like to acknowledge Hideo Iwai for his supervision during this thesis work. I greatly appreciate that you invited me to come to Finland and accepted me in your Lab. Hideo has inspired me with his keen intelligence and clever ideas which have meant a lot to me in my work and how my skills have developed. I would also acknowledge Reinhard Wimmer who was my supervisor during my master and who recommended me to join Hideo's Lab and provided the contact.

I appreciate my follow up committee members Professor Adrian Goldman and Professor Kari Keinänen for thoughtful insight and comments at our meetings. In particular would acknowledge Adrian for his strong opinions and his influence on me to pursue X-ray crystallography.

My pre-examiners Rikkert W. Wierenga and Ilkka Kilpeläinen are acknowledged for reviewing and providing valuable comments to improve my thesis – Thank you.

I would like to acknowledge Dongwen Zhou and Alexander Wlodawer for a good collaboration during this work and for the help preparing the co-author manuscripts.

Several people have been a part of the group and I would like to thank the present and former members for a good time in the group: Sesilja, Fernando, Anniina, Brita, Valtteri, Sandra, Cathrin, Justus, Lynn, Gerrit, Kimmo, Krista, Daniel, Lauri, Mikael, Katariina, Simo, Roosa, and Anne. In particular I would mention Sesilja for good discussions and collaborations over the years and the valuable comments you provided for my thesis. I would also like to express my gratitude to all the people from Adrian's, Sarah's, Tommi's, and Perttu's groups for interesting group seminars and a beer on occasions. I would especially like to express thanks to Seija Mäki for setting up my protein crystallization experiments and Tuomas Niemi-Aro for help when the spectrometers were on "strike".

My thesis work has been funded by The National Doctoral Programme in Informational and Structural Biology graduate school and I would like to thank the selection committee that approved my application and Professor Mark Johnson for accepting me. Additionally I would acknowledge CIMO and Academy of Finland for financial support.

I would like to thank my family for their support while I went to Finland for "three months". Finally I give the warmest thanks goes to my loved Ružica for your constant loving support, and patience.

Helsinki, November 2013
Jesper S. Øemig

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N,W., Read, R. J., Sacchettini, J. C., Sauter, N. K. and Terwilliger, T. C. (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. Sect. D: Biol. Crystallogr*. **58**, 1948–1954.

Amitai, G., Belenkiy, O., Dassa, B., Shainskaya, A. and Pietrokovski, S. (2003) Distribution and function of new bacterial intein-like protein domains. *Mol. Microbiol.* **47**, 61-73.

Amitai, G., Callahan, B.P., Stanger, M.J., Belfort, G. and Belfort, M. (2009) Modulation of intein activity by its neighboring extein substrates. *Proc. Natl. Acad .Sci. USA*. **106**, 11005-11010.

Amitai, G., Dassa, B. and Pietrokovski, S. (2004) Protein splicing of inteins with atypical glutamine and aspartate C-terminal residues. *J. Biol. Chem.* **279**, 3121-3131.

Appleby-Tagoe, J.H., Thiel, I.V., Wang, Y., Wang, Y., Mootz, H.D. and Liu, X.Q. (2011) Highly efficient and more general cis- and trans-splicing inteins through sequential directed evolution. *J. Biol. Chem*. **286**, 34440-34447.

Appleby, J.H., Zhou, K., Volkmann, G. and Liu, X.Q. (2009) Novel split intein for trans-splicing synthetic peptide onto C terminus of protein. *J. Biol. Chem*. **284**, 6194-6199.

Aranko, A. S., Züger, S., Buchinger, E. and Iwaï, H. (2009) *In Vivo* and *In Vitro* protein ligation by naturally occurring and engineered split DnaE inteins. *PLos One*. **4**, e5185.

Aranko, A.S., Oeemig, J.S. and Iwaï, H. (2013a) Structural basis for Protein Trans-Splicing by a Bacterial Intein-Like domain: Protein Ligation without nucleophilic side-chains. *FEBS J*. 280, 3256-3269.

Aranko, A.S., Oeemig, J.S., Kajander, T. and Iwaï, H. (2013b) Increased protein diversity by intein-mediated protein alternative splicing. *Nat. Chem. Biol*. Accepted.

Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*. **289**, 905-920.

Barzel, A., Naor, A., Privman, E., Kupiec, M. and Gophna. U. (2011) Homing endonucleases residing within inteins: evolutionary puzzles awaiting genetic solutions. *Biochem. Soc. Trans*. **39**, 169-173.

Belfort, M. and Boberts, R. J. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res*. **25**, 3379–3388.

Berg, J.M., Tymoczko, J.L. and Stryer, L. (2002). *Biochemistry*, 5th ed. W. H. Freeman and Company, New York, NY.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*. **28**, 235-242.

Bertini, I., Case, D. A., Ferella, L., Giachetti, A. and Rosato, A. (2011) A Grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics*. **27**, 2384–2390.

Binschik, J. and Mootz, H.D. (2013) Chemical bypass of intein-catalyzed N-S acyl shift in protein splicing. *Angew. Chem. Int. Ed. Engl*. **52**, 4260-4.

Boehr, D.D., Dyson, H.J. and Wright, P.E. (2006) An NMR perspective on enzyme dynamics. *Chem. Rev.* **106**, 3055-3079.

Borra, R., Dong, D., Elnagar, A.Y., Woldemariam, G.A. and Camarero, J.A. (2012) In-cell fluorescence activation and labeling of proteins mediated by FRET-quenched split inteins. *J. Am. Chem. Soc.* **134**, 6344-6353.

Brace, L.E., Southworth, M.W., Tori, K., Cushing, M.L. and Perler, F. (2010) The Deinococcus radiodurans Snf2 intein caught in the act: detection of the Class 3 intein signature Block F branched intermediate. *Protein Sci.* **19**, 1525-1533.

Brange, J., Ribel, U., Hansen, J.F., Dodson, G., Hansen, M.T., Havelund, S., Melberg, S.G., Norris, F., Norris, K., Snel, L., Sørensen, A.R. and Voight, H.O. (1988) Monomeric insulins obtained by protein engineering and their medical implications. *Nature*. **333**, 679-682.

Busche, A.E., Aranko, A.S., Talebzadeh-Farooji, M., Bernhard, F., Dötsch, V. and Iwaï, H. (2009) Segmental isotopic labeling of a central domain in a multidomain protein by protein trans-splicing using only one robust DnaE intein. *Angew. Chem. Int. Ed. Engl*. **48**, 6128-6131.

Callahan, B.P., Topilina, N.I., Stanger, M.J., Van Roey, P. and Belfort, M. (2011) Structure of catalytically competent intein caught in a redox trap with functional and evolutionary implications. *Nat. Struct. Mol. Biol.* **18**, 630-633.

Camarero, J.A. and Muir, T.W. (1999) Biosynthesis of a head-to-tail cyclized protein with improved biological activity. *J. Am. Chem. Soc*. **121**, 5597-5598.

Camarero, J.A., Kimura, R.H., Woo, Y.H., Shekhtman, A. and Cantor, J. (2007) Biosynthesis of a fully functional cyclotide inside living bacterial cells. *Chembiochem*. **8**, 1363-1366.

Caspi, J., Amitai, G., Belenkiy, O. and Pietrokovski, S. (2003) Distribution of split DnaE inteins in cyanobacteria. *Mol. Microbiol*. **50**, 1569-1577.

Cavanagh, J., Fairbrother, W.J., Palmer III, A.G., Skelton, N.J. and Rance, M. (2007) *Protein NMR Spectroscopy, Principles and Practice*. 2nd ed. Elsevier Academic Press: Oxford.

Chayen, N.E. and Saridakis, E. (2008) Protein crystallization: from purified protein to diffraction-quality crystal. *Nat. Methods*. **5**, 147-153.

Chen, L., Benner, J. and Perler, F.B. (2000) Protein splicing in the absence of an intein penultimate histidine. *J. Biol. Chem*. **275**, 20431-20435.

Chen, W., Li, L., Du, Z., Liu, J., Reitter, J.N., Mills, K.V., Linhardt, R.J. and Wang, C. (2012) Intramolecular disulfide bond between catalytic cysteines in an intein precursor. *J. Am. Chem. Soc.* **134**, 2500-2503.

Chen, X., Tukachinsky, H., Huang, C.H., Jao, C., Chu, Y.R., Tang, H.Y., Mueller, B., Schulman, S., Rapoport, T.A. and Salic, A. (2011) Processing and turnover of the Hedgehog protein in the endoplasmic reticulum. *J. Cell Biol*. **192**, 825-838.

Cheriyan, M., Pedamallu, C.S., Tori, K. and Perler, F. (2013) Faster protein splicing with the *Nostoc punctiforme* DnaE intein using non-native extein residues. *J. Biol. Chem.* **288**, 6202-6211.

Chong, S. and Xu, M.Q. (1997) Protein splicing of the *Saccharomyces cerevisiae* VMA intein without the endonuclease motifs. *J. Biol. Chem*. **272**, 15587-15590.

Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F.B. and Xu, M.Q. (1996) Protein splicing involving the *Saccharomyces cerevisiae* VMA intein. The steps in the splicing pathway, side reactions leading to protein cleavage, and establishment of an *in vitro* splicing system. *J. Biol. Chem*. **271**, 22159-22168.

Chong, S., Williams, K.S., Wotkowicz, C. and Xu, M.Q. (1998) Modulation of protein splicing of the *Saccharomyces cerevisiae* vacuolar membrane ATPase intein. *J. Biol. Chem*. **273**, 10567-10577.

Cooper, A.A., Chen, Y.J., Lindorfer, M.A. and Stevens, T.H. (1993) Protein splicing of the yeast TFP1 intervening protein sequence: a model for self-excision. *EMBO J*. **12**, 2575-2583.

Cordier, F. and Grzesiek, S. (1999) Direct Observation of Hydrogen Bonds in Proteins by Interresidue $^{3h}J_{NC'}$ Scalar Couplings. *J. Am. Chem. Soc.* **121**, 1601-1602.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*. **13**, 289-302.

Craik, D.J. (2006) Seamless proteins tie up their loose ends. *Science*. **311**, 1563-1564.

Dalgaard, J.Z., Moser, M.J., Hughey, R. and Mian, I.S. (1997) Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J. Comput. Biol*. **4**, 193-214.

Dassa, B., Amitai, G., Caspi, J., Schueler-Furman, O. and Pietrokovski, S. (2007) Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. *Biochemistry*. **46**, 322-330.

Dassa, B., Haviv, H., Amitai, G. and Pietrokovski, S. (2004a) Protein splicing and auto-cleavage of bacterial intein-like domains lacking a C'-flanking nucleophilic residue. *J. Biol. Chem*. **279**, 32001-32007.

Dassa, B., Yanai, I. and Pietrokovski, S. (2004b) New type of polyubiquitin-like genes with intein-like autoprocessing domains. *Trends Genet*. **20**, 538-542.

Davis, E.O., Jenner, P.J., Brooks, P.C., Colston, M.J. and Sedgwick, S.G. (1992) Protein splicing in the maturation of M. tuberculosis recA protein: a mechanism for tolerating a novel class of intervening sequence. *Cell*. **71**, 201-210.

Dawson, P.E., Muir, T.W., Clark-Lewis, I. and Kent, S.B. (1994) Synthesis of proteins by native chemical ligation. *Science*. **266**, 776-779.

Dearden, A.K., Callahan, B., Van Roey, P., Li, Z., Kumar, U., Belfort, M. and Nayak, S.K. (2013) Conserved threonine spring-loads precursor for intein splicing. *Protein Sci*. **22**, 557-563.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*. **6**, 277–293.

Delaglio, F., Torchia, D.A. and Bax, A. (1991) Measurement of ^{15}N-^{13}C J couplings in *staphylococcal* nuclease. *J. Biomol. NMR*. **1**, 439-446.

Derbyshire, V. and Belfort, M. (1998) Lightning strikes twice: intron-intein coincidence. *Proc. Natl. Acad. Sci. USA*. **95**, 1356-1357.

Derbyshire, V., Wood, D.W., Wu, W., Dansereau, J.T., Dalgaard, J.Z. and Belfort, M. (1997) Genetic definition of a protein-splicing domain: Functional mini-inteins support structure predictions and a model for intein evolution. *Proc. Natl. Acad. Sci. USA*. **94**, 11466-11471.

DiMaio, F., Terwilliger, T.C., Read, R.J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H.L., Das, D., Vorobiev, S.M., Iwaï, H., Pokkuluri, P.R. and Baker, D. (2011) Increasing the radius of convergence of molecular replacement by density and energy guided protein structure optimization. *Nature*. **473**, 540–543.

Ding, Y., Xu, M., Ghosh, I., Chen, X., Ferrandon, S., Lesage, G. and Rao, Z. (2003). Crystal structure of a mini-intein reveals a conserved catalytic module involved in Side Chain Cyclization of Asparagine during Protein Splicing. *J. Biol. Chem*. **278**, 39133-39142.

Doreleijers, J.F., Sousa da Silva, A.W., Krieger, E., Nabuurs, S.B., Spronk, C.A., Stevens, T.J., Vranken, W.F., Vriend, G. and Vuister, G.W. (2012) CING: an integrated residue-based structure validation program suite. *J. Biomol. NMR.* **54**, 267-283.

Dori-Bachash, M., Dassa, B., Peleg, O., Pineiro, S. A., Jurkevitch, E. and Pietrokovski, S. (2009) Bacterial intein-like domains of predatory bacteria: a new domain type characterized in Bdellovibrio bacteriovorus. *Funct. Integr. Genomics*. **9**, 153-66.

Du, Z., Liu, J., Albracht, C. D., Hsu, A., Chen, W., Marieni, M.D., Colelli, K.M., Williams, J.E., Reitter, J.N., Mills, K.V. and Wang, C. (2011a) Structural and mutational studies of a hyperthermophilic intein from DNA polymerase II of *Pyrococcus abyssi*. *J. Biol. Chem.* **286**, 38638-38648.

Du, Z., Liu, Y., Zheng, Y., McCallum, S., Dansereau, J., Derbyshire, V., Belfort, M., Belfort, G., Van Roey, P. and Wang, C. (2008) ^{1}H, ^{13}C, and ^{15}N NMR assignments of an engineered intein based on *Mycobacterium tuberculosis* RecA. *Biomol. NMR Assign*. **2**, 111-113.

Du, Z., Shemella, P.T., Liu, Y., McCallum, S.A., Pereira, B., Nayak, S.K., Belfort, G., Belfort M, Wang C. (2009) Highly conserved histidine plays a dual catalytic role in protein splicing: a pK_a shift mechanism. *J. Am. Chem. Soc.* **131**, 11581-11589.

Du, Z., Zheng, Y., Patterson, M., Liu, Y. and Wang, C. (2011b) pK(a) coupling at the intein active site: implications for the coordination mechanism of protein splicing with a conserved aspartate. *J. Am. Chem. Soc.* **133**, 10275-10282.

Duan, X., Gimble, F.S., Quiocho, F.A. (1997) Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell*. **89**, 555-564.

Ellisä, S. Jurvansuu, J.M. and Iwaï, H. (2011) Evaluation and comparison of protein splicing by exogenous inteins with foreign exteins in *Escherichia coli*. *FEBS lett*. **585**, 3471-3477.

Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **60**, 2126–2132.

Evans, T.C.Jr., Benner, J. and Xu, M.Q. (1998) Semisynthesis of cytotoxic proteins using a modified protein splicing element. *Protein Sci*. **7**, 2256-2264.

Evans, T.C.Jr., Benner, J. and Xu, M.Q. (1999a) The cyclization and polymerization of bacterially expressed proteins using modified self-splicing inteins. *J. Biol. Chem.* **274**, 18359-18363.

Evans, T.C.Jr., Benner, J. and Xu, M.Q. (1999b) The *in vitro* ligation of bacterially expressed proteins using an intein from *Methanobacterium thermoautotrophicum*. *J. Biol. Chem.* **274**, 3923-3926.

Evans, T.C.Jr., Martin, D., Kolly, R., Panne, D., Sun, L., Ghosh, I., Chen, L., Benner, J., Liu, X.Q. and Xu, M.Q. (2000) Protein trans-splicing and cyclization by a naturally split intein from the dnaE gene of *Synechocystis* species PCC6803. *J. Biol. Chem*. **275**, 9091-9094.

Farrow, N.A., Muhandiram, R., Singer, A.U., Pascal, S.M., Kay, C.M., Gish, G., Shoelson, S.E., Pawson, T., Forman-Kay, J.D. and Kay, L.E. (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry*. **33**, 5984–6003.

Frischkorn, K., Sander, P., Scholz, M., Teschner, K., Prammananan, T. and Böttger, E.C. (1998) Investigation of mycobacterial recA function: protein introns in the RecA of pathogenic mycobacteria do not affect competency for homologous recombination. *Mol. Microbiol*. **29**, 1203-1214.

Frutos, S., Goger, M., Giovani, B., Cowburn, D., and Muir, T.W. (2010) Branched intermediate formation stimulates peptide bond cleavage in protein splicing. *Nat Chem. Biol*. **6**, 527-533.

Fsihi, H., Vincent, V. and Cole, S.T. (1996) Homing events in the gyrA gene of some mycobacteria. *Proc. Natl. Acad. Sci. USA*. **93**, 3410-3415.

Gautier, A., Mott, H.R., Bostock, M.J., Kirkpatrick, J.P. and Nietlispach, D. (2010) Structure determination of the seven-helix transmembrane receptor sensory rhodopsin II by solution NMR spectroscopy. *Nat. Struct. Mol. Biol*. **17**, 768-774.

Ghosh, I., Sun, L. and Xu, M.Q. (2001) Zinc inhibition of protein trans-splicing and identification of regions essential for splicing and association of a split intein. *J. Biol. Chem.* **276**, 24051-24058.

Gimble, F.S. and Thorner, J. (1992) Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. *Nature*. **357**, 301-306.

Giriat, I. and Muir, T.W. (2003) Protein semi-synthesis in living cells. *J. Am. Chem. Soc*. **125**, 7180-7181.

Gorbalenya, A.E. (1998) Non-canonical inteins. *Nucleic Acids Res*. **26**, 1741-1748.

Goto, N.K. and Kay, L.E. (2000) New developments in isotope labeling strategies for protein solution NMR spectroscopy. *Curr. Opin. Struct. Biol.* **10**, 585-592.

Grindl, W., Wende, W., Pingoud, V. and Pingoud, A. (1998) The protein splicing domain of the homing endonuclease PI-sceI is responsible for specific DNA binding. *Nucleic Acids Res*. **26**, 1857-1862.

Güntert, P. (1998) Structure calculation of biological macromolecules from NMR data. *Q. Rev. Biophys*. **31**, 145-237.

Güntert, P. (2009) Automated structure determination from NMR spectra. *Eur. Biophys. J.* **38**, 129–143.

Güntert, P., Braun, W. and Wüthrich, K. (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J. Mol. Biol*. **217**, 517–530.

Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol*. **273**, 283-298.

Habeck, M., Rieping, W., Linge, J.P. and Nilges, M. (2004) NOE assignment with ARIA 2.0: the nuts and bolts. *Methods Mol. Biol*. **278**, 379-402.

Hall, T.M., Porter, J.A., Young, K.E., Koonin, E.V., Beachy, P.A. and Leahy, D.J. (1997) Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins. *Cell*. **91**, 85-97.

He, Z., Crist, M., Yen, H., Duan, X., Quiocho, F.A. and Gimble, F.S. (1998) Amino acid residues in both the protein splicing and endonuclease domains of the PI-SceI intein mediate DNA binding. *J. Biol. Chem.* **273**, 4607-4615.

Hendrickson, W.A., Horton, J.R., LeMaster, D.M. (1990) Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J*. **9**, 1665-1672.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol*. **319**, 209–227.

Hiraga, K., Derbyshire, V., Dansereau, J.T., Van Roey, P. and Belfort, M. (2005) Minimization and stabilization of the *Mycobacterium tuberculosis* recA intein. *J. Mol. Biol*. **354**, 916-926.

Hiraga, K., Soga, I., Dansereau, J. T., Pereira, B., Derbyshire, V., Du, Z., Wang, C., Van Roey, P., Belfort, G. and Belfort M. (2009) Selection and structure of hyperactive inteins: peripheral changes relayed to the catalytic center. *J. Mol. Biol.* **393**, 1106-1117.

Hirata, R., Ohsumk, Y., Nakano, A., Kawasaki, H., Suzuki, K. and Anraku, Y. (1990) Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of Saccharomyces cerevisiae. *J. Biol. Chem*. **265**, 6726-6733.

Hodges, R.A., Perler, F.B., Noren, C.J. and Jack, W.E. (1992) Protein splicing removes intervening sequences in an archaea DNA polymerase. *Nucleic Acids Res*. **20**, 6153-6157.

Holm, L. and Rosenström, P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**:W545–9.

Hu, D., Crist, M., Duan, X., Quiocho, F.A., and Gimble, F.S. (2000) Probing the structure of the PI-SceI-DNA complex by affinity cleavage and affinity photocross-linking. *J. Biol. Chem.* **275**, 2705–2712.

Huet, G., Castaing, J.P., Fournier, D., Daffé, M. and Saves, I. (2006) Protein splicing of SufB is crucial for the functionality of the Mycobacterium tuberculosis SUF machinery. *J. Bacteriol*. **188**, 3412-3414.

Ichiyanagi, K., Ishino, Y., Ariyoshi, M., Komori, K. and Morikawa, K. (2000) Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. *J. Mol. Biol.* **300**, 889-901.

Inglis, A.S. (1983) Cleavage at aspartic acid. *Methods Enzymol*. **91**, 324–332.

Iwai, H. and Plückthun, A. (1999) Circular beta-lactamase: stability enhancement by cyclizing the backbone. *FEBS Lett.* **459**,166-172.

Iwai, H., Lingel, A. and Pluckthun, A. (2001) Cyclic green fluorescent protein produced in vivo using an artificially split PI-*Pfu*I intein from *Pyrococcus furiosus*. *J. Biol. Chem.* **276**, 16548-16554.

Iwai, H., Züger, S., Jin, J. and Tam, P.H. (2006) Highly efficient protein trans-splicing by a naturally split DnaE intein from *Nostoc punctiforme*. *FEBS Lett*. **580**, 1853-1858.

Jagadish, K., Borra, R., Lacey, V., Majumder, S., Shekhtman, A., Wang, L. and Camarero, J.A. (2013) Expression of Fluorescent Cyclotides using Protein Trans-Splicing for Easy Monitoring of Cyclotide-Protein Interactions. *Angew. Chem. Int. Ed. Engl*. **52**, 3126-3131.

Jee, J. and Güntert, P. (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *J. Struct. Funct. Genomics*. **4**, 179-189.

Jennings, C., West, J., Waine, C., Craik, D. and Anderson, M. (2001) Biosynthesis and insecticidal properties of plant cyclotides: the cyclic knotted proteins from *Oldenlandia affinis*. *Proc. Natl Acad. Sci. USA*. **98**, 10614-10619.

Johnson, M.A., Southworth, M.W., Herrmann, T., Brace, L., Perler, F.B. and Wüthrich, K. (2007a) NMR structure of a KlbA intein precursor from *Methanococcus jannaschii*. *Protein Science.* **16**, 1316-1328.

Johnson, M.A., Southworth, M.W., Perler, F.B. and Wüthrich, K. (2007b) NMR assignment of a KlbA intein precursor from *Methanococcus jannaschii*. *Biomol. NMR Assign*. **1**, 19-21.

Jones, P.T., Dear, P.H., Foote, J., Neuberger, M.S. and Winter, G. (1986) Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*. **321**, 522-525.

Juranić, N., Ilich, P.K. and Macura, S. (1995) Hydrogen Bonding Networks in Proteins As Revealed by the Amide $^{1}J_{NC'}$ Coupling Constant. *J. Am. Chem. Soc*. **117**, 405–410.

Kabsch, W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr*. **26**, 795–800.

Kainosho, M., Torizawa, T., Iwashita, Y., Terauchi, T., Mei Ono, A. & Güntert, P. (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature*. **440**, 52-57.

Kane, P.M., Yamashiro, C.T., Wolczyk, D.F., Neff, N., Goebl, M., and Stevens, T.H. (1990) Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. *Science*. **250**, 651-657.

Kawasaki, M., Sarow, Y., Ohya, Y. and Anraku, Y. (1997) Protein splicing in the yeast Vma1 protozyme: evidence for an intramolecular reaction. *FEBS lett.* **412**, 518-520

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Phillips, D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*. **181**, 662-666.

Kerrigan, A.M., Powers, T.L., Dorval, D.M., Reitter, J.N. and Mills, K.V. (2009) Protein splicing of the three *Pyrococcus abyssi* ribonucleotide reductase inteins. *Biochem Biophys. Res. Commun*. **387**, 153-157.

Klabunde, T., Sharma, S., Telenti, A., Jacobs, W. R. and Sacchettini, J. C. (1998) Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing. *Nat. Struct. Biol.* **5**, 31–36.

Kurpiers, T. and Mootz, H.D. (2007) Regioselective cysteine bioconjugation by appending a labeled cystein tag to a protein by using protein splicing in trans. *Angew. Chem. Int. Ed. Engl*. **46**, 5234-5237.

Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*. **8**, 477–486.

Lazarevic, V., Soldo, B., Düsterhöft, A., Hilbert, H., Mauël, C. and Karamata, D. (1998) Introns and intein coding sequence in the ribonucleotide reductase genes of Bacillus subtilis temperate bacteriophage SPbeta. *Proc. Natl. Acad. Sci. USA*. **95**, 1692-1697.

Lee, J.J., von Kessler, D.P., Parks, S. and Beachy, P.A. (1992) Secretion and localized transcription suggest a role in positional signaling for products of the segmentation gene hedgehog. *Cell*. **71**, 33-50.

Lew, B.M., Mills, K.V. and Paulus, H. (1998) Protein splicing in vitro with a semisynthetic two-component minimal intein. *J. Biol. Chem*. **273**, 15887-15890.

Liu, X.Q. (2000) Protein-splicing intein: Genetic mobility, origin, and evolution. *Annu. Rev. Genet*. **34**, 61-76.

Liu, X.Q. and Yang, J. (2004) Prp8 intein in fungal pathogens: target for potential antifungal drugs. *FEBS Lett*. **572**, 46-50.

Liu, X.Q., and Hu, Z. (1997) A DnaB intein in *Rhodothermus marinus*: indication of recent intein homing across remotely related organisms. *Proc. Natl. Acad. Sci. USA*. **94**, 7851-7856.

Liu, X.Q., Yang, J. and Meng, Q. (2003) Four inteins and three group II introns encoded in a bacterial ribonucleotide reductase gene. *J. Biol. Chem*. **278**, 46826-46831.

Liu, Y. and Prestegard, J.H. (2009) Measurement of one and two bond N-C couplings in large proteins by TROSY-based J-modulation experiments. *J. Magn. Reson*. **200**, 109-118.

Lockless, S.W. and Muir, T.W. (2009) Traceless protein splicing utilizing evolved split inteins. *Proc. Natl. Acad. Sci. USA*. **106**, 10999-11004.

Luckett, S., Garcia, R.S., Barker, J.J., Konarev, A.V., Shewry, P.R., Clarke, A.R. and Brady, R.L. (1999) High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. *J. Mol. Biol.* **290**, 525-533.

Ludwig, C., Pfeiff, M., Linne, U. and Mootz, H.D. (2006) Ligation of a synthetic peptide to the N terminus of a recombinant protein using semisynthetic protein trans-splicing. *Angew. Chem. Int. Ed. Engl*. **45**, 5218-5221.

Ludwig, C., Schwarzer, D. and Mootz, H.D. (2008) Interaction studies and alanine scanning analysis of a semi-synthetic split intein reveal thiazoline ring formation from an intermediate of the protein splicing reaction. *J. Biol. Chem.* **283**, 25264-25272.

Mathys, S., Evans, T.C., Chute, I.C., Wu, H., Chong, S., Benner, J., Liu, X.Q. & Xu, M.Q. (1999) Characterization of a self-splicing mini-intein and its conversion into autocatalytic N- and C-terminal cleavage elements: facile production of protein building blocks for protein ligation. *Gene*. **231**, 1-13.

Matsumura, H., Takahashi, H., Inoue, T., Yamamoto, T., Hashimoto, H., Nishioka, M., Fujiwara, S., Takagi, M., Imanaka, T. and Kai, Y. (2006) Crystal structure of intein homing endonuclease II encoded in DNA polymerase gene from hyperthermophilic archaeon *Thermococcus kodakaraensis* strain KOD1. *Proteins*. **63**, 711-715.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. and Read, R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr*. **40**, 658–674.

Mills, K.V. and Paulus, H. (2001) Reversible inhibition of protein splicing by zinc ion. *J. Biol. Chem*. **276**, 10832-10838.

Mills, K.V., Connor, K.R., Dorval, D.M. and Lewandowski, K.T. (2006) Protein purification via temperature-dependent, intein-mediated cleavage from an immobilized metal affinity resin. *Anal. Biochem*. **356**, 86-93.

Mills, K.V., Lew, B.M., Jiang, S. and Paulus, H. (1998) Protein splicing in trans by purified N- and C-terminal fragments of the *Mycobacterium tuberculosis* RecA intein. *Proc. Natl. Acad. Sci. USA*. **95**, 3543-3548.

Mills, K.V., Manning, J.S., Garcia, A.M. and Wuerdeman, L.A. (2004) Protein splicing of a Pyrococcus abyssi intein with a C-terminal glutamine. *J. Biol. Chem*. **279**, 20685-20691.

Minato, Y., Ueda, T., Machiyama, A., Shimada, I. and Iwaï, H. (2012) Segmental isotopic labeling of a 140 kDa dimeric multi-domain protein CheA from Escherichia coli by expressed protein ligation and protein trans-splicing. *J. Biomol. NMR*. **53**, 191-207.

Mizutani, R., Anraku, Y. and Satow, Y. (2004) Protein splicing of yeast VMA1-derived endonuclease via thiazolidine intermediates. *J. Synchrotron. Radiat*. **11**, 109-112.

Mizutani, R., Nogami, S., Kawasaki, M., Ohya, Y., Anraku, Y. and Satow, Y. (2002). Protein-splicing Reaction via a Thiazolidine Intermediate: Crystal Structure of the VMA1-derived Endonuclease Bearing the N and C-terminal Propeptides. *J. Mol. Biol*. **316**, 919–929.

Mossessova, E. and Lima, C.D. (2000) Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol. Cell*. 5, 865-876.

Moure, C.M., Gimble, F.S. and Quiocho, F.A. (2002) Crystal structure of the intein homing endonuclease PI-SceI bound to its recognition sequence. *Nat. Struct. Biol*. **9**, 764-770.

Muir, T.W., Sondhi, D. and Cole, P.A. (1998) Expressed protein ligation: a general method for protein engineering. *Proc. Natl. Acad. Sci. USA*. **95**, 6705-6710.

Muona, M., Aranko, A.S. and Iwai, H. (2008) Segmental isotopic labelling of a multidomain protein by protein ligation by protein trans-splicing. *Chembiochem*. **9**, 2958-2961.

Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. Sect. D: Biol. Crystallogr.* **53**, 240–255.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. **247**, 536-540.

Niehaus, F., Frey, B. and Antranikian, G. (1997) Cloning and characterisation of a thermostable alpha-DNA polymerase from the hyperthermophilic archaeon Thermococcus sp. TY. *Gene*. **204**, 153-158.

Noren, C.J., Wang, J. and Perler, F.B. (2000) Dissecting the Chemistry of Protein Splicing and Its Applications. *Angew. Chem. Int. Ed. Engl*. **,** 450-466.

O'Brien, K.M., Schufreider, A.K., McGill, M.A., O'Brien, K.M., Reitter, J.N. and Mills, K.V. (2010) Mechanism of protein splicing of the *Pyrococcus abyssi* lon protease intein. *Biochem. Biophys. Res. Commun.* **403**, 457-461.

Ohki, S. and Kainosho, M. (2008) Stable isotope labeling methods for protein NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **53**, 208–226.

Oksanen, E. and Goldman, A. (2010) Introduction to Macromolecular X-Ray Crystallography in "Comprehensive Natural Products Chemistry II" Vol. 9. Mander, L. & Liu, B. (eds.). Amsterdam: Elsevier.

Olschewski, D., Seidel, R., Miesbauer, M., Rambold, A.S., Oesterhelt, D., Winklhofer, K.F., Tatzelt, J., Engelhard, M. and Becker, C.F. (2007) Semisynthetic murine prion protein equipped with a GPI anchor mimic incorporates into cellular membranes. *Chem. Biol*. **14**, 994-1006.

Otomo, T., Teruya, K., Uegaki, K., Yamazaki, T., and Kyogoku, Y. (1999) Improved segmental isotope labeling of proteins and application to a larger protein. *J. Bio. NMR*. **15**, 105-114.

Otwinowski, Z. & Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol*. **276**, 307–326.

Page, R., Peti, W., Wilson, I.A., Stevens, R.C. and Wüthrich, K. (2005) NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. *Proc. Natl. Acad. Sci. USA*. **102**, 1901-1905.

Palmer, A.G. III (2004) NMR characterization of the dynamics of biomacromolecules. *Chem. Rev*. **104**, 3623-3640.

Papavinasasundaram, K.G., Colston, M.J. and Davis, E.O. (1998) Construction and complementation of a recA deletion mutant of *Mycobacterium smegmatis* reveals that the intein in *Mycobacterium tuberculosis* recA does not affect RecA function. *Mol. Microbiol*. **30**, 525-534.

Paulus, H. (2003) Inteins as targets for potential antimycobacterial drugs. *Front Biosci*. **8**, s1157-1165.

Pearl, E.J., Bokor, A.A., Butler, M.I., Poulter, R.T. and Wilbanks, S.M. (2007a) Preceding hydrophobic and β-branched amino acids attenuate splicing by the *Cne*PRP8 intein. *Biochim. Biophys. Acta*. **1774**, 995-1001.

Pearl, E.J., Tyndall, J.D., Poulter, R.T. and Wilbanks, S.M. (2007b) Sequence requirements for splicing by the Cne PRP8 intein. *FEBS Lett*. **581**, 3000-3004.

Perler, F.B. (2002) InBase: the Intein Database. *Nucleic Acids Res*. **30**, 383-384.

Perler, F.B., Comb, D.G., Jack, W.E., Moran, L.S., Qiang, B., Kucera, R.B., Benner, J., Slatko, B.E., Nwankwo, D.O., Hempstead, S.K., Clotilde, K.S., Carlow, C.K. and Jannasch, H. (1992) Intervening sequences in an Archaea DNA polymerase gene. *Proc. Natl. Acad. Sci. USA*. **89**, 5577-5581.

Perler, F.B., Davis, E.O., Dean, G.E., Gimble, F.S., Jack, W.E., Neff, N., Noren, C.J., Thorner, J. and Belfort, M. (1994) Protein splicing elements: inteins and exteins - a definition of terms and recommended nomenclature. *Nucleic Acids Res*. **22**, 1125-1127.

Perler, F.B., Olsen, G.J. and Adam, E. (1997) Compilation and analysis of intein sequences. *Nucleic Acids Res.* **25**, 1087-1093.

Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G. and North, A.C. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature*. **185**, 416-422.

Pervushin, K., Riek, R., Wider, G. and Wüthrich, K. (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci. USA*. **94**, 12366-12371.

Pietrokovski, S. (1994) Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Sci*. **3**, 2340-50.

Pietrokovski, S. (1998a) Identification of a virus intein and a possible variation in the protein-splicing reaction. *Curr. Biol*. **8**, R634-635.

Pietrokovski, S. (1998b) Modular organization of inteins and C-terminal autocatalytic domains. *Protein Sci*. **7**, 64-71.

Pietrokovski, S. (2001) Intein spread and extinction in evolution. *Trends Genet*. **17**, 465–472.

Poland, B.W., Xu, M.Q. and Quiocho, F.A. (2000) Structural insights into the protein splicing mechanism of PI-*Sce*I. *J. Biol. Chem*. **275**, 16408–16413.

Porter, J.A., von Kessler, D.P., Ekker, S.C., Young, K.E., Lee, J.J., Moses, K. and Beachy, P.A. (1995) The product of hedgehog autoproteolytic cleavage active in local and long-range signalling. *Nature*. **374**, 363-366.

Ramirez, M., Valdes, N., Guan, D. and Chen, Z. (2013) Engineering split intein DnaE from *Nostoc punctiforme* for rapid protein purification. *Protein Eng. Des. Sel*. **26**, 215-223.

Roberts, R.J. and Macelis, D. (1997) REBASE-restriction enzymes and methylases. *Nucleic Acids Res*. **25**, 248-262.

Romanelli, A., Shekhtman, A., Cowburn, D. and Muir, T.W. (2004) Semisynthesis of a segmental isotopically labeled protein splicing precursor: NMR evidence for an

unusual peptide bond at the N-extein-intein junction. *Proc. Natl. Acad. Sci. USA*. **101**, 6397-402.

Rossmann, M.G. (1990) The Molecular Replacement Method. *Acta Cryst*. **A46**, 73-82.

Rupp, B. (2010) *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Science, New York, NY.

Saether, O., Craik, D.J., Campbell, I.D., Sletten, K., Juul, J. and Norman, D.G. (1995) Elucidation of the primary and three-dimensional structure of the uterotonic polypeptide kalata B1. *Biochemistry*. **34**, 4147-4158.

Sattler, M., Schleucher, J. and Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. NMR Spectroc*. 34, 93–158.

Scott, C.P., Abel-Santos, E., Wall, M., Wahnon, D.C. and Benkovic, S.J. (1999) Production of cyclic peptides and proteins in vivo. *Proc. Natl. Acad. Sci. USA*. **96**, 13638-13643.

Shah, N.H., Dann, G.P., Vila-Perelló, M., Liu, Z. and Muir, T.W. (2012) Ultrafast protein splicing is common among cyanobacterial split inteins: implications for protein engineering. *J. Am. Chem. Soc*. **134**, 11338-11341.

Shah, N.H., Vila-Perelló, M. and Muir, T.W. (2011) Kinetic control of one-pot trans-splicing reactions by using a wild-type and designed split intein. *Angew. Chem. Int. Ed Engl*. **50**, 6511-6515.

Shao, Y. and Paulus, H. (1997) Protein splicing: estimation of the rate of 0-N and S-N acyl rearrangements, the last step of the splicing process. *J. Pept. Res*. **50**, 193-198.

Shao, Y., Xu, M.Q. and Paulus, H. (1995) Protein splicing: characterization of the aminosuccinimide residue at the carboxyl terminus of the excised intervening sequence. *Biochemistry*. **34**, 10844-10850.

Shao, Y., Xu, M.Q. and Paulus, H. (1996) Protein splicing: evidence for an N-O acyl rearrangement as the initial step in the splicing process. *Biochemistry*. **35**, 3810-3815.

Shemella, P., Pereira, B., Zhang, Y., Van Roey, P., Belfort, G., Garde, S. and Nayak, S.K. (2007) Mechanism for intein C-terminal cleavage: a proposal from quantum mechanical calculations. *Biophys. J*. **92**, 847-853.

Shen, Y., Delaglio, F., Cornilescu, G. and Bax, A. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*. **44**, 213–223.

Shingledecker, K., Jiang, S.Q. and Paulus, H. (1998) Molecular dissection of the Mycobacterium tuberculosis RecA intein: design of a minimal intein and of a trans-splicing system involving two intein fragments. *Gene*. **207**, 187-195.

Shingledecker, K., Jiang, Sq. and Paulus, H. (2000) Reactivity of the cysteine residues in the protein splicing active center of the *Mycobacterium tuberculosis* RecA intein. *Arch. Biochem. Biophys*. **375**, 138-144.

Shub, D.A. and Goodrich-Blair, H. (1992) Protein introns: a new home for endonucleases. *Cell*. **71**, 183-186.

Snyder, D.A., Chen, Y., Denissova, N.G., Acton, T., Aramini, J.M., Ciano, M., Karlin, R., Liu, J., Manor, P., Rajan, P.A., Rossi, P., Swapna, G.V., Xiao, R., Rost, B., Hunt, J.

and Montelione, G.T. (2005) Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. *J. Am. Chem. Soc.* **127**, 16505-16511.

Southworth, M.W., Adam, E., Panne, D., Byer, R., Kautz, R. and Perler, F.B. (1998) Control of protein splicing by intein fragment reassembly. *EMBO J.* **17**, 918-926.

Southworth, M.W., Amaya, K., Evans, T.C., Xu, M.Q. and Perler, F.B. (1999) Purification of proteins fused to either the amino or carboxy terminus of the *Mycobacterium xenopi* gyrase A intein. *Biotechniques*. **27**, 110-114, 116, 118-120.

Southworth, M.W., Benner, J. and Perler, F.B. (2000) An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. *EMBO J.* **19**, 5019-5026.

Southworth, M.W., Yin, J. and Perler, F.B. (2004) Rescue of protein splicing activity from a *Magnetospirillum magnetotacticum* intein-like element. *Biochem. Soc. Trans*. **32**, 250-254.

Sun, P., Ye, S., Ferrandon, S., Evans, T. C., Xu, M.Q. and Rao, Z. (2005) Crystal structures of an intein from the split dnaE gene of Synechocystis sp. PCC6803 reveal the catalytic model without the penultimate histidine and the mechanism of zinc ion inhibition of protein splicing. *J. Mol. Biol*. **353**, 1093–1105.

Sun, W., Yang., J. and Liu, W.Q. (2004) Synthetic two-piece and three-piece split inteins for protein trans-splicing. *J. Biol. Chem*. **279**, 35281-35286.

Swithers, K.S., Senejani, A.G., Fournier, G.P. and Gogarten, J.P. (2009) Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC Evol. Biol*. **9**:303.

Tang, Y.Q., Yuan, J., Osapay, G., Osapay, K., Tran, D., Miller, C.J., Ouellette, A.J. and Selsted, M.E. (1999) A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated α-defensins. *Science*. **286**, 498-502.

Telenti, A., Southworth, M., Alcaide, F., Daugelat, S., Jacobs, W.R. Jr and Perler, F.B. (1997) The Mycobacterium xenopi GyrA protein splicing element: characterization of a minimal intein. *J. Bacteriol*. **179**, 6378-6382.

Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc. Natl. Acad. Sci. USA*. **92**, 9279-9283.

Torda, A.E., Brunne, R.M., Huber, T., Kessler, H. and van Gunsteren, W.F. (1993) Structure refinement using time-averaged J-coupling constant restraints. *J. Biomol. NMR*. **3**, 55-66.

Tori, K., Cheriyan, M., Pedamallu, C.S., Contreras, M.A. and Perler, F.B. (2012) The *Thermococcus kodakaraensis Tko* CDC21-1 intein activates its N-terminal splice junction in the absence of a conserved histidine by a compensatory mechanism. *Biochemistry*. **51**, 2496-505.

Tori, K., Dassa, B., Johnson, M.A., Southworth, M.W., Brace, L.E., Ishino, Y., Pietrokovski, S., and Perler, F.B. (2010) Splicing of the mycobacteriophage Bethlehem DnaB intein: identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. *J. Biol. Chem*. **285**, 2515-2526.

Tugarinov, V., Choy, W.Y., Orekhov, V.Y. and Kay, L.E. (2005) Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc. Natl. Acad. Sci. USA*. **102**, 622-627.

Van Roey, P., Pereira, B., Li, Z., Hiraga, K., Belfort, M. and Derbyshire, V. (2007) Crystallographic and mutational studies of *Mycobacterium tuberculosis* recA mini-inteins suggest a pivotal role for a highly conserved aspartate residue. *J. Mol. Biol.* **367**, 162–173.

Vitali, F., Henning, A., Oberstrass, F.C., Hargous, Y., Auweter, S.D., Erat, M., and Allain, F.H. (2006) Structure of the two most C-terminal RNA recognition motifs of PTB using segmental isotope labeling. *EMBO J*. **25**, 150-162.

Volkmann, G. and Liu, X.Q. (2009) Protein C-terminal labeling and biotinylation using synthetic peptide and split-intein. *PLoS One*. **4**, e8381.

Volkmann, G., Murphy, P.W., Rowland, E.E., Cronan, J.E. Jr., Liu, X.Q., Blouin, C. and Byers, D.M. (2010) Intein-mediated cyclization of bacterial acyl carrier protein stabilizes its folded conformation but does not abolish function. *J. Biol. Chem*. **285**, 8605-8614.

Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J. and Laue, E.D. (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*. **59**, 687–696.

Wang, S. and Liu, X.Q. (1997) Identification of an unusual intein in chloroplast ClpP protease of *Chlamydomonas eugametos*. *J. Biol. Chem*. **272**, 11869-11873.

Watson, J.D. and Crick, F.H.C. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*. **171**, 737-738.

Werner, E., Wende, W., Pingoud, A., and Heinemann, U. (2002) High resolution crystal structure of domain I of the *Saccharomyces cerevisiae* homing endonuclease PI-*Sce*I. *Nucleic Acids Res*. **30**, 3962-71.

Wider, G. and Wüthrich, K. (1999) NMR spectroscopy of large molecules and multimolecular assemblies in solution. *Curr. Opin. Struct. Biol*. **9**, 594-601.

Williamson, M.P., Havel, T.F. and Wüthrich, K. (1985) Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. *J. Mol. Biol*. **182**, 295-315.

Wishart, D.S. and Sykes, B.D. (1994) The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR*. **4**, 171-180.

Wood, D.W., Wu, W., Belfort, G., Derbyshire, V. and Belfort M. (1999) A genetic system yields self-cleaving inteins for bioseparations. *Nat. Biotechnol*. **17**, 889-892.

Wu, H., Hu, Z. and Liu, X.Q. (1998a) Protein trans-splicing by a split intein encoded in a split DnaE gene of Synechocystis sp. PCC6803. *Proc. Natl. Acad. Sci. USA*. **95**, 9226–9231.

Wu, H., Xu, M.Q. and Liu, X.Q. (1998b) Protein trans-splicing and functional mini-inteins of a cyanobacterial dnaB intein. *Biochim. Biophys. Acta*. **1387**, 422-432.

Wynne, S.A., Crowther, R.A. and Leslie, A.G. (1999) The crystal structure of the human hepatitis B virus capsid. *Mol Cell*. **3**, 771-780.

Xu, M.Q. and Perler, F.B. (1996) The mechanism of protein splicing and its modulation by mutation. *EMBO J*. **15**, 5146-5153.

Xu, M.Q., Comb, D.G., Paulus, H., Noren, C.J., Shao, Y. and Perler, F.B. (1994) Protein splicing: an analysis of the branched intermediate and its resolution by succinimide formation. *EMBO J*. **13**, 5517-5522.

Xu, M.Q., Paulus, H. and Chong, S. (2000) Fusions to self-splicing inteins for protein purification. *Methods Enzymol*. **326**, 376-418.

Xu, M.Q., Southworth, M.W., Mersha, F.B., Hornstra, L.J. and Perler, F.B. (1993) In Vitro Protein Splicing of Purified Precursor and the Identification of a Branched intermediate. *Cell*. **75**, 1371-1377.

Yagi, H., Tsujimoto, T., Yamazaki, T., Yoshida, M. and Akutsu, H. (2004) Conformational Change of H+-ATPase ß Monomer Revealed on Segmental Isotope Labeling NMR Spectroscopy. *J. Am. Chem. Soc*. **126**, 16632-16638.

Yamazaki, T., Otomo, T., Oda, N., Kyogoku, Y., Uegaki, K., Ito, N., Ishino, Y. and Nakamura, H. (1998) Segmental isotope labeling for protein NMR using peptide splicing. *J. Am. Chem. Soc*. **120**, 5591-5592.

Yee, A.A., Savchenko, A., Ignachenko, A., Lukin, J., Xu, X., Skarina, T., Evdokimova, E., Liu, C.S., Semesi, A., Guido, V., Edwards, A.M. and Arrowsmith, C.H. (2005) NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins. *J. Am. Chem. Soc*. **127**, 16512-16517.

Zettler, J., Schütz, V. and Mootz, H.D. (2009) The naturally split *Npu*DnaE intein exhibits an extraordinarily high rate in the protein trans-splicing reaction. *FEBS Lett*. **583**, 909-914.

Zhang, X., Settembre, E., Xu, C., Dormitzer, P.R., Bellamy, R., Harrison, S.C. and Grigorieff, N. (2008) Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. *Proc. Natl. Acad. Sci. USA*. **105**, 1867-1872.

Zheng, Y., Wu, Q., Wang, C., Xu, M.Q. and Liu, Y. (2012) Mutual synergistic protein folding in split intein. *Biosci. Rep.* **32**, 433-442.

Züger, S. and Iwai, H. (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. *Nat. Biotechnol*. **23**, 736-740.