Division of Molecular Structural Biology
Department of Medical Biochemistry and Biophysics
Karolinska Institutet, Stockholm, Sweden

STRUCTURAL STUDIES OF GLYCOPROTEIN SORTING AND PROCESSING COMPLEXES IN THE EARLY SECRETORY PATHWAY

Dominik Possner

Karolinska Institutet

Stockholm 2016

Published by Karolinska Institutet.
Printed by E-Print AB 2016

ISBN 978-91-7676-240-0

Structural Studies of Glycoprotein Sorting and Processing Complexes in the Early Secretory Pathway

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Dominik Possner

Principal Supervisor:
Dr. Jodie Guy
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Molecular Structural Biology

Co-supervisor:
Prof. Ylva Lindqvist
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Molecular Structural Biology

Opponent:
Prof. Petri Kursula
Universitetet i Bergen
Department of Biomedicine

Examination Board:
Prof. Adnane Achour
Karolinska Institutet
Department of Medicine, Solna

Dr. Ylva Ivarsson
Uppsala Universitet
Department of Chemistry - BMC

Dr. Herwig Schüler
Karolinska Institutet
Department of Medical Biochemistry and Biophysics

ABSTRACT

The secretory pathway of a eukaryotic cell exerts a stringently regulated quality control system for the correct folding and transport of newly synthesised proteins and their subsequent transfer to their final destinations. While this pathway has been mostly studied in yeast, it has become increasingly clear that aberrations in its function are the main causes of a range of human disorders, and we are still in the process of gathering knowledge of the underlying molecular structures and mechanisms of the disease-causing agents in order to fully understand their impact. ERGIC-53, Erv41p and Ktr4p are all membrane anchored proteins involved in the transport or processing of glycoproteins, and form the main focus of this thesis. We have used X-ray crystallography and small-angle X-ray scattering, together with complementary biophysical and biochemical methods, to provide detailed descriptions of these proteins.

The human glycoprotein transporter ERGIC-53 is responsible for the export of specific cargo proteins, which it binds in the endoplasmic reticulum (ER) and releases in the ER-Golgi intermediate compartment whilst cycling between the two organelles. Association of ERGIC-53 with the co-transporter MCFD2 is required for the transport of a subset of cargo proteins. ERGIC-53 is only able exert its function when present within the cell as a hexamer, but the details regulating its oligomeric state are still debated, and the structure of the hexameric protein remains unknown. We show that the oligomerisation of ERGIC-53 is independent of disulfide-bond formation and, based on small-angle X-ray scattering experiments, propose two alternative shapes describing the structure of the soluble lumenal part of the protein in its hexameric state.

Erv41p is a glycoprotein transporter found in complex with its homologue Erv46p in yeast, and, similarly to ERGIC-53, the complex governs the export of cargo proteins from the ER to the Golgi apparatus, as well as the retrieval of escaped proteins back towards the ER. We have determined the structure of the soluble domain of Erv41p by X-ray crystallography, and show that the protein is comprised of a twisted β-sandwich. With almost the entirety of the concave face of Erv41p being negatively charged, this could be the site of interaction with its cargo or another interaction partner.

One cargo protein that has recently been revealed to be transported by the Erv41p/Erv46p-complex is Ktr4p, a protein localised to the Golgi apparatus. Ktr4p is a member of a protein family associated with glycoprotein processing. The structures of the Ktr4p apo-protein and its complex with GDP were determined by X-ray crystallography and show that the protein is comprised of a central β-sheet surrounded by α-helices, and that it belongs to the GT-A fold class of glycosyltransferases In addition, we have biochemically characterised the protein's function and show that it indeed possesses mannosyltransferase activity.

LIST OF SCIENTIFIC PAPERS

I. Biterova, E. I., Svärd, M., Possner, D. D. D., & Guy, J. E. (2013). Purification, crystallization and preliminary X-ray crystallographic analysis of the lumenal domain of the ER-vesicle protein Erv41p from *Saccharomyces cerevisiae*. *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications*, *69*(5), 544–546. http://doi.org/10.1107/S1744309113008063

II. Biterova, E. I., Svärd, M., Possner, D. D. D., & Guy, J. E. (2013). The crystal structure of the lumenal domain of Erv41p, a protein involved in transport between the endoplasmic reticulum and Golgi apparatus. *Journal of Molecular Biology*, *425*(12), 2208–2218. http://doi.org/10.1016/j.jmb.2013.03.024

III. Possner, D. D. D., Claesson, M., & Guy, J. E. (2015). Structure of the glycosyltransferase Ktr4p from *Saccharomyces cerevisiae*. *PLoS ONE*, *10*(8), e0136239. http://doi.org/10.1371/journal.pone.0136239

IV. Possner, D. D. D., Abate, F. & Guy, J. E. (2016). *Trichoplusia ni* ferritin: a warning of a recurrent crystallisation contaminant from the High Five expression strain. Manuscript.

V. Possner, D. D. D., Wigren E., Guy J. E. & Lindqvist Y. (2016). Recombinant production of human ERGIC-53 in insect cells yields assembly-competent protein. Manuscript.

CONTENTS

LIST OF ABBREVIATIONS

BEVS	Baculovirus expression vector system
CD	Circular dichroism
COP	Coat protein complex
CRD	Carbohydrate recognition domain
DLS	Dynamic light scattering
EM	Electron microscopy
ER	Endoplasmic reticulum
ERES	ER exit sites
ERGIC	ER-Golgi intermediate compartment
ERGIC-53	ER-Golgi intermediate compartment 53 kDa protein
Erv41p	ER vesicle protein 41 kDa
Erv46p	ER vesicle protein 46 kDa
F5F8D	Combined deficiency of factor V and factor VIII
IEX	Ion exchange chromatography
IMAC	Immobilised metal ion chromatography
Ktr4p	Kre-two related protein 4
LMAN1	Lectin mannose-binding 1
MCFD2	Multiple coagulation factor deficiency protein 2
MCS	Multiple cloning site
RMSD	Root mean square deviation
SAXS	Small-angle X-ray scattering
SEC	Size exclusion chromatography
SPR	Surface plasmon resonance

1 INTRODUCTION

1.1 GENERAL INTRODUCTION

In the evolution from the prokaryotic to the eukaryotic cell, several instrumental aspects of life changed fundamentally, and one of the most important was the development of specialized organelles that perform dedicated functions within the cell. These physiologically separated compartments, each surrounded by semi-permeable membranes that allow communication between the cytosol and the organelles, as well as direct communication between the different organelles, provide an increased level of order in the eukaryotic cell and give it the capacity to perform a range of complex cellular processes simultaneously.

The nucleus was the first organelle to be discovered, having been first observed by Antonie van Leeuwenhoek (1632 – 1723) in 1719, but it was only in the early 20th century, with the advent of our understanding of chromosomes and heredity, that its function became clear. Around the same time, in 1910, Camillo Golgi first observed what he termed the *internal reticular apparatus* under a light microscope. This organelle, which is now commonly referred to as the Golgi apparatus, is usually located close to the nucleus and in most eukaryotes it consists of several cisternae, a network of connected, membrane-enclosed compartments which appear as flattened disks in electron micrographs. The Endoplasmic Reticulum was not identified until much later, after development of the first electron microscopes. In 1945, Keith R. Porter, Albert Claude, Brody Meskers and Ernest R. Fullam, using a magnification of 1600x, first observed a network of membranes between the nucleus and the Golgi apparatus, which they termed *reticulum*, and which is now known as the Endoplasmic Reticulum.

These three organelles have been studied and characterised extensively since their respective discoveries, and an intricate functional and physical connectivity between them has become evident. Together, they not only govern the correct folding of proteins by providing a suitable environment, but also provide crucial enzymes regulating post-translational modifications, and a means of transporting the newly synthesised proteins to their final destinations. This organised system for the maturation, sorting and transport of newly-synthesised proteins is known as the secretory pathway.

The importance of the secretory pathway in the human cell is emphasised by the fact that several diseases are associated with failures in the folding and transport of proteins, such as different types of cancers and combined deficiency of blood coagulation factors V and VIII. Research conducted in the past few decades has significantly improved our understanding of the molecular mechanisms causing these diseases. However, in order to pinpoint the exact root of the aberrant molecular behaviour, every step of the pathway, and in particular each of the disease-causing proteins, must be studied in detail at the molecular level.

In this thesis we have applied X-ray crystallography together with biochemical and biophysical methods to characterise several different proteins involved in protein transport in

the early secretory pathway: ERGIC-53 is, together with its interaction partner MCFD2, responsible for the secretion of the blood clotting proteins factor V and VIII, and a number of other proteins. Erv41p has been shown to travel from the ER to the Golgi in vesicles, and has also been implicated in protein transport. Lastly, Ktr4p is involved in the glycosylation of secreted proteins. With these studies we hope to have provided new insights into the structures and functions of the proteins themselves, but also to have contributed towards a greater understanding of protein transport in the early secretory pathway.

1.2 THE EARLY SECRETORY PATHWAY

The secretory pathway – sometimes also called the exocytotic pathway - consists of several physiologically separated compartments, each of which provides a unique environment that is required for the correct folding, sorting, and post-translational modification of a newly synthesised protein. Recent estimates show that about a third of all synthesised proteins travel through the secretory pathway (Stolz & Wolf, 2010), and these are collectively referred to as secretory cargo proteins. After synthesis of the proteins on the surface of the ER, they are folded, receive post-translational modifications, and are sorted into different compartments.

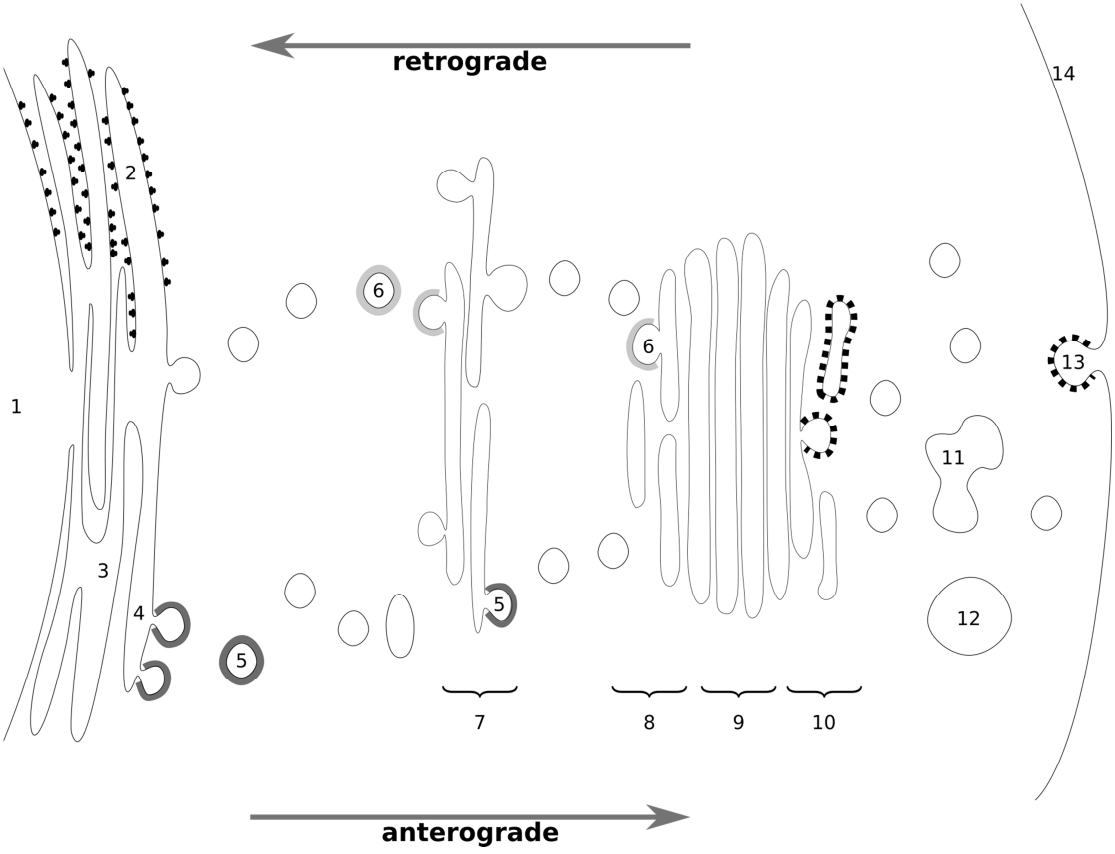

Figure 1: A schematic representation of the components of the secretory pathway. 1: Nucleus, 2: rough ER, 3: smooth ER, 4: ER exit sites; 5: COPII-coated vesicle, 6: COPI-coated vesicle, 7: ERGIC, 8: cis-Golgi network, 9: median Golgi-network, 10: trans-Golgi network, 11: endosome, 12: lysosome, 13: clathrin-coated vesicle 14: extracellular space

While some proteins reside in the ER, others exit it in vesicles and commence an anterograde journey along the secretory pathway towards the ER-Golgi Intermediate Compartment (ERGIC), the Golgi apparatus and ultimately lysosomes or secretory vesicles leading to the plasma membrane or the extracellular space (figure 1). The transport of proteins between the different organelles of the secretory pathway is directed by vesicles that bud off from the membrane loaded with cargo molecules and travel towards the target compartment, where, after fusion of the vesicle with the target membrane, the cargo is released. The formation and directionality of the vesicles is mediated by their coat proteins: COPI, COPII and clathrin each are responsible for the sorting of vesicles at various stages of the secretory pathway. The secretory pathway is directly linked to the endocytotic pathway, which allows for protein and nutrient uptake into the cell *via* endosomes budding from the plasma membrane.

The secretory pathway is a very dynamic system which is able to respond to the varying tasks posed to cells in different environments and, as such, it has the capability of rapidly increasing or decreasing the protein turnover according to the cell's demands (Kondylis *et al.*, 2009).

1.2.1 The Endoplasmic Reticulum

The endoplasmic reticulum was first observed in 1945, as the advancement of the electron microscope allowed for greater sample magnifications, and it was described as a "lace-like reticulum" by Porter *et al.* (Porter *et al.*, 1945). Later, this structure was termed the endoplasmic reticulum, a term derived from the words endon (gr. *ἔνδον*, inner, within), plasma (gr. *Πλάσμα*, something formed) and reticulum (lat. network). Since its discovery, the ER has been studied in detail and observed in almost every cell type, although it is notably absent in erythrocytes and spermatozoa. In cell types with a high protein turnover, such as pancreatic exocrine cells, the ER can take up a major portion of the cells volume. Functionally, the ER is responsible for folding of secretory proteins, quality control and post-translational modifications, as well as protein sorting. In addition, it is the organelle where the synthesis of fatty acids, the major components of membranes, and steroid hormones takes place. Furthermore, the ER acts as a storage compartment for calcium ions, which serve a central function in signal transduction. The free Ca^{2+} concentration in the ER has been measured at 0.5 – 1 mM, with another 1 – 5 mM bound to proteins (Bygrave & Benedetti, 1996), compared to nanomolar concentrations in the cytosol. In this sense, the ER is therefore topologically similar to the extracellular space, where the free Ca^{2+} concentration is in the low millimolar range (Clapham, 2007).

1.2.2 The ER-Golgi Intermediate Compartment

As its name implies, the ER-Golgi intermediate compartment (ERGIC) is situated between the ER and Golgi apparatus. It is a highly dynamic cluster of vesicles emerging from the ER and assembling into loosely connected tubular structures, and is therefore also referred to as vesiculotubular clusters (VTC). First described in 1984 (Saraste & Kuismanen, 1984), its existence was debated until a specific marker for the ERGIC was discovered in 1988 (Schweizer *et al*., 1988). The presence of this marker, a protein called p53, was later also confirmed by the identification of its rat homologue, now called p58/ERGIC-53 (Saraste *et al*., 1987; Lahtinen *et al*., 1992). Today, the ERGIC is recognised as a stationary site for the anterograde and retrograde sorting of proteins in ER-to-Golgi traffic (Appenzeller-Herzog & Hauri, 2006) and has been shown to differ biochemically from both the ER and the Golgi apparatus, specifically in its pH and calcium concentration. While the exact pH value of the ERGIC has so far not been determined, it is estimated to lie between the values encountered within the ER (pH 7.1 – 7.4) and the Golgi apparatus (pH 5.9 – 6.3) (Paroutis *et al*., 2004). The linear decrease of the pH value from the ER to the Golgi apparatus stands in contrast with the calcium concentration in each compartment. The calcium concentrations within both the ER and Golgi apparatus have been determined to be relatively high (0.5 – 1 mM), whereas the concentration within the ERGIC was below the detection limit (Pezzati *et al*., 1997). It is now believed that these physiological differences along the early secretory pathway provide a means for the sorting of proteins for onward transport towards the Golgi apparatus or retrieval back towards the ER.

1.2.3 The Golgi Apparatus

Camillo Golgi was the first to observe a basket-like network surrounding the nucleus in Purkinje cells in 1898 and he called it the “apparato reticolare interno”, or internal reticular apparatus. Today, this organelle is referred to as the Golgi apparatus. Until the 1960's, his description of the Golgi apparatus remained heavily debated, and the function of the proposed organelle remained obscure. As was the case with the discovery of the ER, it was the development of electron microscopy methods, together with autoradiography, which put the debate to rest. The Golgi apparatus serves two basic functions: it is the host organelle for further glycosylation of proteins, and it serves as another sorting hub for proteins travelling along the secretory pathway, directing them either along the anterograde direction towards the plasma membrane, or along the retrograde pathway back towards the ER. The Golgi is comprised of several tightly stacked cisternae which have an inherent directionality, the *cis*-side which faces the ER and transiently fuses with vesicles arriving from the ERGIC, the medial Golgi, and the *trans*-side which faces the plasma membrane. It is not a stable organelle, but rather exhibits a continuous flux of protein-loaded vesicles from the *cis*-Golgi network (CGN) towards the *trans*-Golgi network (TGN). The exact manner in which proteins traverse the cisternae, be it vesicular transport, percolation, or cisternal progression, is to this day still controversial (Pelham & Rothman, 2000; Patterson *et al*., 2008). In addition to

anterograde and retrograde protein sorting, the Golgi apparatus is responsible for O-linked glycosylation and extensive remodelling of the branched N-linked glycan, which proteins receive in the ER. The complexity and diversity of the final products are vast; the glycans are often branched and can be comprised of more than 200 sugar molecules, which, in addition, are potentially modified with phosphate, sulfate, acetate, or phosphorylcholine (Stanley, 2011). Finally, after this extensive posttranslational processing, secretory proteins exit the Golgi at the TGN and continue their journey to their final destinations.

1.3 PROTEIN TRANSLATION, FOLDING AND SORTING

All nascent proteins destined to enter the secretory pathway are guided towards the ER *via* a 16-30 amino acid long signal sequence at their N-terminus (Blobel & Dobberstein, 1975). This sequence often contains a stretch of hydrophobic amino acids in its centre which, during its synthesis, is recognised by the signal recognition particle (SRP), a ribonucleoprotein, leading the translating ribosome to the SRP-receptor located on the surface of the rough endoplasmic reticulum (Akopian *et al*., 2013). The signal peptide is subsequently translocated into the ER membrane through a complex of proteins, which is generally referred to as the translocon (Denks *et al*., 2014).

Within the ER lumen, the hydrophobic regions of the nascent peptide are often recognised and bound by proteins of the hsp70 and hsp40 families, which help to prevent the growing protein chain from uncontrolled aggregation and instead assist in protein folding (Ellgaard & Helenius, 2003). Additionally, the branched core oligosaccharide $GlcNAc_2Man_9Glu_3$ is covalently attached to the majority of proteins at asparagine residues within the consensus sequence -NXS/T-, where X is any amino acid except proline (Gavel & von Heijne, 1990). This reaction is performed by a membrane protein complex called oligosaccharyltransferase (OST) (Mohorko *et al*., 2011), and the attachment of the branched glycan also helps prevent aggregation of the folding-intermediates. The glycan furthermore acts as a maturation signal during the ensuing cyclic folding process. The glucosidases I and II sequentially trim the oligosaccharide of its terminal glucose moieties, producing $GlcNAc_2Man_9Glu_1$ (Grinna & Robbins, 1980), which is a substrate for the transmembrane protein calnexin (CNX) and its soluble paralog calreticulin (CRT) (Helenius *et al*., 1997). The folding intermediates also associate with the protein disulfide-isomerase Erp57, which aids in the formation of intramolecular disulfide bonds, *via* calnexin and calreticulin. During multiple rounds of association and dissociation from calnexin and calreticulin, the protein undergoes attempts to fold into its native structure and if the process succeeds, glucosidase II removes the last of the three glucose moieties, which allows the folded protein to exit the ER. In contrast, non-natively folded proteins are recognised and re-glucosylated by the UDP-glucose:glycoprotein glycosyltransferase (UGGT), which again allows interaction with calnexin/calreticulin (Parodi, 2000). This cycle continues until the protein succeeds in achieving its native state, or, if it fails to do so, is targeted for ER-associated degradation.

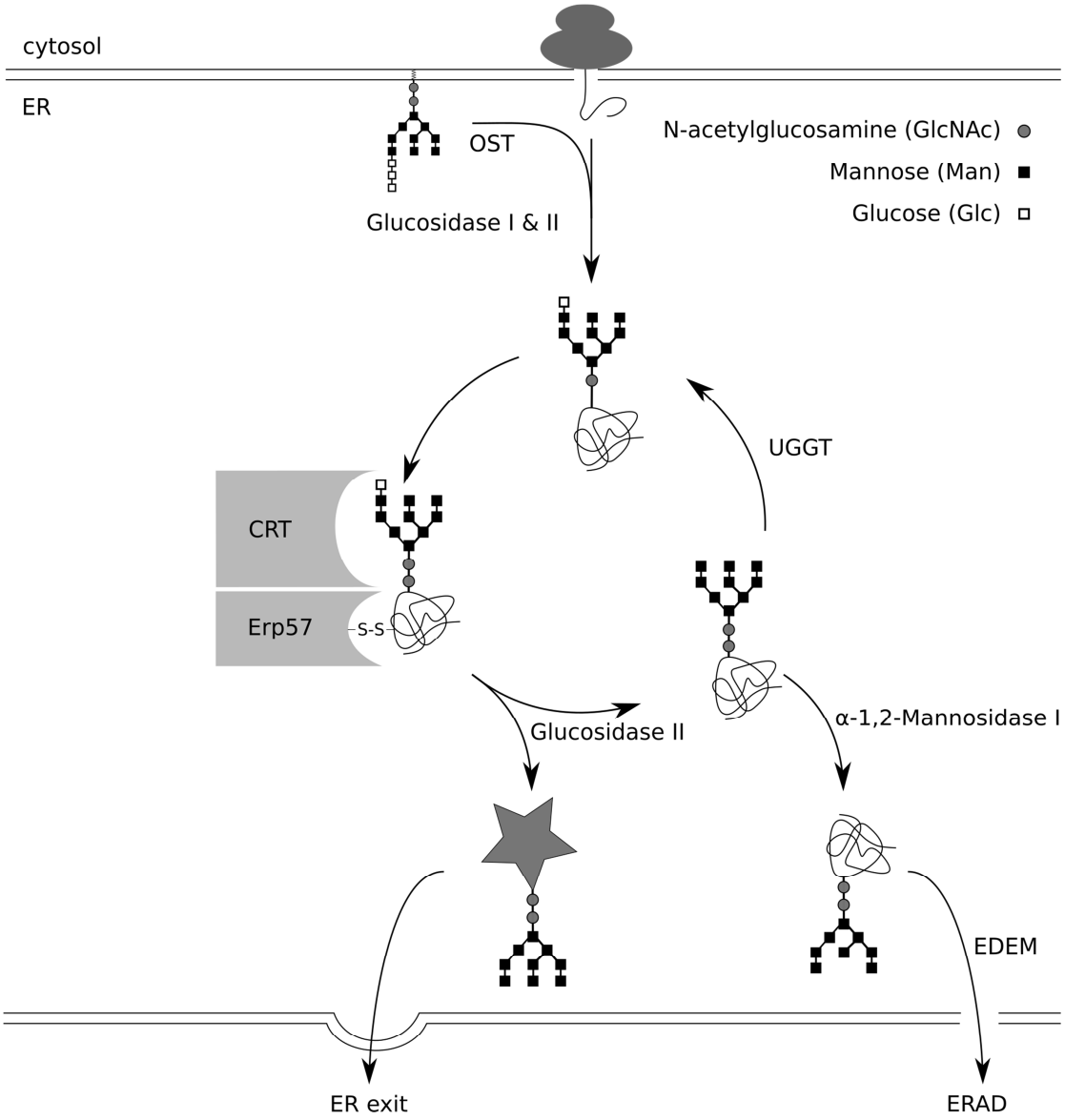

Figure 2: Schematic overview of the glycoprotein folding pathway. For clarity, calnexin has been omitted from the figure.

In this process, α-1,2-mannosidase I removes one terminal mannose moiety from the branched oligosaccharide, producing $GlcNAc_2Man_8$ (Tremblay & Herscovics, 1999), which interacts with the ER-degradation-enhacing 1,2-mannosidase (EDEM) (Kanehara *et al*., 2007). This enzyme is thought to remove the terminally misfolded protein from the calnexin/calreticulin cycle and to direct it, with the help of XTP-3B, OS-9, Sel1L and Hrd1, towards the retrotranslocation channel (Gauss *et al*., 2011). The retrotranslocation channel exports the protein through the ER membrane into the cytosol, where it is poly-ubiquitinylated and degraded by the proteasome (figure 2). The process of ER associated degradation (ERAD) is still very poorly understood, and the nature of the retrotranslocation channel is heavily debated (Nakatsukasa & Brodsky, 2008; Hampton & Sommer, 2012). Overall, these complex processes protect the cell from accumulating misfolded and dysfunctional proteins.

1.3.1 Protein Exit from the ER along the Secretory Pathway

After the functional integrity of a newly synthesised protein is established, it is sorted from ER resident proteins and moved to ER exit sites (ERES) (Budnik & Stephens, 2009), which are highly curved parts of the smooth ER (Okamoto *et al*., 2012). Even though the ERES are clearly distinguishable on electron micrographs (Bannykh *et al*., 1996), it remains unknown how their composition differs from other parts of the ER. The distinction of the ERES from the rest of the ER, however, serves a crucial function in protein sorting. This distinction may be established by the retention of incorrectly folded proteins to other sites of the ER *via* their binding to chaperones. Additionally, numerous cargo receptors, responsible for the recognition and transport of proteins approved for ER exit, are present at the ER exit sites, and they might provide another selective barrier for the ER exit of folding-intermediates by being incapable of binding proteins that are unfit for release. A third mechanism to uphold the integrity of the secretory pathway might be the prevention of protein aggregates being packed into budding vesicles. The vesicles are typically only ~50 nm in diameter, which would likely inhibit the exit of very large protein aggregates from the ER.

Following their localisation to the ERES, secretory proteins are packed into COPII-coated vesicles. The COPII-coat (COP refers to the coat protein complex) consists of an inner layer, which is comprised of the Sec23-Sec24 heterodimer, recruited to the ER-membrane by the small GTPase Sar1, and an outer layer, comprised of the heterotrimeric Sec13-Sec31 complex (Jensen & Schekman, 2011). The fission of the COPII-coated extrusion from the ER-membrane to form an intact vesicle appears to be governed by the Sec13-Sec31 complex (Matsuoka *et al*., 1998). At the target membrane, the force enabling the fusion of the vesicle with the membrane is exerted by SNARE and SM (Sec1/Munc18-like) proteins (Südhof & Rothman, 2009).

1.3.2 Protein Sorting

The balance between anterograde and retrograde vesicle traffic maintains the integrity of the different organelles in the early secretory pathway both by keeping the membranes of the organelles intact, as well as by retrieving proteins that have escaped from the ER. While anterograde vesicle traffic is governed by COPII-coated vesicles, retrograde traffic is maintained by COPI-coated vesicles. Soluble ER-resident proteins generally contain a C-terminal ER-retention signal (KDEL in humans, HDEL in yeast), and, *via* the interaction of the signals with the COPI-associated KDEL-receptors, are retrotranslocated towards the ER in the case of escape (Munro & Pelham, 1987; Pelham *et al*., 1988). The transport of membrane-anchored proteins is guided by their N- or C-terminal sorting sequences; a di-basic (KK, RR) motif interacts directly with COPI and is thereby retrieved to the ER (Jackson *et al*., 1990), while proteins containing a more diverse di-acidic (e.g. YENE, DID) or di-hydrophobic (e.g. FF, LL, YL, FY) motif are sorted to the ERGIC and Golgi apparatus *via* COPII vesicles (Barlowe, 2003).

1.4 TRANSPORT RECEPTORS AND THEIR CARGO

The active shuttling of soluble proteins within COP-coated vesicles requires the mediation of the interaction of the cargo to the COP-coat *via* transmembrane receptors. A growing number of these transport receptors, as well as their specific cargo molecules, have been identified over the course of the last two decades (table 1), but their cargo recognition processes remain poorly understood in many cases. Most of the early work on the identification and cargo specificity of transport proteins was performed in the yeast *S. cerevisiae*, and major contributions to this research area are provided by the work performed in the laboratories of C. Barlowe and H.-P. Hauri.

Receptor	Species	Cargo	Reference
Multispanning transmembrane receptors			
Erv29p	yeast	gpαf, CPY, PrA	(Belden & Barlowe, 2001; Caldwell *et al*., 2001; Otte & Barlowe, 2004)
Erv26p	yeast	P-ALP, Ktr3p	(Bue *et al*., 2006; Bue & Barlowe, 2009)
Erv14p	yeast	Axl2p, Sma2p	(Powers & Barlowe, 2002; Nakanishi *et al*., 2007)
KDEL receptor	mammals	-KDEL sequence	(Munro & Pelham, 1987)
Protein cornichon homolog 4	mammals	GPCR	(Sauvageau *et al*., 2014)
p24 proteins			
Emp24p-Erv25p	yeast	Gas1p, Suc2p	(Muñiz *et al*., 2000)
ER vesicle proteins			
Erv41p/Erv46p	yeast	Ktr4p, Gls1, Fpr2	(Noda *et al*., 2014; Shibuya *et al*., 2015)
ERGIC1/ERGIC2/ERGIC3	mammals	unknown	
L-type lectins			
Emp46p/Emp47p	yeast	Ssp120	(Margulis *et al*., 2015)
ERGIC-53/MCFD2	mammals	FV, FVIII, catC, catZ, A1AT, IgM, Mac2-bp, nicastrin, FGFR3, SUMF1	see text
ERGL	mammals	unknown	
VIP-36	mammals	unknown	
VIPL	mammals	unknown	

Table 1: Protein sorting receptors of the early secretory pathway and their cargo proteins.

One of the first transport receptors to be identified in yeast was the conserved, multispanning transmembrane receptor Erv29p, which has been isolated from COPII-coated vesicles and found to be directly required for the export of a number of cargo proteins from the ER to the Golgi apparatus (Belden & Barlowe, 2001). Following this discovery, Erv26p and Erv14p were identified. Erv26p is non-essential, but it has been found to cycle between the ER and the Golgi apparatus and to interact with the proteins of the COPI-complex (Bue & Barlowe, 2009). Similarly, Erv14p has been shown to cycle between the ER and Golgi, but this protein is involved in the packaging of cargo into COPII-coated vsicles (Powers & Barlowe, 2002).

The members of the p24 protein family are type I membrane proteins with a lumenal GOLD (Golgi dynamics) domain, a transmembrane helix and C-terminal tail containing binding

motifs to COPI and COPII proteins (Fiedler *et al*., 1996). They are highly conserved from yeast to mammals and were first identified as abundant constituents of COPI and COPII-coated vesicles (Stamnes *et al*., 1995), and shortly after their discovery their role in protein transport and sorting was established (Schimmöller *et al*., 1995). As a knock-out of all eight family members in the yeast *S. cerevisiae* showed no visible difference in protein transport (Springer *et al*., 2000), the precise role of these proteins remained puzzling for a substantial period of time and it was only more than a decade after their discovery that their function as receptors for glycophosphatidylinositol (GPI)-anchored proteins was revealed (Takida *et al*., 2008).

In humans, the ERGIC family is comprised of four members, namely ERGIC-53, ERGL (ERGIC-53 like), VIP-36 and VIPL (VIP-36 like). All four members of the family are type I transmembrane proteins comprised of an N-terminal lumenal part, one transmembrane helix and a short, C-terminal tail exposed to the cytoplasm. On the lumenal side they all contain a carbohydrate recognition domain (CRD) belonging to the class of L-type lectins, which is defined by the structural similarities and the glycan-binding abilities of its members. Structurally, the CRD consists of a β-sandwich structure with a concave and a convex sheet, and two conserved metal-ion binding sites proximal to the carbohydrate binding site. The members of this family of L-type lectins bind their cargo proteins by recognising high-mannose carbohydrates, but in some cases *via* protein-protein interactions as well (Appenzeller *et al*., 1999; Carrière *et al*., 1999; Satoh *et al*., 2007).

1.4.1 ERGIC-53

In 1988, Schweizer *et al*. produced an antibody with the goal of specifically labelling a tubulovesicular compartment near the *cis* side of the Golgi apparatus of a human intestinal cell line, and were able to show that this monoclonal antibody bound to a specific membrane protein of 53 kDa (Schweizer *et al*., 1988). Independently, Saraste and co-workers identified a protein of 58 kDa by a similar approach, namely by the generation of polyclonal antibodies against a fraction of the Golgi apparatus of rat pancreas cells (Saraste *et al*., 1987). Furthermore, a third group isolated an intracellular protein in search for lectins by mannose-column chromatography (Pimpaneau *et al*., 1991). It was later revealed that these three newly identified proteins were in fact the same protein, which is now termed ERGIC-53 or, less commonly, LMAN1 (Arar *et al*., 1995; Itin *et al*., 1996).

The discovery of ERGIC-53 was therefore a key factor for the identification of the ER-Golgi intermediate compartment (ERGIC), and this is where it is most abundant. It is also found, albeit in lower abundance, at ER exit sites (ERES) and in the first *cis*-cisternae of the Golgi apparatus (Schweizer *et al*., 1988; Chavrier *et al*., 1990; Klumperman *et al*., 1998), where it is associated with COPII and COPI proteins (Tisdale *et al*., 1997; Wendeler *et al*., 2007). Since its discovery, ERGIC-53 has been extensively used to study the transport routes of the early secretory pathway.

ERGIC-53 is now known to be a non-glycosylated type I integral membrane protein, which assembles into homodimers and homohexamers immediately after its synthesis (Schweizer *et al*., 1988; Lahtinen *et al*., 1992). Sequence comparisons have revealed that it contains a lumenal domain resembling the carbohydrate recognition domain (CRD) of plant leguminous lectins (Fiedler & Simons, 1994) at its N-terminus. An oligomerisation domain is located to the C-terminus of the CRD, and this has been proposed to be instrumental for the oligomerisation of the protein (Lahtinen *et al*., 1999; Zhang *et al*., 2006; Zheng *et al*., 2010). The oligomerisation domain is followed by a short linker, connecting the protein to the ER membrane, which contains two cysteine residues (Cys-466 and Cys-475). On its cytosolic side, a short tail with the very C-terminal KKFF motif contains information for the recycling of the protein.

The CRD is the best characterised part of ERGIC-53, and its modes of carbohydrate recognition and release are well understood. After the first crystal structure of the CRD from the rat protein had been determined (Velloso *et al*., 2002), it became clear that the domain contains two Ca^{2+}-binding sites (Velloso *et al*., 2003) and a large carbohydrate binding site able to accommodate complex glycans (Zheng *et al*., 2013; Satoh *et al*., 2014). It was therefore hypothesised that ERGIC-53 functions as a glycoprotein receptor involved in the export of glycoproteins from the ER. The discovery that the correct localisation of the thiol-proteases Cathepsin C and Cathepsin Z to the lysosomes depends on ERGIC-53 (Vollenweider *et al*., 1998) further strengthened the hypothesis. Since then, an additional, albeit small, number of proteins dependent on ERGIC-53 for their ER export have been identified. Most prominently, the blood coagulation factors FV and FVIII rely on ERGIC-53 for their secretion from the cell (Nichols *et al*., 1998), and a disruption of their interaction with the transport receptor leads to the bleeding disorder F5F8D (combined deficiency of factors FV and FVIII). Patients suffering from this disorder exhibit a concentration of the blood coagulation factors in the blood stream that is only 5 – 30% of the level in healthy patients. The other, thus far identified, cargo proteins of ERGIC-53 are the serine protease inhibitor α1-antitrypsin (Nyfeler *et al*., 2008b), the antibody IgM (Mattioli *et al*., 2006; Cortini & Sitia, 2010), the galectin binding protein Mac2-bp (Chen *et al*., 2013), nicastrin, a component of the γ-secretase complex (Morais *et al*., 2006), the fibroblast growth factor receptor FGFR3 (Lievens *et al*., 2008), and the sulfatase modifying factor I (Fraldi *et al*., 2008).

The interaction of the ERGIC-53 CRD with the N-linked glycans of the cargo molecules depends on two factors, namely the Ca^{2+}-concentration and the pH of the local environment. While glycan binding is promoted at a slightly basic pH and at a high Ca^{2+}-concentration, as it is the case within the ER, substrate release is triggered by a slightly acidic pH and a lower Ca^{2+}-concentration, an environment provided by the ERGIC (figure 3a) (Appenzeller-Herzog *et al*., 2004).

While the CRD of ERGIC-53 has been studied in detail, the oligomerisation domain of the protein is less well described at present and there is some confusion over the mechanism of

oligomerisation. Two studies have shown that two cysteine residues present in the membrane-proximal linker region mediate the oligomerisation of the protein (Appenzeller *et al.*, 1999; Lahtinen *et al.*, 1999), but a different study contradicts this claim (Neve *et al.*, 2005). As the hexamerisation of ERGIC-53 is a prerequisite for its function (Nufer *et al.*, 2003), further insights into the structure and oligomerisation properties of this domain are needed in order to better understand the roles of the oligomeric states of ERGIC-53 in protein transport.

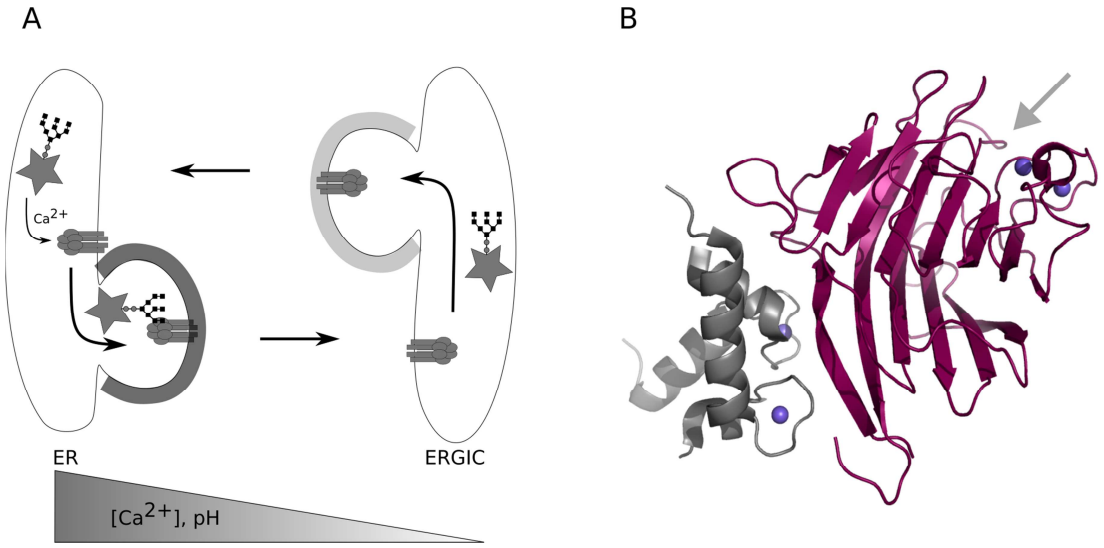

Figure 3: (a) Schematic overview of protein transport facilitated by ERGIC-53. The $GlcNAc_2Man_9$ glycan of correctly folded proteins is recognised by hexameric ERGIC-53 in a Ca^{2+}-dependent manner and the multi-protein complex is packed into COPII-coated vesicles at ER exit sites. After fission from the ER, the vesicle is transported to the ERGIC where the two membranes are fused. The lower Ca^{2+}-concentration within the ERGIC triggers the release of the cargo molecule from ERGIC-53, which is subsequently recycled to the ER *via* COPI-coated vesicles. **(b)** The crystal structure of the ERGIC-53 CRD (magenta) in complex with MCFD2 (grey). Calcium-ions are coloured in violet and the carbohydrate recognition site is indicated by an arrow (Wigren *et al.*, 2010).

1.4.2 MCFD2

A connection between ERGIC-53 and MCFD2 was first discovered in 2003, when Zhang *et al.* identified MCFD2 as a second agent causing F5F8D (Zhang *et al.*, 2003). They discovered that mutations in the protein led to a phenotype which was indistinguishable from the one caused by mutations in ERGIC-53, suggesting it could be acting as a co-transporter for these cargo proteins. It was then shown that an MCFD2 knock-down resulted in no adverse effect on the localisation of Cathepsin C and Z, indicating that MCFD2 is disposable for their transport along the secretory pathway (Nyfeler *et al.*, 2006). It was only in 2013 that another protein whose secretion depends on MCFD2 as well as ERGIC-53, namely the Mac2 binding protein, was discovered by a protein fragment complementation assay (Chen *et al.*, 2013).

In studies designed to investigate whether ERGIC-53 and MCFD2 require each other for correct localisation in the cell, siRNA-based knock-down experiments showed that the absence of ERGIC-53 led to the secretion of MCFD2, whereas a knock-down of MCFD2 had no effect on the localisation of ERGIC-53 (Nyfeler *et al.*, 2006). This result, together with the fact that MCFD2 contains no C-terminal KDEL retrieval signal, lead to the conclusion that the correct localisation of MCFD2 is dependent on its interaction with ERGIC-53.

MCFD2 is a soluble, 14 kDa protein found in the ER and ERGIC. The structure of MCFD2 has been determined by NMR and it revealed that, while the protein is intrinsically disordered, it partially folds upon the incorporation of two calcium ions into the two calcium-binding EF-hand motifs at its C-terminus. The N-terminus, however, remains disordered even in the Ca^{2+}-bound state (Guy *et al.*, 2008), and the function of this disordered N-terminal region is not currently known. The C-terminal EF-hand domains mediate the interaction with the ERGIC-53 CRD, and the crystal structure of the ordered region of MCFD2 in complex with the ERGIC-53 CRD has been determined (Nishio *et al.*, 2010; Wigren *et al.*, 2010). MCFD2 binds to the ERGIC-53 CRD at a site opposite to the carbohydrate recognition site (figure 3b), and the study also showed that most F5F8D-causing mutations lie at the interface of the two proteins, suggesting that a disruption of the protein complex is responsible for the reduced plasma concentrations of the two coagulation factors.

1.4.3 The ER Vesicle Protein Complex Erv41p/Erv46p

While a great deal is already known about the role of ERGIC-53 as a glycoprotein transport receptor, comparatively little was known about the other main target of this thesis, the Erv41p/Erv46p complex, when the study was initiated. In 2001, Otte *et al.* first described the Erv41p/Erv46p protein complex in yeast. By performing a reconstituted COPII vesicle budding assay, aiming at the identification of novel proteins of the early secretory pathway, combined with mass-spectrometry experiments, both Erv41p and Erv46p were identified, and they were named based on their cellular localisation and apparent mass. Further experiments determined that the two proteins are conserved across species, that they co-localise to the ER and Golgi apparatus and that they are present within the cell as a protein-protein complex. Knock-out experiments showed that their expression levels in yeast are interdependent, and that yeast cells lacking either protein are viable but susceptible to cold-shock. Further, cell-free, experiments suggested that the protein complex is involved in membrane fusion (Otte *et al.*, 2001). It is now however suspected that this effect is caused by indirect consequences of the double knock-out mutants (Shibuya *et al.*, 2015).

Since their initial discovery, the protein complex consisting of Erv41p and Erv46p has been investigated more thoroughly. Both proteins are integral membrane proteins, consisting of a large, lumenal domain flanked by one transmembrane helix on each side and short cytoplasmic tails. While both Erv41p and Erv46p contain a di-hydrophobic sorting motif, IL

and FT respectively, for the anterograde transport from the ER to the Golgi apparatus, only Erv46p contains a di-acidic KKXX motif signalling for the retrieval of the protein from the Golgi apparatus to the ER (Otte & Barlowe, 2002).

The function of the Erv41p/Erv46p protein complex was initially not well understood, and suggestions ranged from an involvement in the sorting of proteins into transport vesicles, to a role in the retention and/or retrieval of transport machinery to the early secretory pathway, a role in the transport of lipids or, finally, a role in the posttranslational maturation of secretory proteins such as protein folding or glycosylation (Otte *et al*., 2001). It was only in 2014 that Noda *et al*. first showed by domain-switching experiments that the ER-exit of the putative mannosyltransferase Ktr4p relied directly on the Erv41p/Erv46p complex, thereby confirming a role for the complex in protein sorting (Noda *et al*., 2014). Since then, another study has identified the Erv41p/Erv46p protein complex as a new retrograde receptor for the retrieval of non-HDEL bearing ER-resident proteins. Erv41p/Erv46p double knock-out strains led to the secretion or mislocalisation of the Mannosyl-oligosaccharide glucosidase Gls1 and the Peptidyl-prolyl *cis-trans* isomerase Fpr2, two ER-resident proteins (Shibuya *et al*., 2015). Moreover, this study showed that the interactions between the Erv41p/Erv46p complex and Gls1 are regulated by pH. The interaction is strong in the slightly acidic environment provided by the Golgi apparatus and substantially weaker at the slightly basic pH provided by the ER. The conclusion drawn from the combination of these experiments is therefore that the Erv41p/Erv46p complex binds escaped ER-resident proteins within the Golgi apparatus and retrieves them to the ER, where they are released.

In humans, homologous proteins of Erv41p and Erv46p can be found as members of a three-component protein complex consisting of ERGIC1, ERGIC2 and ERGIC3. While Erv41p and Erv46p share 30% and 41% sequence identity with ERGIC2 and ERGIC3, respectively, ERGIC-1 lacks a homologous protein in yeast. ERGIC1 shares the domain topology of the other members of the complex, and interacts directly with ERGIC3 (Breuza *et al*., 2004) but is mainly localised to the ERGIC. While the human proteins are much less well characterised, studies have suggested an involvement of all three proteins in different types of cancer. The *ERGIC1* gene has been shown to be highly expressed in prostate cancer tissue and suggested to be a potential drug target (Vainio *et al*., 2012), the *ERGIC2* gene has been found to upregulate interferon-β in a prostate cancer cell line (Kwok *et al*., 2006), and the *ERGIC3* gene has been suggested as a potential biomarker for lung cancer (Lin *et al*., 2015).

1.4.4 Ktr4p, a Cargo of the Erv41p/Erv46p Complex

The first protein to be identified as cargo of the Erv41p/Erv46p complex was the putative glycosyltransferase Ktr4p, a member of the Kre2/Mnt1 family of in *S. cerevisiae*. Our knowledge of this family of proteins dates back to 1991, when Hill *et al*. first cloned and sequenced the *KRE2* gene from the yeast *S. cerevisiae* and found that it shares significant sequence identity with two previously identified proteins, termed KTR1 and Yur1p, and the

common domain topology consistent with type II membrane proteins, in concert with the location of the most conserved regions, suggested a related function. The three proteins were therefore classified as a family (Hill *et al*., 1992), and since their first classification, six additional proteins have been identified as members of the Kre2/Mnt1-family.

The enzymatic activities of six members of the family have been studied, and, while five have been shown to possess glycosyltransferase activity, one enzyme, namely Ktr6p, has been shown to instead have mannosylphosphate transferase activity (Wang *et al*., 1997; Jigami & Odani, 1999). Kre2/Mnt1 is the best characterised member of the family. It catalyses the addition of the second and third α-1,2-linked mannose residues in linear O-linked oligosaccharides, and has also been shown to be involved in the synthesis of the outer chains of N-linked oligosaccharides (Hill *et al*., 1992; Lussier *et al*., 1996, 1997). A crystal structure of the lumenal, catalytically active, domain of Kre2/Mnt1 has also been determined, and the structure shows the lumenal domain to be comprised of a single, Rossmann-fold-type domain with an adjacent second domain of α/β-structure, and thus to belong to the GT-A fold family of glycosyltransferases, rather than the GT-B fold family which has two, less tightly-packed, Rossmann-like domains with the active site in a cleft between them (Lobsanov *et al*., 2004).

When initiating our studies, the knowledge of Ktr4p was very limited in comparison to Kre2. Sequence alignments showed that it exhibits a 32% identity to Kre2p/Mnt1p , with the residues constituting the active site, including those responsible for the binding of the GDP-nucleotide and the Mn^{2+}-ion, being largely conserved. This comparison implied that the function of Ktr4p is similar to that of Kre2p/Mnt1p, but no *in vitro* enzyme activity assays have been performed prior to the experiments described in this thesis. Gene knock-out studies have been performed in the filamentous fungi *Beauvaria bassiana* and these led to growth defects, a decrease in cell wall components, reduced tolerance to stress and lower virulence (Wang *et al*., 2014), reinforcing the potential importance of the Ktr4p protein.

2 AIM OF THE THESIS

The general aim of this thesis was the characterisation of proteins or protein-protein complexes involved in glycoprotein transport or processing in the early secretory pathway. More specifically, the aims were

- to elucidate the three-dimensional structure of Erv41p and study its interaction with Erv46p,
- to enzymatically and crystallographically characterise Ktr4p, and
- to study the structure of the full lumenal part of ERGIC-53 and to investigate its oligomerisation properties.

3 RESULTS AND DISCUSSION

3.1 EXPRESSION, PURIFICATION AND X-RAY STRUCTURE DETERMINATION OF ERV41P (PAPERS I & II)

The membrane protein complex comprised of Erv41p and Erv46p in yeast initially attracted our interest when it was first hypothesised to be involved in the transport of cargo proteins between the ER and the Golgi apparatus. While only a handful of studies on the yeast protein complex were published at the time, even fewer articles described the role of the human orthologues ERGIC1, ERGIC2 and ERGIC3. Most strikingly, our bioinformatic analyses showed little sequence similarity between these proteins and any protein structure deposited in the PDB, which severely hindered prediction of their functions. In order to better understand their potential functions, we therefore initiated a project to investigate the structures of these proteins, as well as the interactions between them.

3.1.1 Expression and Purification of Erv41p and Erv46p

As extensive trials to produce soluble Erv41p and Erv46p in *E. coli* in our laboratory had previously failed, we decided to instead utilise the baculovirus expression vector system for the recombinant expression of these proteins. In this system, ovarian cells from either of two organisms, *Spodoptera frugiperda* or *Trichoplusia ni*, are transfected with a viral bacmid, engineered to contain the gene of interest. This gene is cloned at the location of the dispensable polH gene, which, in the wildtype virus, encodes for the protein polyhedrin, a major structural component of baculovirus occlusion bodies. In this eukaryotic expression host it is possible to direct expression of recombinant proteins through the secretory pathway to the extracellular medium, thus employing the folding and quality control mechanisms in the host cell ER, by attaching a signal sequence to the N-terminus. We therefore chose to include the signal sequence of honey bee mellitin - the active component in bee venom, and a highly expressed and efficiently secreted protein – at the N-terminus of each construct, and this was followed by a His_6-purification tag. For this work, we decided to omit the N- and C-terminal transmembrane helices of both Erv41p and Erv46p, as well as their cytoplasmic tails, and aimed to produce only the lumenal domain of each protein. Several constructs were

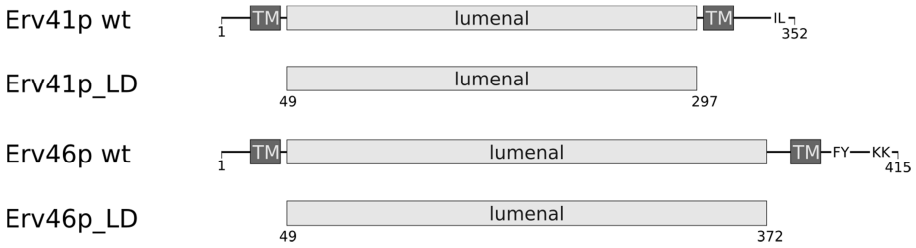

Figure 4: Schematic representation of the domain topology of Erv41p and Erv46p. Both wildtype proteins are comprised of a single lumenal domain flanked by a transmembrane helix (TM) on each side and short, cytoplasmatically exposed, N- and C-termini. While Erv41p only contains a C-terminal anterograde transport motif, Erv46p contains both antero- and retrograde transport motifs at its C-terminus. The constructs studied in this work, comprised only of the lumenal domain of each protein, also are depicted.

tested for expression and, ultimately, the longest constructs, spanning residues 49 - 297 and 49 - 372, for Erv41p and Erv46p, respectively, were used in this study (figure 4).

Following the transfection of *S. frugiperda* Sf9 cells with the bacmid, and confirmation of production and secretion of the recombinant protein, each protein was produced in larger scale by infecting *T. ni* High Five cells with the respective virus. Both proteins could then be purified from the expression medium by standard techniques.

Using these methods, we were able to successfully express and purify Erv41p_LD and Erv46p_LD of suitable qualities and in suitable quantities to proceed with crystallographic studies.

3.1.2 Structure Determination of Erv41p

Crystallisation trials with Erv41p were performed using commercially-available screens, and resulted in one condition producing diffracting crystals. Optimisation of this condition was successful, and X-ray diffraction data to 2.0 Å resolution was recorded. Since no phase information from a homologous or structurally similar protein was available, the single anomalous dispersion method was employed to experimentally obtain the information for initial phasing of the structure. To achieve this, crystals of the native protein were soaked with a range of metals or heavy metal containing compounds, and many soaked crystals were screened for diffraction on synchrotron beamlines. Ultimately, only crystals soaked with $YbCl_3$ diffracted to a high enough resolution (2.7 Å) and provided sufficient anomalous signal to enable the placement of the strongly diffracting heavy atoms and make initial model building possible. At this stage, elongated sections of electron density, clearly corresponding to β-strands, could be identified and a poly-alanine model was built to fit the density. After several rounds of phase combination, model building and refinement, this model was of high enough quality to be used as a molecular replacement search model for the higher-resolution dataset recorded from the native crystals, and model building was completed with this dataset. The final model at 2.0 Å resolution was refined to R/R_{free} values of 16.5% and 20.6%, respectively, and comprises the full lumenal domain of Erv41p, with the exception of three flexible loops.

	Erv41p_LD/$YbCl_3$	Erv41p_LD native
Data collection and phasing		
Wavelength [Å]	1.3852	0.9334
Resolution range [Å]	48.24-2.7 (2.85-2.7)	48.14-1.99 (2.10-1.99)
Space group	$P2_1$	$P2_1$
Unit cell parameters		
a, b, c [Å]	48.77, 76.65, 65.47	49.78, 76.93, 65.11
α, β, γ [°]	90, 98.49, 90	90, 104.7, 90
Total reflections	77603	127463
Unique reflections	12968	31843
Multiplicity	6 (6.2)	3.9 (3.2)
Completeness [%]	98.4 (97.2)	99.1 (94.2)
Mean I/σI	18.6 (7.3)	19.41 (4.41)
Anomalous completeness [%]	94.9 (93.5)	
Anomalous multiplicity	3.1 (3.1)	
Wilson B-factor [$Å^2$]	34.8	23.1
R_{merge}	0.064 (0.19)	0.05 (0.27)
Phasing Figure of Merit	0.432	
Refinement		
R/R_{free} [%]		16.5/20.6 (18.8/23.0)
Number of atoms		3868
macromolecules		3628
water		240
Protein residues		432
Average B-factor, including TLS contribution [$Å^2$]		36.40
macromolecules		33.4
solvent		39.8
R.M.S. deviations from ideal		
RMSD bond lengths[Å]		0.009
RMSD bond angles [°]		1.21
Ramachandran plot		
Ramachandran favored [%]		95
Ramachandran outliers [%]		0

Table 2: Summary of data collection, phasing and refinement statistics for Erv41p_LD. Statistics for the highest-resolution shell are shown in parentheses.

3.1.3 The Structure of Erv41p

The lumenal domain of Erv41p consists of a tightly packed and twisted β-sandwich, whose concave and convex faces are comprised of an eight-stranded β-sheet (sheet A) and a six-stranded β-sheet (sheet B), respectively. On one side of the β-sandwich the β-strands are connected by loops, some of which contain short helices, while on the other side, short N- and C-terminal β-strands protrude from the core. Since the native protein contains transmembrane helices at the N- and C-termini, this would be the membrane-proximal side (figure 5a,b).

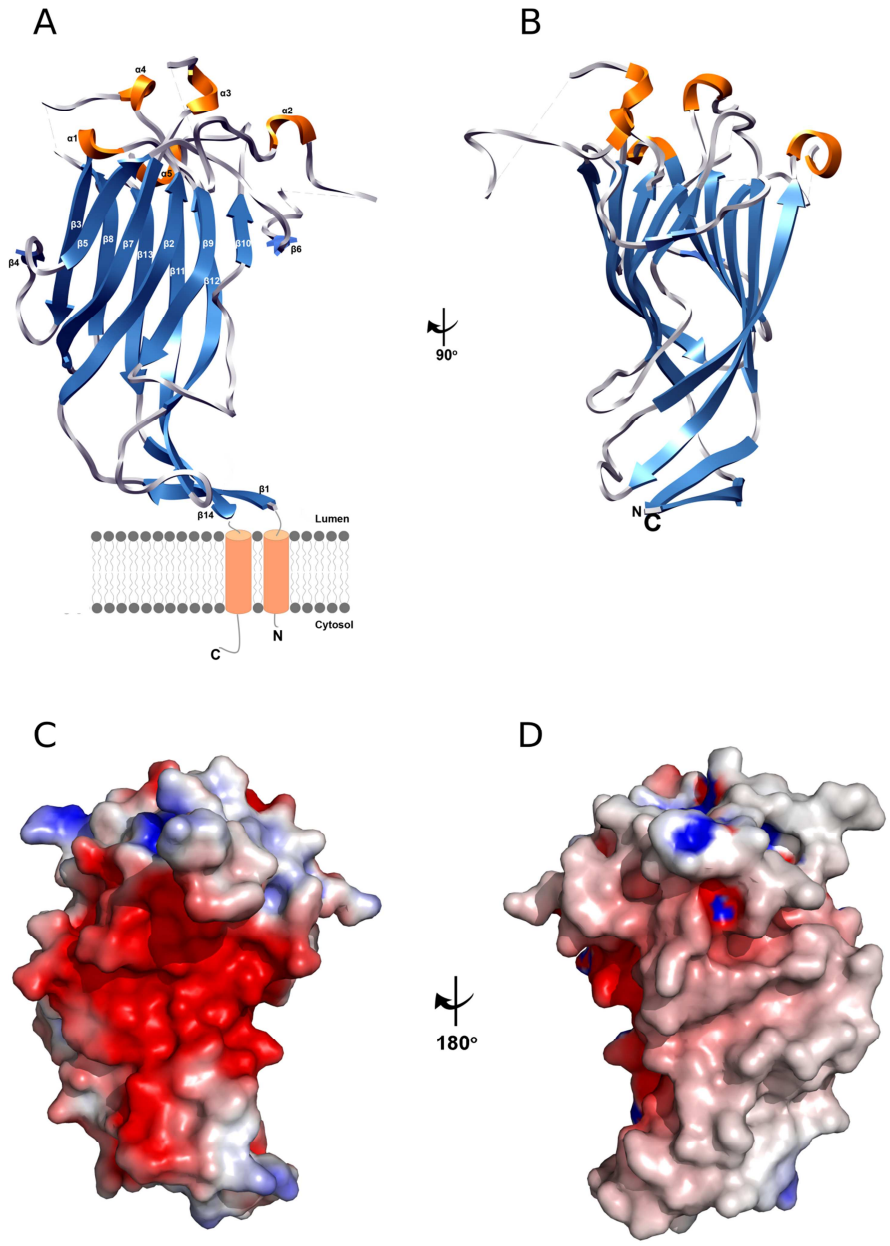

Figure 5: The structure of Erv41p_LD. **(a)** The structure of Erv41p_LD is represented as a cartoon and coloured according to secondary structure elements, with α-helices in gold, β-sheets in blue and loops in grey. The secondary structure elements are labelled, and missing sections of loops are indicated by dashed lines. A schematic representation of the ER-membrane shows the position of the lumenal domain in the context of the full protein. **(b)** A representation of the structure rotated by 90º around a vertical axis. **(c)** The electrostatic potential of the Erv41p_LD surface. The orientation of the protein is equivalent to (a). Negative potential is denoted in red and positive potential is denoted in blue. The maps were calculated using the APBS plug-in to PyMOL and are contoured at the 4-kT level. **(d)** A view of the structure rotated by 180º around a vertical axis.

Calculation of the electrostatic surface potential of the Erv41p lumenal domain reveals a large, negatively charged area covering almost the entire sheet B, while on the opposite sheet A, only singular charged residues stand out from the largely neutral surface. The size of this

negatively charged area led us to the hypothesis that it is likely to be physiologically relevant, possibly as a site of interaction with other proteins or smaller ligands (figure 5c,d).

Searching the PDB for structurally similar proteins identified, among others, acid-sensing ion channels and other structures of proteins with carbohydrate binding modules. However, due to the low similarity of even the closest structural relatives with root mean square deviations (RMSD) of 4.4 Å and 3.7 Å, respectively, no meaningful conclusion about the function of Erv41p can be drawn from these comparisons.

3.1.4 Expression, Purification and Crystallisation of Erv46p

As with Erv41p, exhaustive trials to express Erv46p in *E. coli* yielded only insoluble protein and the gene was therefore cloned for expression in insect cells. A construct lacking the transmembrane helices, but covering the full lumenal domain, with an N-terminal honeybee mellitin signal sequence was engineered and insect cells were transfected with the resulting bacmid. The protein could be purified from the culture medium *via* its N-terminal His_6-tag, but upon removal of this purification tag a large fraction of the protein formed soluble aggregates, which could not be recovered. Nevertheless, some monomeric protein remained in solution, and several commercial crystallisation screens were set up to identify suitable conditions for the crystallisation of Erv46p_LD. Microcrystals could indeed be produced, however these crystals did not diffract at synchrotron beamlines. After optimisation of the condition to produce larger and diffracting crystals failed, further screening for different conditions was performed. Three-dimensional crystals were obtained at this point, however they were later revealed to be of a contaminant protein (described in further detail in paper IV).

3.1.5 Interaction Studies of Erv41p and Erv46p

In order to gain further insight into the strength and mechanism of interaction between Erv41p and Erv46p, we performed surface plasmon resonance (SPR) experiments with the two proteins. Due to the tendency of Erv46p_LD to form soluble aggregates however, our studies were ultimately limited to showing the direct and concentration-dependent interaction of the two proteins *in vitro*.

The experiments were performed on a Biacore 3000 instrument (GE Healthcare, Uppsala, Sweden) at 25 ºC. Erv41p_LD was immobilised on a CM5 chip by amine-coupling according to standard procedures (GE Healthcare) in acetate buffer, pH 4, leading to a response of 700 RU, and Erv46p_LD was used as the analyte and injected onto the chip in varying concentrations and in a random order at a flow-rate of 100 µl/min. An unmodified flow-cell was used as a reference surface. After each analyte injection, the surface was regenerated with short pulses of 5 mM NaOH. Evaluation of the sensorgrams was performed using

BIAevaluation 4.1 (GE Healthcare), and the reference signal was subtracted from the raw data (figure 6).

Since Erv41p and Erv46p have been observed as a part of a larger complex of 200 - 400 kDa (Welsh *et al*., 2006), the possibility that as-yet unidentified adapter proteins maintain an indirect interaction between Erv41p and Erv46p had to be considered. In addition, the transmembrane helices might be involved in the assembly of the complex. While the SPR experiments confirm that the interaction between the lumenal domains of the two proteins is direct, the low response upon binding points towards a weak and possibly transient interaction between the two proteins.

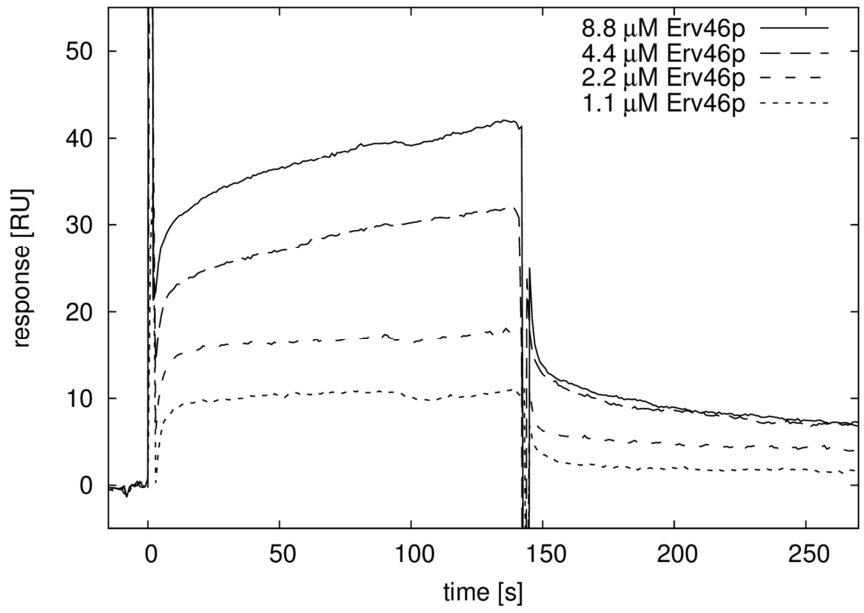

Figure 6: SPR sensorgram of Erv41p_LD interacting with Erv46p_LD. Erv41p_LD was immobilised on a CM5 chip and Erv46p_LD was used as the analyte. Four injections of analyte concentrations ranging from 1.1 µM to 8.8 µM were performed in random order and reference data from an unmodified flow-cell was subtracted. The results indicate a direct binding of Erv46p_LD to Erv41p_LD.

3.1.6 Expression and Crystallisation of ERGIC1

ERGIC1, the third member of the human protein complex, which lacks a yeast homologue, was cloned, expressed and purified in the same manner as Erv41p, and attempts to crystallise the protein were undertaken. Unfortunately, as for Erv46p, only very small crystals could be obtained, which diffracted poorly, and optimisation of these crystals was not successful despite extensive efforts.

3.1.7 Co-expression of ERGIC1, ERGIC2 and ERGIC3

In addition to the expression of Erv41p and Erv46p, we also attempted to co-express the tripartite complex of the homologous human proteins in insect cells. This was facilitated by the use of the co-expression system developed for the BEVS by Fitzgerald *et al*., (Fitzgerald *et al*., 2006), which provides two plasmids, each containing two multiple cloning sites (MCS). This allows one to make use not only of the polyhedrin promotor, but also the promotor of the gene encoding for p10 - another viral protein dispensable for infectious virus formation which is replaced by the second gene of interest. The two plasmids can be recombined using the Cre-Lox technology. After engineering a plasmid containing all three human proteins, transfection of insect cells and protein production is carried out as usual.

Using these techniques, we successfully co-expressed the lumenal domains of ERGIC1, ERGIC2 and ERGIC3, which could then be co-purified in small amounts. Unfortunately, the very low yields resulting from the co-expression of the proteins prevented further studies to investigate the complex.

3.2 THE CRYSTAL STRUCTURE OF KTR4P FROM *S. CEREVISIAE* (PAPER III)

In yeast, the family of Kre2/Mnt1 glycosyltransferases consists of nine members, which have been shown to play roles in N- and O-linked glycosylation of newly synthesised proteins. Several of the members of this family have been characterised and shown to catalyse similar and partially redundant mannosyltransferase reactions (Lussier *et al*., 1999), and one member has been shown to be a mannosylphosphate transferase (Wang *et al*., 1997; Jigami & Odani, 1999). These enzymes are mostly located to the Golgi apparatus, and in 2014, the location of one member of the family, Ktr4p, was shown to directly depend on the presence of the ER vesicle proteins Erv41p and Erv46p (Noda *et al*., 2014). This study brought Ktr4p to our attention and we decided to produce recombinant Ktr4p with the aim of both studying the protein itself, and of potentially performing interaction studies together with Erv41p and Erv46p.

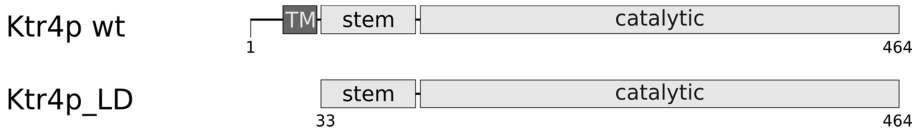

Figure 7: Schematic representation of the domain topology of Ktr4p. The wildtype protein is comprised of a short, cytoplasmatically exposed tail at the N-terminus, followed by a transmembrane helix, a lumenal stem domain and a large catalytic domain. The construct studied in this work, comprised only of the lumenal part of the protein is also depicted. TM: transmembrane helix.

3.2.1 Expression and Purification of Ktr4p

For the production of the putative glycosyltransferase Ktr4p, attempts to express the protein in *E. coli* succeeded. Twelve constructs of the lumenal domain of the protein, with the longest one ranging from residues Asn-33 to Tyr-464 (figure 7), were each cloned into the pNIC28-Bsa4 plasmid, which contains a His_6-purification tag N-terminal of the MCS. The actual cloning was carried out by the Protein Science Facility at Karolinska Institutet, although all subsequent work was performed in our laboratory. Given the eukaryotic origin of the protein and its subcellular location in the Golgi apparatus, combined with our previous experience of expressing these proteins in a bacterial host, the *E. coli* Rosetta-gami 2 strain was chosen as the expression host. The Rosetta-gami 2 strain combines the features of the Rosetta strain, allowing for enhanced disulfide bond formation, and the Origami strain, enhancing the expression of eukaryotic proteins containing rare codons scarcely used in *E. coli*. We were pleasantly surprised to learn that all correctly cloned constructs led to soluble protein using this *E. coli* strain. After the successful small-scale expression trials, the production of all tested constructs could be scaled up and approximately 20 mg of pure protein could be obtained from a 1 litre *E. coli* culture by following standard expression and purification protocols. While the purification of some constructs resulted in the presence of a degradation product, the longest construct, ranging from residues Asn-33 to Tyr-464 could be isolated to

purity without signs of degradation. We therefore chose to focus only on this construct in subsequent studies. The integrity of the protein preparation was investigated with circular dichroism (CD), which confirmed that the protein was folded and showed that its secondary structure content was mostly α-helical, and dynamic light scattering (DLS), which indicated a monodisperse sample.

3.2.2 Crystallisation and Structure Determination

Crystallisation trials using the sitting drop method were set up in 300 nl drops with commercially available screens, and crystals were harvested directly from one condition of the screening plates. These crystals diffracted to 2.2 Å resolution, and data were collected at beamline ID23-1 of the ESRF. The structure was solved by molecular replacement, using the already available structure of Kre2/Mnt1 (PDB id 1S4N, (Lobsanov *et al*., 2004)) as a search model, and after model building refined to R/R_{free} values of 0.16/0.20, respectively.

	Ktr4p_LD apo structure	**Ktr4p_LD-GDP complex**
Data collection		
Beamline	ESRF ID23-1	ESRF ID23-2
Wavelength [Å]	0.8726	0.97241
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit cell parameters		
a, b, c [Å]	60.19, 102.37, 156.91	61.215, 102.62, 162.65
α, β, γ [º]	90, 90, 90	90, 90, 90
Resolution [Å]	50-2.2 (2.28-2.20)	50-1.9 (1.94-1.90)
R_{merge}	0.127 (0.651)	0.104 (0.867)
Mean I/σI	13.8 (3.0)	15.5 (2.3)
Mean $CC_{1/2}$	0.997 (0.856)	0.999 (0.833)
Completeness [%]	99.7 (97.0)	99.1 (98.9)
Multiplicity	7.8 (8.0)	9.8 (9.2)
Number of reflections	387144 (34874)	790203 (41691)
Number of unique reflections	49596 (4381)	80667 (4508)
Wilson B-factor [$Å^2$]	10.5	17.5
Refinement		
Resolution [Å]	50 - 2.2	50 - 1.9
R/R_{free}	0.163/0.202	0.156/0.191
Number of non-hydrogen atoms (protein)	7081	7278
Mean B value [$Å^2$]	24.96	26.51
Number of waters	498	577
R. M. S. deviations from ideal		
RMSD bond lengths [Å]	0.018	0.20
RMSD bond angles [º]	1.678	1.832
Ramachandran plot		
Residues in favoured regions [%]	97.8	98.2
Residues in allowed regions [%]	2.2	1.8
Residues in disallowed regions [%]	0	0

Table 3: Summary of data collection and refinement statistics of Ktr4p_LD and Ktr4p_LD bound to GDP and Mn^{2+}. Statistics for the highest-resolution shell are shown in parentheses.

By incubating the protein with both GDP and $MnCl_2$ for 1h, followed by the same crystallisation procedure as for the apo-protein, crystals of Ktr4p_LD bound to both Mn^{2+} and GDP could be obtained, and diffraction data to 1.9 Å were recorded at beamline ID23-2 of the ESRF. The previously obtained apo-structure of Ktr4p_LD was used as a search model for molecular replacement. During refinement of the protein chain, clear density corresponding to a GDP-molecule became apparent in the active site of both chains in the asymmetric unit. The structure was ultimately refined to R/R_{free} values of 0.16/0.19, respectively (data collection and refinement statistics are shown in table 3).

3.2.3 The Crystal Structure of Ktr4p

The structure of Ktr4p is comprised of a central seven-stranded β-sheet, which is surrounded by several α-helices and flanked by two β-strands (figure 8a). With this structure it belongs to the glycosyltransferase subfamily GT-A, and it aligns to Kre2/Mnt1, the closest homologue with a determined structure, with an RMSD of 1.2 Å over 235 residues of the catalytic domain. This striking similarity, despite the relatively low sequence identity of 32%, is also apparent in the active site of the enzymes, which led to the hypothesis that Ktr4p binds GDP, coordinated by a Mn^{2+} ion, two ligands that are both present in the structure of Kre2/Mnt1. This hypothesis was tested by incubating the protein with GDP and $MnCl_2$ prior to crystallisation, and both ligands were indeed clearly identifiable in the active site of the resulting structure (figure 8b). In contrast, the soaking experiments performed with a relatively high concentration of methyl-α-mannoside led to only very weak density visible at the proposed binding site for the acceptor, which would suggest a very weak binding of the monosaccharide at this position. Similarly weak density was observed when analogous experiments were performed with Kre2/Mnt1 (Lobsanov *et al.*, 2004). This weak binding could either be an indication of a non-native acceptor substrate, which possibly decreases its binding affinity to the active site substantially, a very fast dissociation rate of the acceptor substrate, or a combination of the two.

Upon comparison of the apo-structure with the GDP-complex, the only striking differences are observed in the active site, with the main rearrangement being the displacement of Arg-142 to allow the binding of GDP in the active site. The preformed nature of the active site is relatively unusual in glycosyltransferases; it is more common that conformational rearrangements of loops are observed upon ligand binding. The loops covering the active site of the enzyme are thought to sequester the active site from the solvent and potentially assist in product release (Unligil & Rini, 2000).

3.2.4 Activity assay

Since the activity of Ktr4p had not been previously studied and had been only speculated to be that of a glycosyltransferase, experiments to confirm its function were performed using a

coupled malachite-green assay. In this experiment, the glycosyltransferase reaction of Ktr4p (figure 8c) is assayed *via* GDP, the proposed leaving product of Ktr4p. GDP is further hydrolysed by the nucleotidase ENTPD3/CD39L3, giving rise to free inorganic phosphate, which can subsequently be detected and quantified using the malachite green reagent. This activity assay was performed using GDP-mannose as the sugar donor and three potential acceptor substrates were tested, namely mannose, α-1,2-mannobiose and methyl-α-mannoside. The choice of these donor- and acceptor-substrates was based on suggestions from sequence identity to Kre2/Mnt1 and commercial availability. Our results show that Ktr4p indeed is an active mannosyltransferase and the most efficient mannosyltransfer was observed for methyl-α-mannoside, while the observed reaction rates for mannose and α-1,2-mannoside were very low (figure 8d).

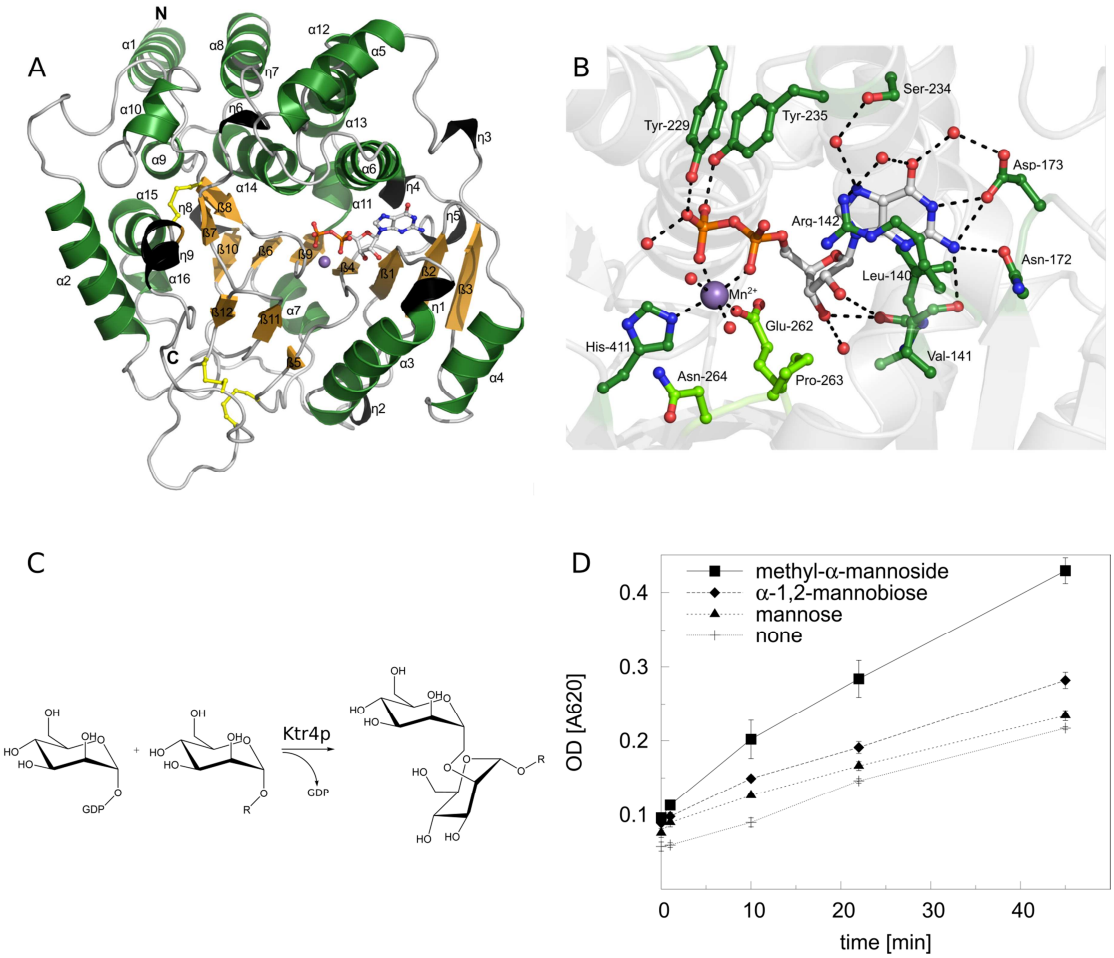

Figure 8: The crystal structure and enzymatic activity of Ktr4p_LD. **(a)** The structure of Ktr4p_LD in complex with GDP and Mn^{2+}. α-helices are coloured in green, 3_{10}-helices in black and β-strands in orange, and all secondary structure elements are numbered. The GDP is shown in ball-and-stick representation, as are the cysteines forming disulfide-bonds. **(b)** Ball-and-stick representation of GDP (white) and Mn^{2+} (purple) located in the active site and residues involved in their binding (green). **(c)** The glycosyltransfer reaction catalysed by Ktr4p. **(d)** Activity of Ktr4p_LD. The enzyme is active using methyl-α-mannoside (■) as acceptor substrate, and the signal observed using α-1,2-mannobiose (♦) and D-mannose (▲), respectively, is comparable to the background reading in the absence of acceptor substrate (+). The blank reading, measured in the absence of enzyme, has been deducted from all experimental readings.

These three acceptor substrates might only represent a fraction of the native substrate of Ktr4p. While our results clearly prove the mannosyltransferase activity of the enzyme, further experiments using the biological acceptor substrate, possibly attached to a target protein or lipid, might lead to further insight into the selectivity and activity of the protein. The identification of this substrate was however a task beyond the scope of this study for several reasons. Generally, the substrates of glycosyltransferases are tremendously hard to identify, and intricate knowledge of their roles in a specific pathway is required. This in turn would demand knowledge of the individual glycosylation remodelling steps of the substrate, leading to a Catch-22 situation. While an *in silico* approach to identify classes of substrates based on the protein's coding region is available for bacteria (Sánchez-Rodríguez *et al.*, 2014), in most cases the only readily available option is a microarray-based glycan screen. For this, a collection of glycans is immobilised on a microarray and binding of a protein in question to specific substrates can be tracked by antibodies.

We have undertaken this approach and screened for binding of Ktr4p_LD in the presence of Mn^{2+} at a pH of 6.5, the physiologial pH within the Golgi apparatus (Paroutis *et al.*, 2004), to a selection of over 600 natural and synthetic mammalian glycans in collaboration with the Consortium for Functional Genomics. The results of this assay were however negative, as no binding of Ktr4p_LD to any of the screened glycans was observed. It therefore has to be considered that Ktr4p might indeed not interact with complex glycans. Other possible explanations for the lack of an observed signal are however manifold: the interactions could be weak and transient and not give rise to a detectable signal. This might be caused by non-native experimental conditions, such as a too high or too low pH-value, improper ionic strength of the solvent or the lack of a co-factor. While the early steps of N-linked glycosylation in the ER of yeasts resemble those of humans, later stages of glycan remodelling, especially of O-linked glycans, differ greatly between the two organisms (Van den Steen *et al.*, 1998). It is therefore possible that the chosen array of mammalian glycans, which does not cover the complete human glycome, does not incorporate the actual substrate of Ktr4p from yeast.

3.3 FERRITIN FROM THE EXPRESSION HOST *T. NI* IS A COMMON CRYSTALLISATION CONTAMINANT (PAPER IV)

In our laboratory, the application of the baculovirus expression vector system for the production of eukaryotic proteins localised to the early secretory pathway has proven to be successful in many cases, and a substantial number of proteins that had previously proven unamenable for expression in *E. coli* strains could be produced using this technology. For the expression of secretory proteins, an N-terminal honeybee mellitin secretion signal was added to the protein sequence, followed by a His_6-purification tag and a TEV cleavage site. For large-scale production of these proteins, the BTI-TN5B1-4 (High Five) insect cell strain, a cultured cell line originated from the ovarian cells of the cabbage looper *Trichoplusia ni*, was used. This cell line had been previously reported to produce higher expression yields, especially for secreted proteins (Krammer *et al.*, 2010; Wilde *et al.*, 2014), an observation confirmed in our studies, as the High Five strain routinely increased expression yields up to 10-fold compared to the Sf9 cell line.

The recombinant proteins were purified by standard techniques, small-scale crystallisation screens were set up, and in the cases of Erv46p and the ERGIC-53 4H/MCFD2-ΔN complex, small crystals were obtained (figure 9b). In the case of the Erv46p protein preparation, crystals grown in five different conditions diffracted to ~2.2 Å resolution without optimisation. For the ERGIC-53 4H/MCFD2-ΔN complex, the crystals could reproducibly be obtained and increased in size after drops of larger volumes were set up, finally leading to diffraction up to ~4 Å resolution.

As molecular replacement search models for both proteins were available, this method was employed in order to solve the structures. It however never led to any convincing solutions, and rather to the suspicion that a protein contaminant had crystallised instead of the target proteins. In the case of Erv46p, this contaminant was readily identified by querying the PDB

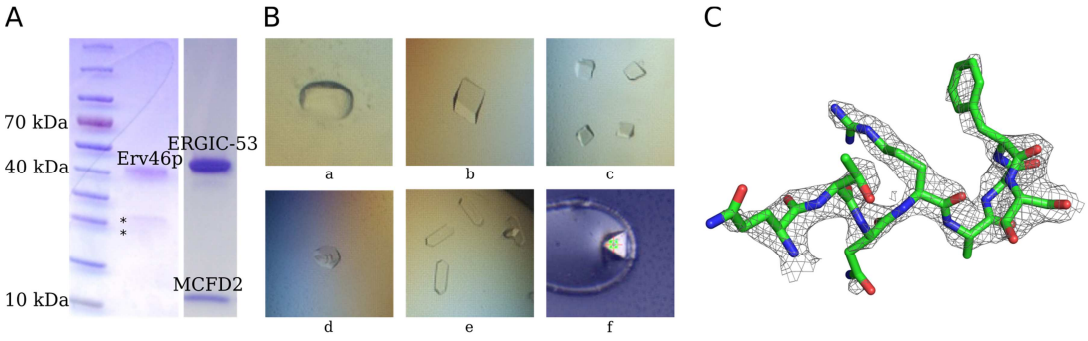

Figure 9: (a) SDS-PAGE analysis of purified Erv46p_LD and hexameric ERGIC-53 4H in complex with MCFD2-ΔN. Asterisks indicate the molecular mass of the ferritin heavy and light subunit. **(b)** Crystals of ferritin grown in six different conditions. The crystals labelled a-e were obtained from contaminated Erv46p preparations, and the one labelled f from ERGIC-53 preparations. **(c)** A section of the unbiased Fo-Fc omit map after molecular replacement, confirming that attempts to crystallise Erv46p_LD instead resulted in *T. ni* ferritin crystals. A section of the *T. ni* ferritin structure is shown in green and the Fo-Fc omit map, which was produced by running refinement after deleting this section of the structure from the model, is depicted as grey mesh.

for similar unit cell dimensions, which led to the previously-published structure of ferritin from *T. ni*. The structure could subsequently be solved using this molecule as a search model, and it was refined to R/R_{free}-values of 0.24/0.26, respectively, and the quality of the electron density confirmed beyond doubt that the crystals indeed were of *T. ni* ferritin (figure 9c).

In the case of the complex of ERGIC-53 4H and MCFD2-ΔN however, no protein with similar unit cell parameters could be found in the PDB and the structure solution process was therefore more cumbersome. Ultimately though, with our previous experience in mind, an attempt to solve the structure using *T. ni* ferritin as a search model for molecular replacement was undertaken and the structure could readily be solved and refined to R/R_{free}-values of 0.30/0.34, respectively. In both cases, no extensive structure refinement was undertaken as a higher-resolution structure of *T. ni* ferritin is already available (Hamburger *et al*., 2005).

The iron-storage protein ferritin of insects is a spherical 24-mer consisting of twelve heavy (26 kDa) and twelve light (24 kDa) chains. It is mostly found in the rER, and ferritin crystals can even be observed in secretory vesicles by electron microscopy (Locke & Leung, 1984; Nichol *et al*., 2002). In *D. melanogaster* fly extract, ferritin was found to be a major constituent, representing 0.8% of the total protein content (Li, 2010). Clearly, ferritin is highly abundant in baculovirus infected cells as well, and it co-elutes from Ni-NTA Sepharose beads with the His_6-tagged target protein (figure 9a). The fact that it is still present in the sample after removal of the His_6-purification tag from the target protein by incubation with TEV protease, followed by a reversed IMAC step, is surprising. However it might be attributed to its large quantities; weak binding of large amounts of ferritin to a Ni-NTA matrix might lead to removal of a large portion in the flow-through, but some ferritin could still be retained during the purification. Similarly, during the reversed Ni-NTA purification step, ferritin might be found in the flow through. Erv46p_LD was further purified by SEC and the fractions eluting at the volume corresponding to ~400 kDa were pooled and concentrated. While the exact molecular mass of *T. ni* ferritin remains unknown, it can be estimated to ~600 kDa. In the case of the ERGIC-53 4H/MCFD2-ΔN complex, the protein peak eluting at a volume corresponding to ~100 kDa was collected and re-injected onto the SEC column, yielding a fraction of hexameric protein eluting at a volume corresponding to ~300 kDa. This fraction was then incubated with MCFD2-ΔN in the presence of $CaCl_2$ and once again analysed by SEC, which clearly showed the presence of a stable complex eluting at a volume corresponding to ~360 kDa.

As the shapes of both Erv46p and the ERGIC-53 4H/MCFD2-ΔN complex might be non-globular, their elution volumes might differ more profoundly from their theoretical molecular masses. This could be an explanation for the presence of ferritin in Erv46p preparations, it does however not hold in the case of the ERGIC-53 4H/MCFD2-ΔN complex, as the latter protein sample was originally obtained from the peak eluting at 100 kDa.

In addition to its persistent presence during protein preparation, *T. ni* ferritin seems very prone to crystallisation in a number of different conditions and at very low concentrations, presumably aided by its high internal symmetry.

Altogether, we conclude that *T. ni* ferritin poses a problem when working with High Five cells as an expression host, and that protein preparations originating from High Five cells should be meticulously inspected for the presence of ferritin at an early stage of experimentation.

3.4 INVESTIGATING THE OLIGOMERISATION PROPERTIES OF ERGIC-53 (PAPER V)

While the carbohydrate recognition function of ERGIC-53 has been well studied (Velloso *et al*., 2002, 2003; Wigren *et al*., 2010; Zheng *et al*., 2013), the mode of oligomerisation provided by the membrane proximal oligomerisation domain of the protein remains more obscure. It has been established that ERGIC-53 is only functional as a hexamer (Nufer *et al*., 2003), but we do not yet understand why this is the case, or how the hexamer is built up. Experiments aimed at elucidating the structure of the hexameric form, and the mechanisms of its interactions with the co-receptor MCFD2 and with cargo, are therefore of key importance to increase our understanding of how ERGIC-53 functions as a cargo transport receptor in the early secretory pathway. However, despite the efforts of several research groups, it has not previously been possible to produce oligomerisation-competent ERGIC-53 protein in a recombinant host. We therefore set out to address this problem.

3.4.1 Expression and Purification of ERGIC-53

While the expression of the CRD of ERGIC-53 in *E. coli* was feasible, longer constructs including the full oligomerisation domain or only parts of it produced either insoluble protein or insufficient yields, and the expression of the isolated oligomerisation domain led to similar results. Only a construct comprising the CRD and three of the four helices belonging to the oligomerisation domain (ERGIC-53 3H) could be produced as soluble protein, and this was in minute amounts.

Two constructs of ERGIC-53, one including the full lumenal part (ERGIC-53 4H+L) and one truncation variant lacking the membrane-proximal linker (ERGIC-53 4H), were therefore prepared for expression trials in insect cells. Crucially, the ERGIC-53 4H+L construct includes the membrane-proximal linker, which contains two cysteine residues (Cys-466 and Cys-575) that are the subject of an ongoing debate. Several studies have suggested that these cysteines form intermolecular disulfide-bonds and trigger the trimerisation of ERGIC-53 dimers (Appenzeller *et al*., 1999; Carrière *et al*., 1999; Lahtinen *et al*., 1999), while others have suggested that they do not play a crucial role in the assembly of hexameric ERGIC-53 (Neve *et al*., 2005). Comparisons of these two constructs would allow for the further study of the role of the two cysteine-residues *in vitro,* and we have therefore focused our work on these two (figure 10).

Each construct was cloned into a plasmid containing a honeybee mellitin secretion signal followed by a His_6-purification tag N-terminal of the MCS. The bacmid preparation, insect cell transfection and large scale production of the constructs were performed using the same protocols as for the production of the ER vesicle proteins, and soluble protein could be obtained for all tested constructs.

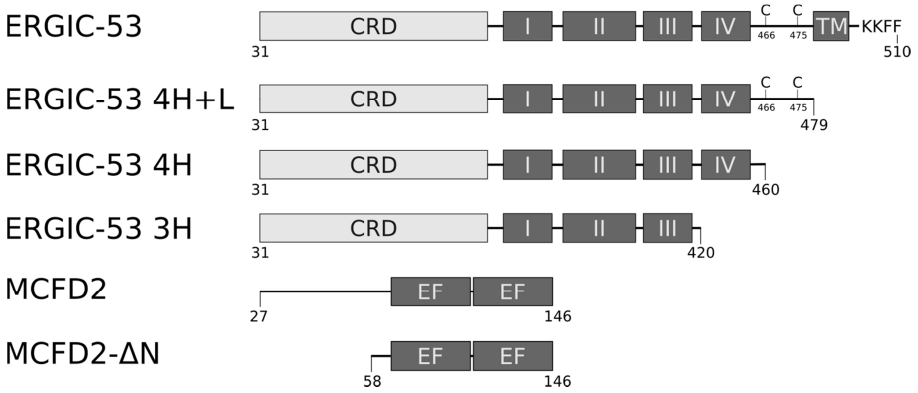

Figure 10: Schematic representation of the domain topologies of ERGIC-53 and MCFD2. ERGIC-53 is a 53 kDa protein comprised of a lumenal carbohydrate recognition domain (CRD) at its N-terminus and an oligomerisation domain predicted to consist of four α-helices (I - IV). This domain is followed by a short region linking the lumenal part of the protein to the ER membrane, which includes two cysteine residues (Cys-466 and Cys-475). On the cytoplasmic side, ERGIC-53 exposes a short C-terminal tail including the sorting motifs for both antero- and retrograde transport. The different C-terminal truncation variants of ERGIC-53 used in this study are represented below.
Wildtype MCFD2 is a soluble, 14 kDa protein consisting of a disordered N-terminal region, followed by two EF-hand domains. The MCFD2-ΔN construct is depicted below the wildtype.

To purify the different constructs, the growth medium was separated from the cells by centrifugation and the protein isolated from the medium by immobilised metal affinity chromatography (IMAC) *via* the N-terminal His_6-tag, followed by size exclusion chromatography (SEC). This protocol resulted in pure protein, as judged by SDS-PAGE. Both protein constructs consistently eluted as two distinct species from the SEC column and, by comparing their elution volumes to a standard curve produced based on a set of proteins

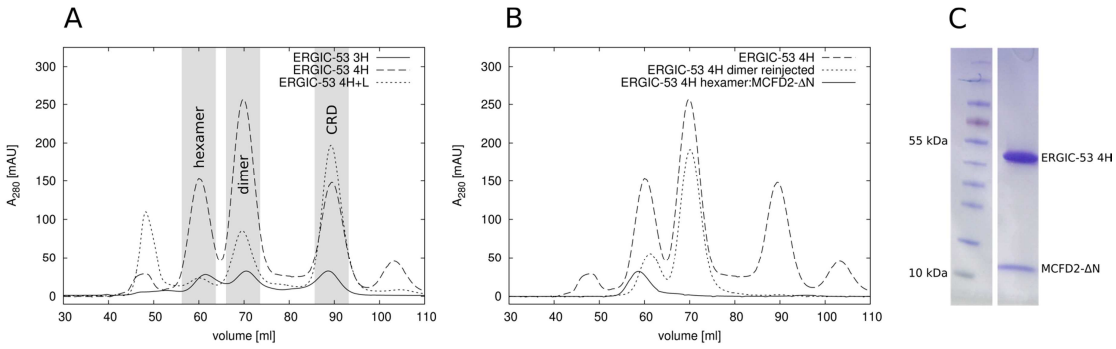

Figure 11: Purification of different ERGIC-53 constructs and complex formation with MCFD2-ΔN. **(a)** Size exclusion chromatograms of ERGIC-53 4H+L, ERGIC-53 4H and ERGIC-53 3H obtained from the final purification step on a Superdex S200 16/60 column (GE Healthcare). The three constructs elute at volumes corresponding to a hexameric and a dimeric species, and the isolated CRD is present as a degradation product. **(b)** SEC analysis of dimeric ERGIC-53 4H. Reinjection of the purified dimer leads to the formation of a hexameric protein. After incubation of this fraction with MCFD2-ΔN, a shift towards a volume corresponding to a larger molecular mass is observed, proving the stable formation of the complex. **(c)** SDS-PAGE analysis of hexameric ERGIC-53 4H after complex formation with MCFD2-ΔN and SEC.

with known molecular weight, the masses of the two species could be calculated to approximately 100 kDa and 300 kDa, respectively, corresponding to the theoretical masses of ERGIC-53 dimers and hexamers. This was also the case for the ERGIC-53 3H construct produced in *E. coli* (figure 11a).

To increase the purity of the dimeric ERGIC-53 4H, anion exchange chromatography (IEX) was performed and again, two distinct fractions eluted from the IEX column. These were later identified by DLS to correspond to monomers and dimers. In contrast, when performing the same experiment with hexameric ERGIC-53 4H, no dissociation of the protein was observed.

3.4.2 The Oligomerisation of ERGIC-53

The two ERGIC-53 oligomers were further investigated by SEC, native PAGE and glutaraldehyde crosslinking experiments. Both hexameric and dimeric species of ERGIC-53 4H were re-injected into the SEC column and the resulting elution profiles showed that, while the hexamer neither dissociated into smaller oligomers nor formed higher oligomeric species, the re-injected dimer eluted at the volumes of both hexamer and dimer, indicating that the dimers are competent of forming a stable hexamer (figure 11b).

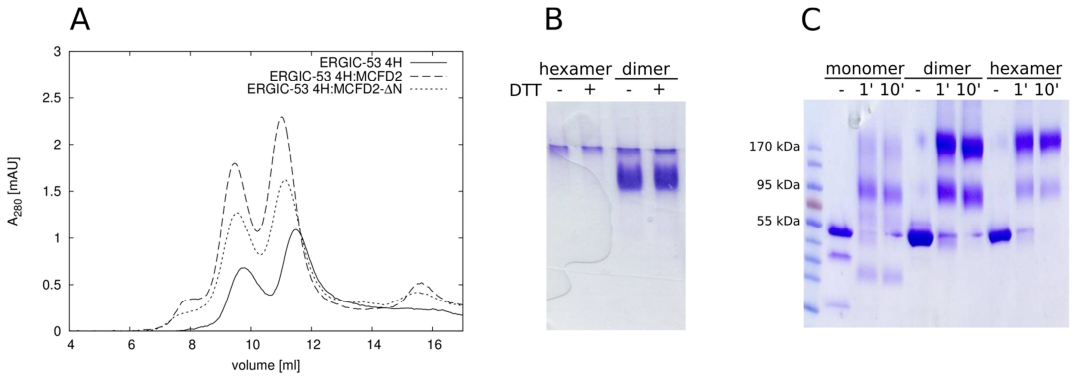

Figure 12: Analyses of the oligomerisation behaviour of ERGIC-53. **(a)** SEC runs of ERGIC-53 4H alone and after incubation with MCFD2 and MCFD2-ΔN, respectively, on a Superdex S200 10/300 column (GE Healthcare). Neither construct of MCFD2 changes the ratio between dimer and hexamer, indicating no consequence of the presence of MCFD2 on the oligomeric state of ERGIC-53 4H. **(b)** Native PAGE analyses of dimeric and hexameric ERGIC-53 4H+L in the presence and absence of the reducing agent DTT. Neither species dissociates in the presence of DTT, indicating that the two membrane-proximal cysteine residues are not instrumental for keeping oligomeric ERGIC-53 4H+L intact. **(c)** SDS-PAGE analysis of monomeric, dimeric and hexameric ERGIC-53 in the absence of glutaraldehyde, and after 1 and 10 min incubation with 0.025% glutaraldehyde, respectively. A clear tendency of the monomeric and dimeric species to assemble into higher oligomers is observed.

To further elucidate the role of the two cysteine residues present in the membrane-proximal linker region, native PAGE analysis of dimeric and hexameric ERGIC-53 4H+L after incubation with DTT was performed. The results show no difference in the running behaviour between reduced and non-reduced samples of either dimeric or hexameric species (figure 12b). While this seems to confirm that the cysteines are not instrumental for the stability of the oligomers, it cannot be precluded, in particular for the hexamer, that they are in a hydrophobic core and well protected from reduction by DTT.

In order to further strengthen the observation that ERGIC-53 4H+L is assembly-competent, crosslinking studies using glutaraldehyde were performed with the monomeric, dimeric and hexameric species of ERGIC-53 4H. The results confirm our previous observation, namely that the lower oligomeric species are capable of further assembling into higher oligomeric species, and also suggest a dispensable role in oligomerisation for the two cysteine residues present in the C-terminal linker (figure 12c).

3.4.3 Interaction Studies with MCFD2

Previous studies have determined the binding affinity between MCFD2 and the ERGIC-53 CRD to lie in the low-nanomolar region and indicated that this binding is very susceptible to mutations in MCFD2. While some mutations disrupt the interface of the complex, others render MCFD2 unable to bind Ca^{2+} and therefore inhibit its folding, which consequently inhibits binding to the ERGIC-53 CRD (Nyfeler *et al*., 2008a; Nishio *et al*., 2010; Wigren *et al*., 2010; Elmahmoudi *et al*., 2011). The mutations abolishing the interaction are mostly located in the folded EF-hands of MCFD2, and the role of the unfolded N-terminus in this interaction has not been investigated thoroughly. Earlier data have shown that the binding affinity between native MCFD2 or N-terminally truncated MCFD2-ΔN and the ERGIC-53 CRD lies in the same, low-nanomolar, range (Wigren, 2012), but the oligomerisation domain of ERGIC-53 was absent in these studies. We have therefore undertaken different approaches to investigate whether the unfolded N-terminus of MCFD2 plays a role in the interaction with the full lumenal part of ERGIC-53. Several different potential roles for the unfolded N-terminus come to mind: firstly, it might be involved in the oligomerisation of ERGIC-53 or its stabilisation by providing a link between the monomers, binding to one CRD *via* its EF hands and to the next monomer *via* the unfolded tail, which would presumably fold upon interacting with its partner. Secondly, it might bind to the oligomerisation domain of the same or a neighbouring ERGIC-53 molecule or, thirdly, it might be involved in the recruitment of cargo glycoproteins to ERGIC-53. To test the first hypothesis, we have performed SEC studies with dimeric ERGIC-53 and both native MCFD2 and MCFD2-ΔN. After incubating dimeric ERGIC-53 4H with each of the two MCFD2 constructs for 1h in the presence of 5 mM $CaCl_2$, the complexes were analysed by SEC and their elution profiles compared. In this experiment, while the complex formation was clearly confirmed by a shift of the elution volumes towards a higher molecular mass, no trace of newly formed hexamer was observed

for either complex, ruling out the possibility that the unfolded N-terminus of MCFD2 promotes further oligomerisation of dimeric ERGIC-53 to form hexamers (figure 13).

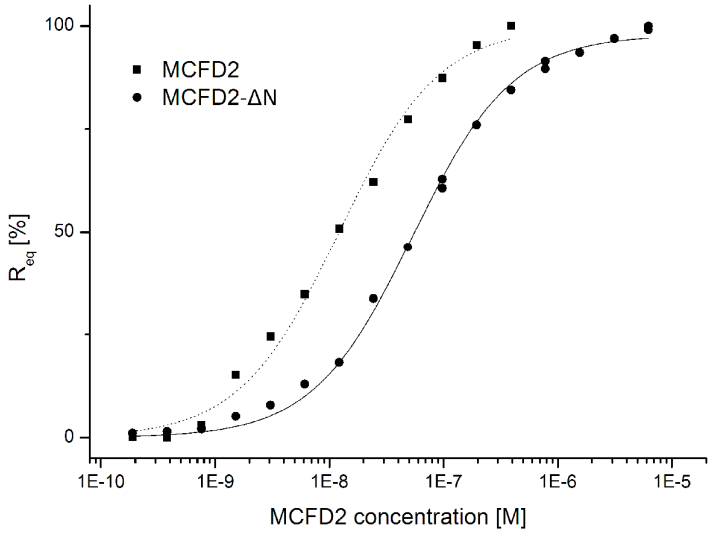

Figure 13: SPR analyses of the interaction between hexameric ERGIC-53 4H and MCFD2 and MCFD2-ΔN, respectively. The steady-state response for each injection in relation to the maximum observed response at steady-state for the highest concentration is plotted against the MCFD2 concentration. The dissociation constant K_D, as determined from each constructs concentration at R_{eq} = 50%, is $5.6*10^{-8}$ M and $1.2*10^{-8}$ M, for MCFD2-ΔN and MCFD2, respectively.

The second hypothesis, that the unfolded tail of MCFD2 is binding to the ERGIC-53 oligomerisation domain, was investigated using SPR. Hexameric ERGIC-53 was immobilised on a CM5 chip by amine coupling, and binding assays were performed using both native and N-terminally truncated MCFD2 as analytes. The resulting data demonstrated a 5-fold increase in the binding affinity of native MCFD2 to hexameric ERGIC-53 ($K_D = 1.2*10^{-8}$ M), compared to MCFD2-ΔN ($K_D = 5.6*10^{-8}$ M), indicating that the N-terminus is indeed also binding to ERGIC-53.

To investigate the third hypothesis, which states that MCFD2 is involved in the recruitment of cargo molecules to ERGIC-53, we have attempted to recombinantly express several constructs of the thus-far identified proteins whose transport from the ER to the ERGIC relies on a concerted function of both ERGIC-53 and MCFD2, namely the blood coagulation factors FV and FVIII, as well as Mac2-bp, using the BEVS system. Unfortunately though, all our efforts to produce these proteins were unsuccessful.

3.4.4 Crystallisation of ERGIC-53

In order to further investigate the structure of the lumenal part of ERGIC-53, crystallisation of different constructs and oligomeric species, as well as their complexes with native and N-terminally truncated MCFD2, was attempted. A total of well over 10 000 small-scale sitting-drop experiments were set up to identify crystallisation conditions for a number of combinations between different ERGIC-53 constructs and oligomeric states in complex with both MCFD2 and MCFD2-ΔN. The extent and success-rate of these experiments are summarised in table 4. Ultimately, two conditions that produced crystals could be identified from the screens performed with the complexes between hexameric ERGIC-53 4H and native MCFD2, and hexameric ERGIC-53 4H and MCFD2-ΔN, respectively. The crystals grown in both conditions initially diffracted only poorly, but their size and diffraction quality could be enhanced by further optimising the initially identified crystallisation conditions.

In total, approximately 250 crystals were cryo-protected, flash frozen and tested for diffraction at various synchrotron sources.

Construct(s)	Tested conditions	Crystals obtained
ERGIC-53 4H+L hexamer	1440	no
ERGIC-53 4H hexamer:native MCFD2	1440	yes (1 condition)
ERGIC-53 4H hexamer:MCFD2-ΔN	576	yes (1 condition)
ERGIC.53 4H dimer:MCFD2-ΔN	1728	no
ERGIC-53 4H dimer	6048	no
ERGIC-53 3H hexamer	576	no
ERGIC-53 3H dimer	576	no

Table 4: The number of tested crystallisation conditions for each construct.

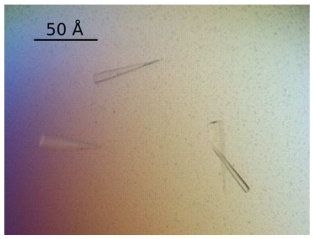

Figure 14: Crystals of ERGIC-53 4H in complex with MCFD2.

The first, cone-shaped crystals, obtained from the ERGIC-53 4H hexamer in complex with native MCFD2, appeared in a 2:1 mixture of 5.7 mg/ml protein and 0.1 M Tris, pH 7.5 – 8.25; 11 – 13% (w/v) PEG4000, 0.15 M $(NH_4)_2SO_4$ after a few days incubation at 4 ºC (figure 14) and - after extensive optimisation - grew to approximately 200 µm in length and 50 µm in width. X-ray diffraction data to a maximum resolution of about 4 Å was recorded at the microfocus beamline P14 of the Deutsches Elektronen-Synchrotron (DESY) in Hamburg, Germany. The protein crystallised in the space group C2 with the unit cell parameters a=275.6 Å, b=49.3 Å, c=305.6 Å, α=γ=90º, β=103.4º. Although the structure of the ERGIC-53 CRD (Nishio *et al.*, 2010; Wigren *et al.*, 2010) was available, extensive molecular replacement searches, making use of differently trimmed models of the CRD in concert with different molecular replacement programs, were not successful in solving the structure. Close inspection of the individual diffraction images revealed several pathologies in the crystal

packing. Smeared spots and possibly a double lattice were observed, in addition to significant radiation damage after only a short exposure of the crystal in the X-ray beam. To circumvent this problem, data was collected by scanning the crystals helically along their long axis, but during data integration and indexing it was revealed that the unit cell was not isomorphous along the length of the crystals. The combination of these problems is probably what prevented a successful determination of the structure.

A different approach was undertaken to solve the structure by experimental phasing and for this, crystals were soaked with Ta_6Br_{14}, a compound previously identified to be suitable for isomorphous replacement, especially in the case of large molecules diffracting to low resolution (Schneider & Lindqvist, 1994; Banumathi *et al*., 2003). These crystals were however not isomorphous, and diffracted to only ~7 Å resolution, with a very weak anomalous signal to ~8.5 Å resolution. The data collection and processing statistics for both native and Ta_6Br_{14}-derivatised crystals are summarised in table 5.

After this unsuccessful attempt to determine the structure of hexameric ERGIC-53 4H in complex with native MCFD2, a new effort was undertaken and screening for suitable crystallisation conditions was resumed. Reasoning that the unfolded N-terminus of MCFD2 might be unfavourable for homogenous crystal packing, this screening was undertaken using the complex of ERGIC-53 4H with the MCFD2-ΔN variant at a concentration of 7.5 mg/ml, and crystals of a bi-pyramidal shape were obtained from the same condition as previously. After observing weak diffraction of these crystals at synchrotron sources, followed by optimisation of the crystallisation condition to increase their size and quality, larger crystals diffracting to 4 Å resolution could be obtained. It became clear that the protein had once more crystallised in the space group C2, but with very different unit cell parameters, namely a=379.9 Å, b=218.6 Å, c=219.5 Å, α=γ=90º, β=125.3º. Once more, extensive trials to solve the structure by molecular replacement using the available model of the CRD as a whole or pruned to its central β-sheet were performed, but were unsuccessful. Later it was discovered that these crystals were in fact of a contaminant, namely ferritin from the expression host *T. ni*, and this observation is discussed in more detail in paper IV. Since this revelation it has of course to be considered that the crystals produced from the preparations of ERGIC-53 4H in complex with native MCFD2 are also of ferritin. Attempts to solve the structure using the available structure of ferritin (Hamburger *et al*., 2005) as a search model have however failed. This might not be surprising considering the very different crystal morphology, as one could expect ferritin crystals grown in the same condition to be identical. However, due to the low quality of these crystals, it has to be considered that the data might have been incorrectly processed and it can therefore not be concluded at this time whether the crystalline protein is in fact ferritin or ERGIC-53 in complex with MCFD2-ΔN.

	Native	Ta_6Br_{14} derivative
Data collection		
Beamline	PETRA III P14	PETRA III P14
Wavelength [Å]	0.97628	1.25492
Space group	C2	C2
Cell axes a, b, c [Å]	275.66, 49.33, 305.42	269.08, 49.76, 302.08
Cell angles α, β, γ [°]	90.00, 103.38, 90.00	90.00, 100.59, 90.00
Resolution [Å]	49.55 – 4.00 (4.19 – 4.00)	49.49 – 7.20 (8.04 – 7.20)
R_{merge}	0.24 (1.07)	0.26 (1.02)
Mean I/σI	6.0 (1.2)	6.6 (1.8)
Mean $CC_{1/2}$	0.983 (0.342)	0.99 (0.53)
Completeness [%]	98.8 (95.8)	94.7 (86.8)
Multiplicity	3.3 (3.2)	4.3 (4.3)
Number of reflections	112834 (13963)	24946 (6284)
Number of unique reflections	34624 (4379)	5778 (1461)
Wilson B-factor [Å²]	102.9	240.6

Table 5: Data collection statistics of native crystals of ERGIC-53 4H in complex with MCFD2 and the Ta_6Br_{14}-derived crystals. Statistics for the highest-resolution shell are shown in parentheses.

3.4.5 Small-angle X-ray Scattering

Concurrently with the crystallographic studies, small-angle X-ray scattering experiments were performed to investigate the structure of ERGIC-53 in solution, with the intention to complement a possible X-ray crystal structure of monomeric or dimeric ERGIC-53. A low-resolution solution structure of hexameric ERGIC-53 might have allowed us to build a higher resolution model by using the structure of the monomeric or dimeric crystal structure of ERGIC-53 as rigid bodies and, by enforcing symmetrical restraints, fit them to the solution scattering curve of the hexamer. Since no structure of either monomeric or dimeric ERGIC-53 could be obtained, this approach could unfortunately not be undertaken. Nevertheless, the SAXS scattering data still allows for some speculation about the structure of hexameric ERGIC-53.

First, SAXS scattering data of dimeric ERGIC-53 4H+L in solution was recorded at beamline X33 of the Deutsches Elektronen-Synchrotron (DESY) in Hamburg, Germany (Blanchet *et al*., 2012). The molecular mass of the dimeric protein was estimated to ~100 kDa using Guinier analysis implemented by AUTORG (Petoukhov *et al*., 2007) and comparison of the scattering intensity of the extrapolated intensity to zero angle I(0) with a BSA standard, a value corresponding well with the theoretical molecular mass of dimeric ERGIC-53 4H+L. The maximum intramolecular distance D_{max} of the particle was determined to be ~21 nm using GNOM (Svergun, 1992) and the paired distance distribution function clearly indicates the presence of two domains (figure 15a,b).

In addition, solution scattering data of the ERGIC-53 4H+L dimer in complex with native MCFD2 was recorded. An increased molecular mass of ~120 kDa was determined, while the maximum distance D_{max} remained approximately the same. The distribution of intramolecular

paired distances however differs strongly from that obtained for the dimeric protein alone, as it does not show the clear presence of two domains, but rather a shape resembling a rod-like structure (figure 15a,c). This observation could be explained by MCFD2 being placed in between the two domains, thereby obscuring their separate peaks.

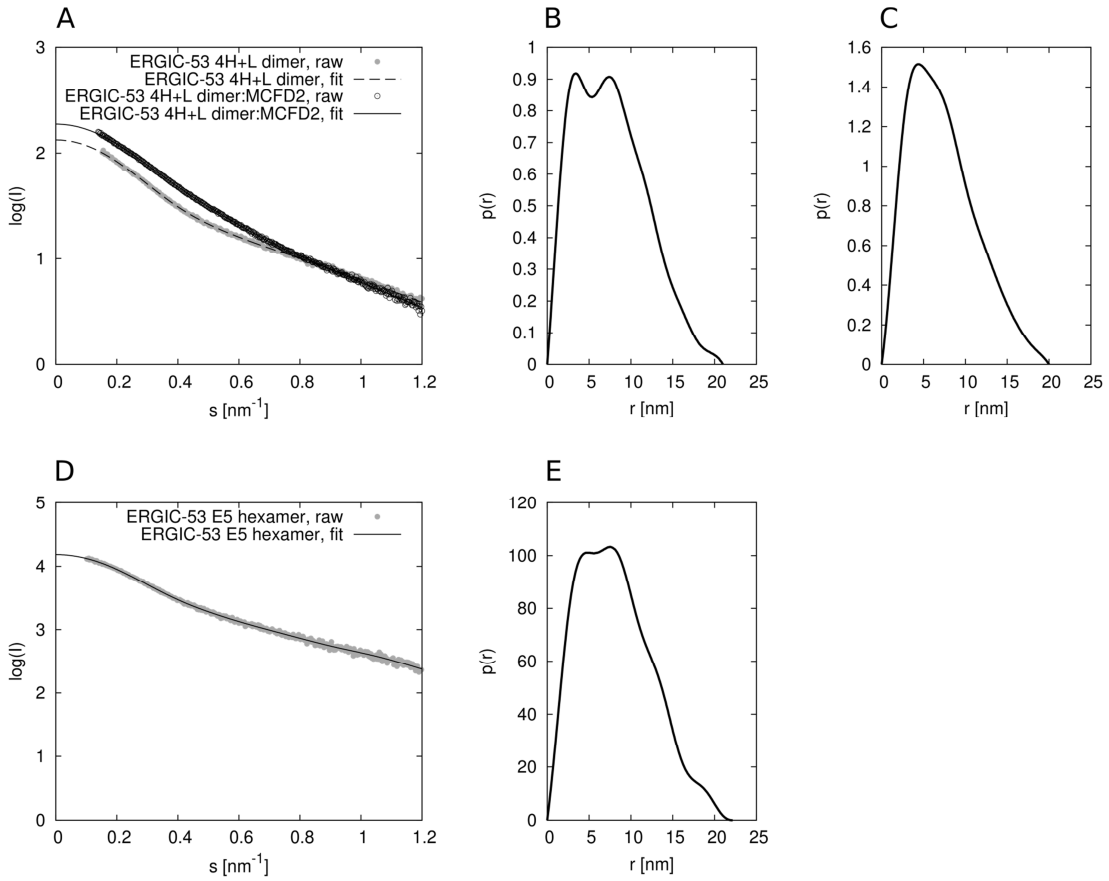

Figure 15: Solution small angle X-ray scattering of ERGIC-53. **(a)** Solution scattering data of dimeric ERGIC-53 4H+L alone and in complex with MCFD2 collected at beamline X33 of DESY. The logarithm of the scattering intensity is plotted as a function of momentum transfer $s=4\pi \sin(\Theta)/\lambda$, where Θ is the scattering angle and λ is the X-ray wavelength. **(b and c)** Paired distance-distribution functions of dimeric ERGIC-53 4H+L alone and in complex with MCFD2, respectively. **(d)** Solution scattering data of hexameric ERGIC-53 4H collected at beamline P12 of DESY. **(e)** Paired distance distribution function of hexameric ERGIC-53 4H. Note that a comparison of the scattering intensity is invalid as the scattering data was recorded at two different beamlines.

By comparing the two shapes obtained from *ab initio* modelling, we aimed to determine the location of the ERGIC-53 CRD within the low-resolution shapes. The difference density obtained from the subtraction of the ERGIC-53 model from the complex model could allow for the placement of MCFD2, and therefore very likely also identify the approximate position of the ERGIC-53 CRD, since the complex structure is well described (Nishio *et al*., 2010; Wigren *et al*., 2010). To achieve this, low-resolution shapes were computed *ab initio* based

on the scattering data of both the dimeric ERGIC-53 4H+L as well as dimeric ERGIC-53 4H+L in complex with MCFD2. This analysis was performed using DAMMIF (Franke & Svergun, 2009) and a P2-symmetry was imposed, resulting in fit-values as reported by DAMMIF of 0.0195 and 0.0158, respectively. One example of a shape of both dimeric ERGIC-53 4H+L and its complex with MCFD2 is shown in figure 16. Both models comprise an elongated shape with similar dimensions and indeed, additional electron density is observed in the centre of each subunit.

As the CRD in complex with MCFD2-ΔN has been shown to be monomeric in the absence of the oligomerisation domain, it is assumed to be represented by the extremities of the low-resolution models and was manually placed there using UCSF Chimera (Pettersen *et al*., 2004). While the location of the CRD can be identified in these two shapes, the variability of other obtained shapes prevents us from drawing sound conclusions from this approach.

SAXS data of hexameric ERGIC-53 E5 in solution was then recorded at beamline P12 of DESY (Blanchet *et al.*, 2015). The molecular weight of the particle is estimated to ~320 kDa, again confirming the hexameric nature of ERGIC-53. Furthermore, the paired distance distribution function indicates the presence of a multi-domain protein with a D_{max} of ~22 nm (figure 15d,e)

Ab initio low-resolution shape determination was performed based on these recordings as well, and a P6 symmetry was enforced. The computation resulted in two structurally distinct shapes, both of which fit the scattering data similarly well with fit-values as reported by DAMMIF of 0.0422 and 0.0446, respectively. While shape 1 resembles a bouquet of flowers with elongated stems and a crown, shape 2 can be described as a planar disc with spherical elements protruding from its side (figure 16). The volume of the protrusions of both models allows for a good fit of one CRD molecule.

While these models provide some insight into the possible structure of hexameric ERGIC-53, they would have to be rigorously validated using other available structural information and, while the high-resolution crystal structure of the CRD is available, any sound interpretation of the *ab initio* models is prohibited by the lack of structural data of the oligomerisation domain. This would not necessarily be a limiting factor for the interpretation of a monomeric model, but since no structural information on the hexameric protein is available, further modelling of the oligomerising region of ERGIC-53 becomes cumbersome and unreliable.

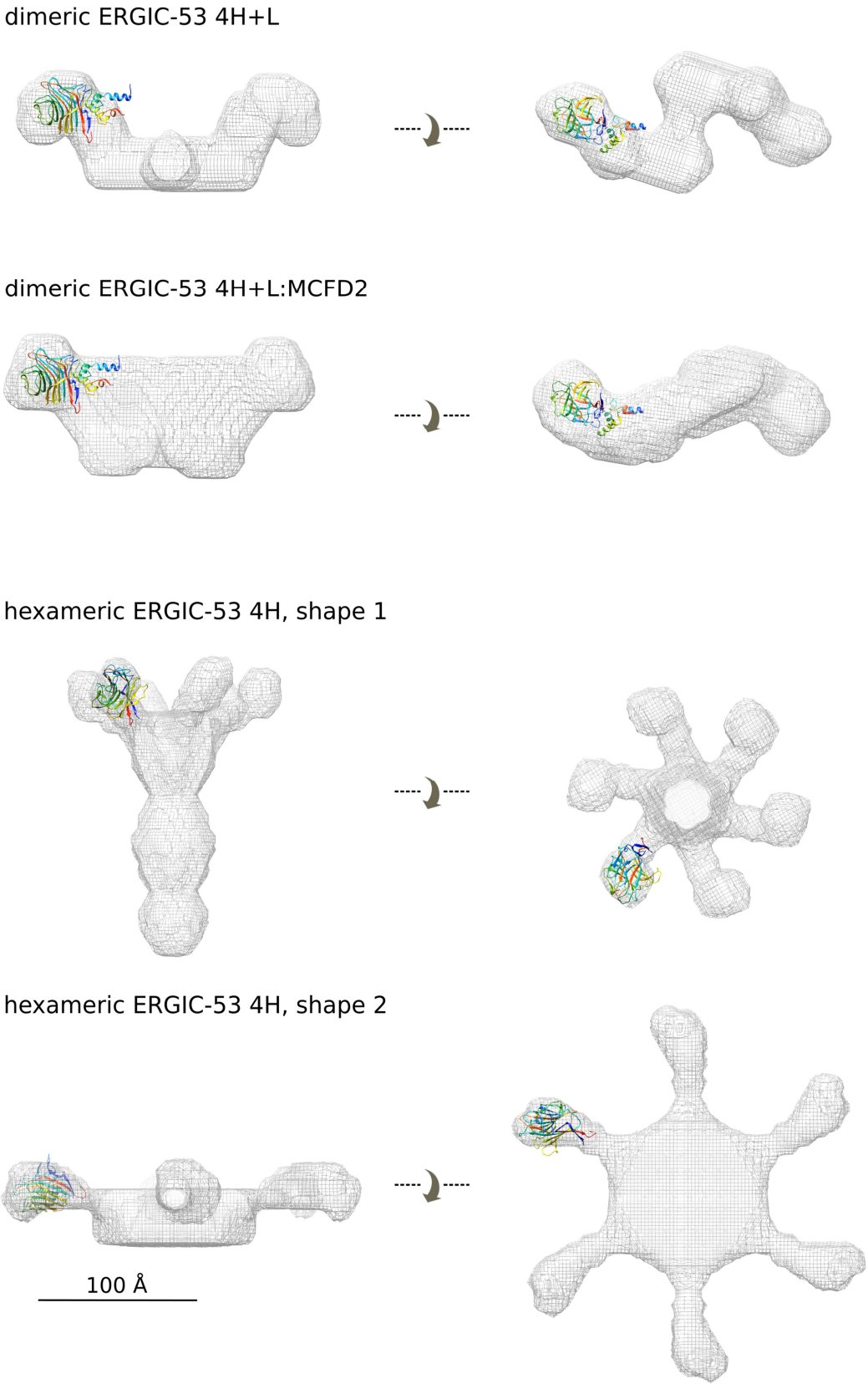

Figure 16: Low-resolution shapes obtained *ab initio* based on the scattering data of dimeric ERGIC-53 4H+L in complex with MCFD2 and hexameric ERGIC-53 4H. The illustrations on the right are rotated 90º around the horizontal axis.

3.4.6 Cryo-EM

To further supplement our X-ray based structural studies of ERGIC-53, attempts to obtain a low-resolution structure of the hexameric ERGIC-53 by cryo-electron microscopy were also initiated. Initial experiments proved the protein to be suitable for cryo-EM studies and images were recorded at a magnification of 60 000x, clearly showing single particles embedded in ice (figure 17). To our surprise however, these results were not reproducible with the subsequent protein preparations and their vitrification, and cryo-EM studies were therefore not continued.

Figure 17: Hexameric ERGIC-53 4H+L embedded in ice, recorded at a magnification of 60 000x using cryo-EM. The lower panel shows a schematic interpretation of the different particles.

3.4.7 Modelling of the ERGIC-53 Oligomerisation Domain

In order to facilitate further analysis of the previously described *ab initio* model of hexameric ERGIC-53 derived from the SAXS studies, we aimed to produce a model of the oligomerisation domain using the protein structure prediction services QUARK (Xu & Zhang, 2012) and Robetta (Raman *et al.*, 2009). Since no structural template of any homologous molecule is available, the structures are based on the amino acid sequence only. Five of the most structurally diverse examples from the ten models obtained from each prediction server are visualised in figure 18.

The QUARK-models resemble each other in their architecture as they all comprise an elongated, ~90 Å long, coiled-coil region encompassing three helices and the most N-terminal helix most often aligned diagonally to the coiled-coil. In contrast to this, the models obtained from Robetta exhibit larger differences in their structures. While the most common model displays two stacked coiled-coil regions encompassing two helices, some models display a four-helix bundle, and one model is comprised of a five-helix bundle. All models are completely α-helical, which is in agreement with data we have obtained from circular dichroism experiments (Wigren, 2012).

A rough estimate of whether these models could resemble the actual structure of the ERGIC-53 oligomerisation domain can be performed based on the low-resolution *ab initio* shapes obtained from SAXS experiments. Shape 1 displays an oligomerisation region extending over approximately 160 Å, which is longer than any model predicted by either the QUARK or Robetta algorithm. In shape 2, the distance ranging from termini of the CRD to the centre of the inner disc is approximately 80 Å and would therefore correspond better to the length of the models obtained from the QUARK algorithm. This approach has however to

be considered very hypothetical, as it is based on two, experimentally unverified, modelled systems

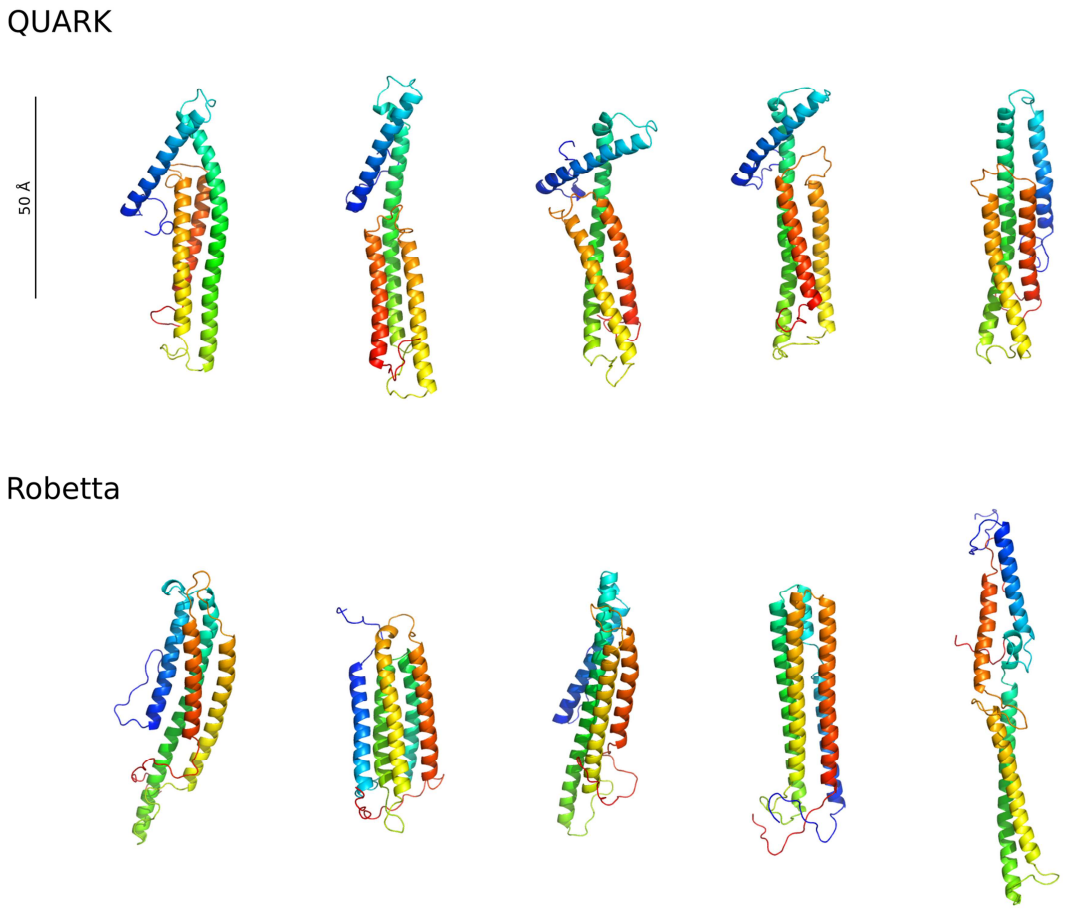

Figure 18: A selection of models of the ERGIC-53 oligomerisation domain produced by *de novo* structure prediction using the QUARK and Robetta servers. The models are represented as rainbow-coloured cartoons with the N-terminus in blue and the C-terminus in red.

One can attempt to use a very different perspective to discuss the role of hexameric ERGIC-53 in the cell by applying our current biological and structural knowledge of the assembly of the COPII-coat proteins, and of the molecules transported *via* ERGIC-53 within secretory vesicles.

Recent cryo-EM studies of the COPII-cage revealed the outer layer of the cage, consisting of the Sec13-Sec31 complex, to be a cuboctahedron with a diameter of 600 Å, which is comparable to the diameter of COPII-coated vesicles observed *in vivo* (Stagg *et al*., 2006). The arrangement of the Sec13-Sec31 complex seems to be modular and expandable to accommodate different cargo sizes. The inner layer, which is recruited to the ER membrane by the GTPase Sar1 and interacts directly with the outer layer (Lee *et al*., 2004), consists of

the Sec23-Sec24 complex, and Sec24 has been shown to be responsible for binding to cargo proteins *via* anterograde ER transport motifs (Nishimura & Balch, 1997; Sato & Nakano, 2002; Wendeler *et al*., 2007). Recently, it has been suggested that the cargo receptor proteins might have a role in directing the structure of the COPII-cage (Gürkan *et al*., 2006).

On the lumenal face of the membrane, ERGIC-53 assembles into hexamers. As we have demonstrated, the hexameric protein itself has a maximum diameter of ~220 Å, about a third of the diameter of the COPII-cage. Only a partial crystal structure of the coagulation factor FVIII, which is a 267 kDa protein recognised by ERGIC-53, is available (Shen *et al*., 2008). However, even this available part, which is lacking the major B-region, already has a maximum dimension of ~250 Å. It is therefore puzzling to imagine how ERGIC-53 and its bound cargo proteins could fit into COPII-coated vesicles of ~500 Å diameter, especially when taking into consideration the fact that hexameric ERGIC-53 can potentially bind up to six cargo-molecules simultaneously. ERGIC-53 might therefore indeed play a role in the expansion of the COPII-cage to allow for the efficient packaging of larger cargo molecules, and it can be speculated that the hexameric organisation of the protein translates across the membrane and enforces a rearrangement of the COPII-proteins to enforce such an expansion (figure 19)

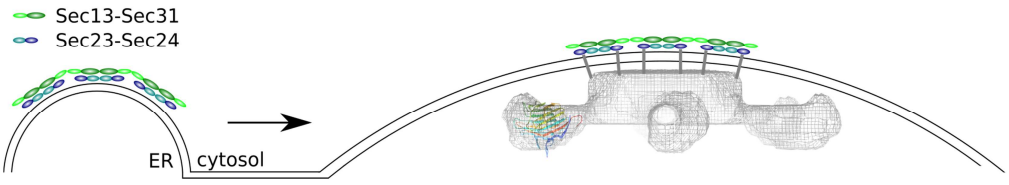

Figure 19: A proposed model for the expansion of COPII-coated vesicles by the incorporation of ERGIC-53. Binding of hexameric ERGIC-53 might lead to conformational changes of the Sec23-Sec24 complex *via* a translation of the hexameric arrangement of ERGIC-53 to the cytosolic face of the vesicles.

A major unresolved question concerning the oligomeric ERGIC-53 is why the protein cannot effectively transport even smaller cargo proteins in its dimeric state. It does not seem plausible that the cell would have an as-yet unknown mechanism to actively hinder dimeric ERGIC-53 from being packed into secretory vesicles. There is however evidence emphasising a key role for glycoprotein receptors and other proteins interacting with Sec24 in the assembly of the coat protein complex (Aridor *et al.*, 1999). Accordingly, failure of ERGIC-53 to hexamerise might lead to the aberrant exposure of the cytoplasmic transport motif with regards to the location of Sec24 and thereby to an interaction that is too weak for both the efficient assembly of the coat-protein complexes and for ERGIC-53 to be efficiently packed into vesicles. According to this speculation, ERGIC-53 would actively partake in the assembly of the COPII-cage. The binding affinity of Sec24 to the cytosolic di-acidic transport motif is, to our best knowledge, yet to be determined.

If it is indeed the case that hexameric ERGIC-53 has an advantage over smaller oligomers of the protein when it comes to incorporation into COPII-coated vesicles, then this raises the possibility that a factor on the lumenal face of the ER membrane extorted some selective pressure onto ERGIC-53 to form hexamers. Obvious candidates to enact such a selective pressure are the cargo proteins transported by ERGIC-53.

One interesting case to argue for a selective pressure towards hexameric ERGIC-53 are the studies performed in the laboratory of R. Sitia (Anelli *et al*., 2007, 2015; Cortini & Sitia, 2010), which showed that the controlled polymerisation and secretion of hexameric IgM relies directly on its interaction with hexameric ERGIC-53. Their findings show that a conserved N-linked glycan on the C-terminal tailpiece of secretory chains is recognised by ERGIC-53, and that the removal of this glycan disturbs the interaction with ERGIC-53 and leads to larger polymeric IgM-assemblies. These results suggest a possible function of ERGIC-53 as a platform governing the assembly of hexameric IgM.

With regards to the other cargo molecules of ERGIC-53 that have been thus-far identified, its hexameric organisation does not seem to play a role in cargo recognition, as neither α1-antitrypsin, Cathepsin C and Z, nor Mac2-bp require their presence as oligomers to enact their function.

Additionally, it is intriguing that ERGIC-53 does not seem to be essential for the secretion or assembly of the blood coagulation factors. While mutations in ERGIC-53 or MCFD2 lead a combined deficiency of the blood coagulation factors V and VIII (F5F8D), patients carrying this disease still have blood levels of these proteins at 5-30 % of the normal levels (Nichols *et al*., 1998), suggesting that exit of native FV and FVIII from the ER can also occur *via* the bulk-flow mechanism (Wieland *et al*., 1987; Thor *et al*., 2009).

One could therefore speculate that the primary function of hexameric ERGIC-53 is to assist the correct assembly of IgM, whose correct and efficient synthesis is of tremendous importance to protect the human body from invasive pathogens, and that other proteins may merely have developed the ability to usurp the function of ERGIC-53, in order to themselves be transported more efficiently from the ER along the secretory pathway.

4 CONCLUSIONS

Within this thesis project, a number of proteins involved in glycoprotein transport and processing in the early secretory pathway have been investigated. X-ray crystallographic studies allowed us to determine the structures of both Erv41p and Ktrp4, and small-angle X-ray scattering experiments performed with ERGIC-53 gave new insights into the shape of this protein in its functional, hexameric form. Additional biochemical and biophysical methods have been applied to further characterise each protein.

X-ray crystallographic studies of Erv41p led to the first structure of any member of the class of ER vesicle proteins, and analyses of the surface properties of the protein's structure together with sequence comparisons with its homologues led us to propose a possible site of interaction with other proteins. Almost the entirety of one face of the protein is negatively charged, and this large surface clearly provides grounds to hypothesise about its potential as a protein-protein interaction interface. We also attempted to determine the crystal structure of the binding-partner Erv46p, but have unfortunately been unsuccessful. We have however been able to show a direct interaction between Erv41p and Erv46p by SPR experiments. As the two proteins act in concert when transporting their cargo, further insight into the structure of Erv46p and the Erv41p/Erv46p complex is of great interest to completely understand the mechanism of protein binding and release by these proteins in the different compartments of the early secretory pathway.

During the course of these studies, Ktr4p was identified as the first protein known to be transported by the ER vesicle protein complex, and we have therefore also investigated its structure and function. The structure of the apo-protein, as determined by X-ray crystallography, shows that it belongs to the GT-A family of glycosyltransferases, and the structure of Ktr4p in complex with GDP and Mn^{2+}, combined with its similarity to the previously characterised homologue Kre2/Mnt1, led to the hypothesis that it acts as a mannosyltransferase. Our biochemical analysis of the catalytic activity confirmed that this is indeed the case. The relatively poor activity of the enzyme towards the provided substrates however led us to conclude that the experiments were performed with incomplete and/or inadequate donor-substrates. Attempts to identify an interaction between Ktr4p and an array of complex, branched mammalian glycans were not successful, and the precise substrates of the enzyme in the Golgi apparatus therefore remain to be elucidated.

The protocols we have developed for the recombinant production of Erv41p and Erv46p in insect cells, as well as of Ktr4p in *E. coli*, will allow for further, detailed interaction studies between the Erv41p/Erv46p complex or its individual components with Ktr4p. Such studies could provide more insight into the mechanism of protein capture within the ER and release within the Golgi apparatus, and thereby increase our understanding of the mechanisms involved in protein transport between the organelles of the early secretory pathway.

Our attempts to determine the structure of Erv46p and the lumenal part of ERGIC-53 led to the identification of a common crystallisation contaminant, present after the purification of the two recombinantly expressed proteins that were secreted by the insect cells into the medium. The expression host *T. ni*, and specifically its ovarian cell line High Five, is becoming more popular as it allows for the production of many eukaryotic proteins whose expression fails in the more traditional expression host *E. coli*. We show that ferritin is a contaminant from the High Five strain that is prone to crystallise even in minute amounts, and warn about its deceiving presence.

Our biophysical and biochemical studies performed with different oligomeric states of ERGIC-53 shed light on its mechanism of oligomerisation. We show that the oligomerisation is independent of disulfide-bond formation, and our results are in agreement with Neve *et al.*, but contradict those of Appenzeller *et al.* and Lahtinen *et al.* While these three studies have been performed *in vivo*, our *in vitro* experiments allow us to exclude significant factors that may influence the assembly of ERGIC-53. In addition, we show that MCFD2 does not promote the oligomerisation of ERGIC-53, but that its unfolded N-terminal tail does interact with ERGIC-53 when the oligomerisation domain is present. By studying dimeric and hexameric ERGIC-53 with small-angle X-ray scattering we were able to propose a low-resolution shape for the dimeric protein, and two potential shapes for the hexameric protein. The exact structure of the oligomerisation domain is still unknown, and only structural information from this domain in its oligomeric state would finally put an end to the discussion about the oligomerisation of ERGIC-53.

Overall, the structural, biophysical and biochemical analyses of this work have led to new insights into glycoprotein transport and processing in the early secretory pathway, and will enable future studies that will hopefully yield a full understanding of each of the individual proteins described here, their roles in the cell and their links to disease.

5 ACKNOWLEDGEMENTS

Very many creative and critical minds have been of immeasurable assistance to me during the last few years and I remain deeply grateful for the advice and support I have received.

First, I would like to thank my supervisor Jodie Guy. You have been my scientific counsellor, reviewer, editor and stubborn reminder of rapidly approaching deadlines and your scientific expertise has helped me overcome many obstacles. I have learned a lot from you, especially in the areas of experimental design and crystallography. Thank you also for giving me the freedom to plan and develop the projects independently, but always being there when I needed advice.

I would also like to thank Professor Ylva Lindqvist. Thank you for first accepting me as a PhD student and for your further support as my co-supervisor. Your enthusiasm for science in general and this project in particular has been truly inspiring and very uplifting in many situations. I would also like to thank Professor Gunter Schneider for sharing his expertise and regularly giving me helpful advice along the way. I want to thank the three of you for the many challenging broad and detailed scientific discussions we have had. Together, you have provided an environment of creativity, curiosity and familiarity.

Edvard, your kindness, patience and persistence from the very first day of my arrival has been of great help and our sometimes derailing speculations about this bunch of ER proteins have been very fun and insightful. Katja, thank you for your guidance and for sharing so many lunches and weekends in the lab. It has been great working with you. Francesca, grazie mille for all the help with insect cells and for so many shared synchrotron trips. Your company has made the travels much more joyous. I would also like to thank Jenna, during the summer you spent in the lab you have helped me a lot, and our co-immunoprecipitation adventures with laboratory-grade yeast from hemköp are something to remember.

I would like to thank everyone who has collaborated with me on the projects of this thesis: Magnus, thank you for lending me your knowledge of the sugary side of protein crystallography and for being a great colleague and science-nerd. Anne, your support with my first steps into the perplexing realm of SAXS was very important for this thesis and I thank you for all your help with teaching me how to make sense of the scattering data. Pasi, thank you for giving me a glimpse into the world of electron microscopy and all your efforts invested into sample preparation and image recording. I would also like to thank everyone at the PSF for their most friendly and competent help, which was crucial for the outcome of this thesis. Martin, thank you for the maintenance of all the robotic crystallography equipment and the sanity of its users, Helena, for advice with construct design and Ida for cloning Ktr4p and surprising me with soluble protein only a couple of days later. Thank you Essam for your support with SPR studies and your hospitality as my former neighbour.

I would also like to thank all the previous and present members of the division for the fantastic times spent together at work and after: Ömer, thank you for your continued honesty and sarcasm and for all the heated conversations. Bernie, for first guiding me towards hjulet

and all the analyses of the goings-on in science and the world during the lunch-breaks. Thank you especially for your help with developing data-collection strategies, data processing and the resulting necessity of troubleshooting. Katharina, your contagious good mood and colourful clothing often brought some balance into the office. We should resume the habit of sharing long-way-home music. Robert, thank you for all the discussions about the big and small questions in science, the lessons on Norse mythology, etymology and nearly forgotten cinematic masterpieces. Rajesh, for always being helpful and the many lovely barbecues you organised to bring everyone together after work. Jason, for late-night discussions about economical superpowers and for being so brave to try to convince me that the Beatles were, in fact, not all that important in musical history. Ming-Wei, it has been a pleasure having you around. I'm still trying to match your photography skills. Eva-Maria, for your enthusiasm for science and organising MBB-pubs. Agata, thank you for the good times here and abroad. A polish wedding can never be forgotten! Doreen, for the help with teaching and gatherings in the old and the new hood. Atsushi, for help with crystallography and insights into the aesthetics of a healthy lunch. Dominic, for deep insights into history and music. I also want to thank all the new- and not-so-new-comers in the lab: Brinda, for keeping me awake during long synchrotron sessions with Bollywood music. Richard, for constantly trying to hack the world and Shoude for your constant smile. Peter, for sharing all your knowledge in such a helpful way. Selma, for wreaking havoc in my close proximity. The whole Petzold-gang, for opening up a window into the bewildering science of RNA and NMR. Lorenzo and Hampus, it has been most fun sharing the office with you and exchanging views on structural biology where we could find a common ground. Grazie och Tack!

I would also like to thank Victoria, Eva and Alessandra for taking care of the to me often diffuse administrative side side of research. Thank you for all the help with organising the many travels abroad, confusing invoices and reseräkningar. Ahmad, thank you for keeping the lab in such a tidy state. You made a big difference.

Many thanks also to everyone outside the division who has made the last years so enjoyable. Juha, for the many climbing and slack-lining sessions and impromptu midnight street-corner discussions about all things relevant. Will and Qing, it was most fun working with you on the thioreductase. If you ever do work on a less troublesome enzyme, let me know! Rozbeh, Michael, Jens, Joe, Lionel, Sergey, Ási, Olle, Fatma, TC, Rosaria, Arwen, Wouter, Ela, Devaraj, Jaakko, Shiromi, Emma, Eleonora, Jen, Barbara, Massa, Stef, Isabel and Elisa, thank you for the fun times all around Stockholm.

Christina, Lasse, Henrik, Lovisa och Märta, tack för de fina dagarna på spåren och på lederna i fjällen och de många, trevliga middagarna hemma hos er.

Ein herzliches Dankeschön an meine Familie. Eure Unterstützung in allen meinen Entscheidungen und eure moralische Hilfe bei den viel zu seltenen Besuchen daheim oder in Stockholm während der letzten Jahre waren unglaublich wichtig. Danke viel, viel mol!

Maria, du betyder så mycket för mig. Tusen tack för allt!

6 REFERENCES

Akopian, D., Shen, K., Zhang, X., & Shan, S. (2013). *Annu. Rev. Biochem.* **82**, 693–721.

Anelli, T., Ceppi, S., Bergamelli, L., Cortini, M., Masciarelli, S., Valetti, C., & Sitia, R. (2007). *EMBO J.* **26**, 4177–4188.

Anelli, T., Sannino, S., & Sitia, R. (2015). *Free Radic. Biol. Med.* **83**, 323–330.

Appenzeller, C., Andersson, H., Kappeler, F., & Hauri, H. P. (1999). *Nat. Cell Biol.* **1**, 330–334.

Appenzeller-Herzog, C. & Hauri, H.-P. (2006). *J. Cell Sci.* **119**, 2173–2183.

Appenzeller-Herzog, C., Roche, A. C., Nufer, O., & Hauri, H. P. (2004). *J. Biol. Chem.* **279**, 12943–12950.

Arar, C., Carpentier, V., Le Caer, J. P., Monsigny, M., Legrand, A., & Roche, A. C. (1995). *J. Biol. Chem.* **270**, 3551–3553.

Aridor, M., Bannykh, S. I., Rowe, T., & Balch, W. E. (1999). *J. Biol. Chem.* **274**, 4389–4399.

Bannykh, S. I., Rowe, T., & Balch, W. E. (1996). *J. Cell Biol.* **135**, 19–35.

Banumathi, S., Dauter, M., & Dauter, Z. (2003). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **59**, 492–498.

Barlowe, C. (2003). *Trends Cell Biol.* **13**, 295–300.

Belden, W. J. & Barlowe, C. (2001). *Science.* **294**, 1528–1531.

Blanchet, C. E., Spilotros, A., Schwemmer, F., Graewert, M. A., Kikhney, A., Jeffries, C. M., Franke, D., Mark, D., Zengerle, R., Cipriani, F., *et al*. (2015). *J. Appl. Crystallogr.* **48**, 431–443.

Blanchet, C. E., Zozulya, A. V., Kikhney, A. G., Franke, D., Konarev, P. V., Shang, W., Klaering, R., Robrahn, B., Hermes, C., Cipriani, F., *et al*. (2012). *J. Appl. Crystallogr.* **45**, 489–495.

Blobel, G. & Dobberstein, B. (1975). *J. Cell Biol.* **67**, 835–851.

Breuza, L., Halbeisen, R., Jenö, P., Otte, S., Barlowe, C., Hong, W., & Hauri, H. P. (2004). *J. Biol. Chem.* **279**, 47242–47253.

Budnik, A. & Stephens, D. J. (2009). *FEBS Lett.* **583**, 3796–3803.

Bue, C. A. & Barlowe, C. (2009). *J. Biol. Chem.* **284**, 24049–24060.

Bue, C. a, Bentivoglio, C. M., & Barlowe, C. (2006). *Mol. Biol. Cell.* **17**, 4780–4789.

Bygrave, F. L. & Benedetti, A. (1996). *Cell Calcium.* **19**, 547–551.

Caldwell, S. R., Hill, K. J., & Cooper, A. A. (2001). *J. Biol. Chem.* **276**, 23296–23303.

Carrière, V., Piller, V., Legrand, A., Monsigny, M., & Roche, A. C. (1999). *Glycobiology*. **9**, 995–1002.

Chavrier, P., Parton, R. G., Hauri, H. P., Simons, K., & Zerial, M. (1990). *Cell*. **62**, 317–329.

Chen, Y., Hojo, S., Matsumoto, N., & Yamamoto, K. (2013). *Glycobiology*. **23**, 904–916.

Clapham, D. E. (2007). *Cell*. **131**, 1047–1058.

Cortini, M. & Sitia, R. (2010). *Traffic*. **11**, 651–659.

Denks, K., Vogt, A., Sachelaru, I., Petriman, N., Kudva, R., & Koch, H. (2014). *Mol. Membr. Biol.* **31**, 58–84.

Ellgaard, L. & Helenius, A. (2003). *Nat. Rev. Mol. Cell Biol.* **4**, 181–191.

Elmahmoudi, H., Wigren, E., Laatiri, a, Jlizi, a, Elgaaied, a, Gouider, E., & Lindqvist, Y. (2011). *Haemophilia*. **17**, e923–e927.

Fiedler, K. & Simons, K. (1994). *Cell.* **77**, 625–626.

Fiedler, K., Veit, M., Stamnes, M. A., & Rothman, J. E. (1996). *Science*. **273**, 1396–1399.

Fitzgerald, D. J. D. J., Berger, P., Schaffitzel, C., Yamada, K., Richmond, T. J. T. J., & Berger, I. (2006). *Nat. Methods*. **3**, 1021–1032.

Fraldi, A., Zito, E., Annunziata, F., Lombardi, A., Cozzolino, M., Monti, M., Spampanato, C., Ballabio, A., Pucci, P., Sitia, R., *et al.* (2008). *Hum. Mol. Genet.* **17**, 2610–2621.

Franke, D. & Svergun, D. I. (2009). *J. Appl. Crystallogr.* **42**, 342–346.

Gauss, R., Kanehara, K., Carvalho, P., Ng, D. T. W., & Aebi, M. (2011). *Mol. Cell.* **42**, 782–793.

Gavel, Y. & von Heijne, G. (1990). *Protein Eng.* **3**, 433–442.

Grinna, L. S. & Robbins, P. W. (1980). *J. Biol. Chem.* **255**, 2255–2258.

Gürkan, C., Stagg, S. M., Lapointe, P., & Balch, W. E. (2006). *Nat. Rev. Mol. Cell Biol.* **7**, 727–738.

Guy, J. E., Wigren, E., Svärd, M., Härd, T., & Lindqvist, Y. (2008). *J. Mol. Biol.* **381**, 941–955.

Hamburger, A. E., West, A. P., Hamburger, Z. A., Hamburger, P., & Bjorkman, P. J. (2005). *J. Mol. Biol.* **349**, 558–569.

Hampton, R. Y. & Sommer, T. (2012). *Curr. Opin. Cell Biol.* **24**, 460–466.

Helenius, A., Trombetta, E. S., Hebert, D. N., & Simons, J. F. (1997). *Trends Cell Biol.* **7**, 193–200.

Hill, K., Boone, C., Goebl, M., Puccia, R., Sdicu, A. M., & Bussey, H. (1992). *Genetics.* **130**, 273–283.

Itin, C., Roche, a C., Monsigny, M., & Hauri, H. P. (1996). *Mol. Biol. Cell*. **7**, 483–493.

Jackson, M. R., Nilsson, T., & Peterson, P. A. (1990). *EMBO J.* **9**, 3153–3162.

Jensen, D. & Schekman, R. (2011). *J. Cell Sci.* **124**, 1–4.

Jigami, Y. & Odani, T. (1999). *Biochim. Biophys. Acta - Gen. Subj*. **1426**, 335–345.

Kanehara, K., Kawaguchi, S., & Ng, D. T. W. (2007). *Semin. Cell Dev. Biol.* **18**, 743–750.

Klumperman, J., Schweizer, a, Clausen, H., Tang, B. L., Hong, W., Oorschot, V., & Hauri, H. P. (1998). *J. Cell Sci.* **111 (Pt 2**, 3411–3425.

Kondylis, V., Pizette, S., & Rabouille, C. (2009). *Semin. Cell Dev. Biol.* **20**, 817–827.

Krammer, F., Schinko, T., Palmberger, D., Tauer, C., Messner, P., & Grabherr, R. (2010). *Mol. Biotechnol.* **45**, 226–234.

Kwok, S. C. M., Liu, X., Mangel, P., & Daskal, I. (2006). *DNA Cell Biol.* **25**, 523–529.

Lahtinen, U., Dahllöf, B., & Saraste, J. (1992). *J. Cell Sci.* **103 (Pt 2**, 321–333.

Lahtinen, U., Svensson, K., & Pettersson, R. F. (1999). *Eur. J. Biochem.* **260**, 392–397.

Lee, M. C. S., Miller, E. A., Goldberg, J., Orci, L., & Schekman, R. (2004). *Annu. Rev. Cell Dev. Biol.* **20**, 87–123.

Lievens, P. M. J., De Servi, B., Garofalo, S., Lunstrum, G. P., Horton, W. A., & Liboi, E. (2008). *Int. J. Biochem. Cell Biol.* **40**, 2649–2659.

Lin, Q.-H., Zhang, K.-D., Duan, H.-X., Liu, M.-X., Wei, W.-L., & Cao, Y. (2015). *Cancer Sci*. **106**, 1463-73.

Li, S. (2010). *Cell Res.* **20**, 1148–1157.

Lobsanov, Y. D., Romero, P. a., Sleno, B., Yu, B., Yip, P., Herscovics, A., & Howell, P. L. (2004). *J. Biol. Chem.* **279**, 17921–17931.

Locke, M. & Leung, H. (1984). *Tissue Cell*. **16**, 739–766.

Lussier, M., Sdicu, A. M., Bussereau, F., Jacquet, M., & Bussey, H. (1997). *J. Biol. Chem.* **272**, 15527–15531.

Lussier, M., Sdicu, A. M., & Bussey, H. (1999). *Biochim. Biophys. Acta - Gen. Subj*. **1426**, 323–334.

Lussier, M., Sdicu, A. M., Camirand, A., & Bussey, H. (1996). *J. Biol. Chem.* **271**, 11001–11008.

Margulis, N. G., Wilson, J. D., Bentivoglio, C. M., Dhungel, N., Gitler, A. D., & Barlowe, C. (2015). *Traffic*. **1**, 1–20.

Matsuoka, K., Orci, L., Amherdt, M., Bednarek, S. Y., Hamamoto, S., Schekman, R., & Yeung, T. (1998). *Cell*. **93**, 263–275.

Mattioli, L., Anelli, T., Fagioli, C., Tacchetti, C., Sitia, R., & Valetti, C. (2006). *J. Cell Sci.* **119**, 2532–2541.

Mohorko, E., Glockshuber, R., & Aebi, M. (2011). *J. Inherit. Metab. Dis.* **34**, 869–878.

Morais, V. A., Brito, C., Pijak, D. S., Crystal, A. S., Fortna, R. R., Li, T., Wong, P. C., Doms, R. W., & Costa, J. (2006). *Biochim. Biophys. Acta.* **1762**, 802–810.

Muñiz, M., Nuoffer, C., Hauri, H. P., & Riezman, H. (2000). *J. Cell Biol.* **148**, 925–930.

Munro, S. & Pelham, H. R. (1987). *Cell*. **48**, 899–907.

Nakanishi, H., Suda, Y., & Neiman, A. M. (2007). *J. Cell Sci.* **120**, 908–916.

Nakatsukasa, K. & Brodsky, J. L. (2008). *Traffic*. **9**, 861–870.

Neve, E. P. a, Lahtinen, U., & Pettersson, R. F. (2005). *J. Mol. Biol.* **354**, 556–568.

Nichol, H., Law, J. H., & Winzerling, J. J. (2002). *Annu. Rev. Entomol.* **47**, 535–559.

Nichols, W. C., Seligsohn, U., Zivelin, A., Terry, V. H., Hertel, C. E., Wheatley, M. a., Moussalli, M. J., Hauri, H. P., Ciavarella, N., Kaufman, R. J., *et al.* (1998). *Cell.* **93**, 61–70.

Nishimura, N. & Balch, W. E. (1997). *Science*. **277**, 556–558.

Nishio, M., Kamiya, Y., Mizushima, T., Wakatsuki, S., Sasakawa, H., Yamamoto, K., Uchiyama, S., Noda, M., McKay, A. R., Fukui, K., *et al.* (2010). *Proc. Natl. Acad. Sci. U. S. A.* **107**, 4034–4039.

Noda, Y., Hara, T., Ishii, M., & Yoda, K. (2014). *Biol. Open*. **3**, 209–224.

Nufer, O., Kappeler, F., Guldbrandsen, S., & Hauri, H.-P. (2003). *J. Cell Sci.* **116**, 4429–4440.

Nyfeler, B., Kamiya, Y., Boehlen, F., Yamamoto, K., Kato, K., de Moerloose, P., Hauri, H.-P., & Neerman-Arbez, M. (2008a). *Blood*. **111**, 1299–1301.

Nyfeler, B., Reiterer, V., Wendeler, M. W., Stefan, E., Zhang, B., Michnick, S. W., & Hauri, H.-P. (2008b). *J. Cell Biol.* **180**, 705–712.

Nyfeler, B., Zhang, B., Ginsburg, D., Kaufman, R. J., & Hauri, H. P. (2006). *Traffic*. **7**, 1473–1481.

Okamoto, M., Kurokawa, K., Matsuura-Tokita, K., Saito, C., Hirata, R., & Nakano, a. (2012). *J. Cell Sci.* **125**, 3412–3420.

Otte, S. & Barlowe, C. (2002). *EMBO J.* **21**, 6095–6104.

Otte, S. & Barlowe, C. (2004). *Nat. Cell Biol.* **6**, 1189–1194.

Otte, S., Belden, W. J., Heidtman, M., Liu, J., Jensen, O. N., & Barlowe, C. (2001). *J. Cell Biol.* **153**, 503–517.

Parodi, A. J. (2000). *Annu. Rev. Biochem.* **69**, 69–93.

Paroutis, P., Touret, N., & Grinstein, S. (2004). *Physiology (Bethesda).* **19**, 207–215.

Patterson, G. H., Hirschberg, K., Polishchuk, R. S., Gerlich, D., Phair, R. D., & Lippincott-Schwartz, J. (2008). *Cell*. **133**, 1055–1067.

Pelham, H. R., Hardwick, K. G., & Lewis, M. J. (1988). *EMBO J*. **7**, 1757–1762.

Pelham, H. R. & Rothman, J. E. (2000). *Cell*. **102**, 713–719.

Petoukhov, M. V, Konarev, P. V, Kikhney, A. G., & Svergun, D. I. (2007). *J. Appl. Crystallogr.* **40**, s223–s228.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.

Pezzati, R., Bossi, M., Podini, P., Meldolesi, J., & Grohovaz, F. (1997). *Mol. Biol. Cell*. **8**, 1501–1512.

Pimpaneau, V., Midoux, P., Monsigny, M., & Roche, A. C. (1991). *Carbohydr. Res.* **213**, 95–108.

Porter, K. R., Claude, A., & Fullam, E. F. (1945). *J. Exp. Med.* **81**, 233–246.

Powers, J. & Barlowe, C. (2002). *Mol Biol Cell*. **13**, 880–891.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., *et al*. (2009). *Proteins*. **77 Suppl 9**, 89–99.

Sánchez-Rodríguez, A., Tytgat, H. L. P., Winderickx, J., Vanderleyden, J., Lebeer, S., & Marchal, K. (2014). *BMC Genomics*. **15**, 349.

Saraste, J. & Kuismanen, E. (1984). *Cell*. **38**, 535–549.

Saraste, J., Palade, G., & Farquhar, M. (1987). *J. Cell Biol.* **105**, 2021–2029.

Satoh, T., Cowieson, N. P., Hakamata, W., Ideo, H., Fukushima, K., Kurihara, M., Kato, R., Yamashita, K., & Wakatsuki, S. (2007). *J. Biol. Chem.* **282**, 28246.

Satoh, T., Suzuki, K., Yamaguchi, T., & Kato, K. (2014). *PLoS One*. **9**, e87963.

Sato, K. & Nakano, A. (2002). *Mol. Biol. Cell*. **13**, 2518–2532.

Sauvageau, E., Rochdi, M. D., Oueslati, M., Hamdan, F. F., Percherancier, Y., Simpson, J. C., Pepperkok, R., & Bouvier, M. (2014). *Traffic*. **15**, 383–400.

Schimmöller, F., Singer-Krüger, B., Schröder, S., Krüger, U., Barlowe, C., & Riezman, H. (1995). *EMBO J.* **14**, 1329–1339.

Schneider, G. & Lindqvist, Y. (1994). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **50**, 186–191.

Schweizer, A., Fransen, J., Bächi, T., Ginsel, L., & Hauri, H. (1988). *J. Cell Biol.* **107**, 1643–1653.

Shen, B. W., Spiegel, P. C., Chang, C., Huh, J., Lee, J., Kim, J., Kim, Y., & Stoddard, B. L. (2008). *Blood*. **111**, 1240–1247.

Shibuya, A., Margulis, N., Christiano, R., Walther, T. C., & Barlowe, C. (2015). *J. Cell Biol.* **208**, 197–209.

Springer, S., Chen, E., Duden, R., Marzioch, M., Rowley, A., Hamamoto, S., Merchant, S., & Schekman, R. (2000). *Proc. Natl. Acad. Sci. U. S. A.* **97**, 4034–4039.

Stagg, S. M., Gürkan, C., Fowler, D. M., LaPointe, P., Foss, T. R., Potter, C. S., Carragher, B., & Balch, W. E. (2006). *Nature*. **439**, 234–238.

Stamnes, M. A., Craighead, M. W., Hoe, M. H., Lampen, N., Geromanos, S., Tempst, P., & Rothman, J. E. (1995). *Proc. Natl. Acad. Sci.* **92**, 8011–8015.

Stanley, P. (2011). *Cold Spring Harb. Perspect. Biol.* **3**, 1–13.

Van den Steen, P., Rudd, P. M., Dwek, R. A., & Opdenakker, G. (1998). *Crit. Rev. Biochem. Mol. Biol.* **33**, 151–208.

Stolz, A. & Wolf, D. H. (2010). *Biochim. Biophys. Acta*. **1803**, 694–705.

Südhof, T. C. & Rothman, J. E. (2009). *Science*. **323**, 474–477.

Svergun, D. I. (1992). *J. Appl. Crystallogr.* **25**, 495–503.

Takida, S., Maeda, Y., & Kinoshita, T. (2008). *Biochem. J.* **409**, 555–562.

Thor, F., Gautschi, M., Geiger, R., & Helenius, A. (2009). *Traffic*. **10**, 1819–1830.

Tisdale, E. J., Plutner, H., Matteson, J., & Balch, W. E. (1997). *J. Cell Biol.* **137**, 581–593.

Tremblay, L. O. & Herscovics, A. (1999). *Glycobiology*. **9**, 1073–1078.

Unligil, U. M. & Rini, J. M. (2000). *Curr. Opin. Struct. Biol.* **10**, 510–517.

Vainio, P., Mpindi, J.-P. P., Kohonen, P., Fey, V., Mirtti, T., Alanen, K. a., Perälä, M., Kallioniemi, O., & Iljin, K. (2012). *PLoS One*. **7**, e39801.

Velloso, L. M., Svensson, K., Pettersson, R. F., & Lindqvist, Y. (2003). *J. Mol. Biol.* **334**, 845–851.

Velloso, L. M., Svensson, K., Schneider, G., Pettersson, R. F., & Lindqvist, Y. (2002). *J. Biol. Chem.* **277**, 15979–15984.

Vollenweider, F., Kappeler, F., Itin, C., & Hauri, H. P. (1998). *J. Cell Biol.* **142**, 377–389.

Wang, J.-J., Qiu, L., Cai, Q., Ying, S.-H., & Feng, M.-G. (2014). *Fungal Genet. Biol.* **70**, 1–10.

Wang, X. H., Nakayama, K. I., Shimma, Y. I., Tanaka, A., & Jigami, Y. (1997). *J. Biol. Chem.* **272**, 18117–18124.

Welsh, L. M., Tong, A. H. Y., Boone, C., Jensen, O. N., & Otte, S. (2006). *J. Cell Sci.* **119**, 4730–4740.

Wendeler, M. W., Paccaud, J.-P., & Hauri, H.-P. (2007). *EMBO Rep.* **8**, 258–264.

Wieland, F. T., Gleason, M. L., Serafini, T. A., & Rothman, J. E. (1987). *Cell*. **50**, 289–300.

Wigren, E. (2012). Structural studies of the ERGIC-53/MCFD2 glycoprotein transport receptor complex.

Wigren, E., Bourhis, J. M., Kursula, I., Guy, J. E., & Lindqvist, Y. (2010). *FEBS Lett.* **584**, 878–882.

Wilde, M., Klausberger, M., Palmberger, D., Ernst, W., & Grabherr, R. (2014). *Biotechnol. Lett.* **36**, 743–749.

Xu, D. & Zhang, Y. (2012). *Proteins.* **80**, 1715–1735.

Zhang, B., Cunningham, M. a, Nichols, W. C., Bernat, J. a, Seligsohn, U., Pipe, S. W., McVey, J. H., Schulte-Overberg, U., de Bosch, N. B., Ruiz-Saez, A., *et al.* (2003). *Nat. Genet.* **34**, 220–225.

Zhang, B., McGee, B., Yamaoka, J. S., Guglielmone, H., Downes, K. A., Minoldo, S., Jarchum, G., Peyvandi, F., De Bosch, N. B., Ruiz-Saez, A., *et al.* (2006). *Blood.* **107**, 1903–1907.

Zheng, C., Liu, H. H., Yuan, S., Zhou, J., & Zhang, B. (2010). *Blood*. **116**, 5698–5706.

Zheng, C., Page, R. C., Das, V., Nix, J. C., Wigren, E., Misra, S., & Zhang, B. (2013). *J. Biol. Chem.* **288**, 20499–20509.