

USING COMPUTATIONAL AND EXPERIMENTAL APPROACHES TO
UNDERSTAND PROTEIN FUNCTION:
APPLICATION TO PROPROTEIN CONVERTASES AND OTHER DISEASE-
RELATED PROTEINS

By

Johannes Elferich

A DISSERTATION

Presented to the Department of Biochemistry & Molecular Biology

and the Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

June 2015

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

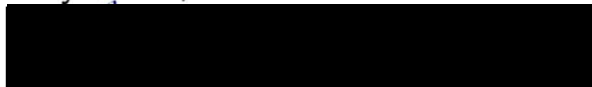
This is to certify that the PhD dissertation of
Johannes Elferich
has been approved



Ujwal Shinde, PhD



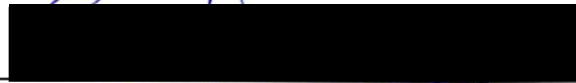
Larry David, PhD



Caroline Erns, PhD



David Farrens, PhD



Eric Gouaux, PhD



Mike Harms, PhD

Table of Contents

1	Introduction	1
1.1	Biology of proprotein convertases and their propeptides	1
1.1.1	Subtilases: Phylogenetic distribution, mechanism and activation	2
1.1.2	Proprotein convertases: Tissue distribution and functions	5
1.1.3	Propeptides: Powerful modulators of protein function	10
1.2	Biological pH-sensors and ways to study them	14
1.2.1	Cellular control of pH	16
1.2.2	The thermodynamics of pH sensing	18
1.2.3	Examples of pH-sensing proteins	23
1.2.4	Approaches to study pH sensing proteins	28
1.3	Computational approaches used in this thesis	39
1.3.1	Sequence based approaches	39
1.3.2	Structure based approaches	45
1.4	Outline of thesis	50
2	Propeptides of eukaryotic proteases encode histidines to exploit organelle pH for regulation	59
2.1	Abstract	60
2.2	Introduction	60
2.3	Material and Methods	62
2.3.1	Conservation Analysis	62

2.3.2	Data acquisition.....	62
2.3.3	Tree construction.....	63
2.3.4	Statistical testing.....	63
2.3.5	Sliding window analysis.....	63
2.4	pH-sensors in the propeptides of subtilases	64
2.5	pH-sensors in the propeptides of cathepsins	70
2.6	The cytosolic caspase family encodes no pH-sensors in their propeptides.....	72
2.7	Implications for other proteins and disease	73
3	The mechanism by which histidine protonation mediates activation	88
3.1	Abstract	89
3.2	Introduction	90
3.3	Materials and Methods	93
3.3.1	Expression and purification	93
3.3.2	CD studies.....	93
3.3.3	MD simulations	94
3.3.4	Amino acid content analysis	96
3.3.5	Enzyme activity assays.....	97
3.3.6	Isolation of <i>in trans</i> propeptide:protease complexes	97
3.4	Results	98
3.4.1	Eukaryotic propeptides harbor an internal cleavage site loop that is missing within their prokaryotic paralogs.....	98

3.4.2	CD spectroscopy demonstrates pH-dependent structural changes in eukaryotic propeptides	100
3.4.3	Swapping propeptides between PC1 and furin reassigns pH-dependent activation	102
3.4.4	Histidine protonation alters conformational dynamics of eukaryotic propeptides	103
3.4.5	The constitutively deprotonated mimic of the pH Sensor, His ₆₉ Leu-PRO ^{FUR} , is more stable than WT-PRO ^{FUR}	107
3.4.6	His ₆₉ Leu-PRO ^{FUR} is more stable toward pH-dependent unfolding ..	110
3.4.7	Molecular Dynamics of WT and His ₆₉ Leu PRO ^{FUR} :.....	112
3.4.8	Protonation of histidine residues in addition to His ₆₉ are involved in pH-sensing	114
3.5	Discussion	116
3.5.1	Propeptides are sufficient to impart pH-sensing	116
3.5.2	Mechanism of pH-dependent activation of furin.....	117
3.5.3	Implications of the pH Sensor in the activation of proprotein convertases.....	121
4	pKa determination of histidine residues in the propeptides of furin and PC1/3 using histidine hydrogen-deuterium exchange mass spectrometry.....	135
4.1	Abstract	136
4.2	Introduction:	137
4.3	Experimental Section:.....	141

4.3.1	Purification of propeptides:.....	141
4.3.2	HD exchange:.....	141
4.3.3	Mass analysis:.....	142
4.3.4	Data processing:.....	143
4.3.5	Homology modeling and pKa prediction:.....	145
4.4	Results and Discussion:.....	145
4.4.1	Identifying ions containing a single histidine residue using whole-protein electron-transfer dissociation and proteolytic digestion:.....	145
4.4.2	Individual histidine residues in PRO ^{FUR} display comparable pKa values, but vary in their solvent accessibility:.....	148
4.4.3	Alkaline pH induces HD exchange of carbon-bound protons in addition to the C2-proton in the imidazole ring:.....	150
4.4.4	His ₇₂ in PRO ^{PC1} displays a pK _{a,app} shifted to a more acidic value:.....	153
4.4.5	A structural interpretation of histidine pKa values:.....	154
4.5	Conclusion:.....	156
5	Using computational techniques to understand protein function in disease	165
5.1	Abstract.....	167
5.2	LdNT1: Understanding purine uptake of parasites on a structural level	168
5.2.1	Introduction.....	168
5.2.2	Materials and Methods.....	169

5.2.3	Results.....	170
5.2.4	Discussion	174
5.3	Otub1: Novel mechanism of p53 stabilization.....	175
5.3.1	Introduction	175
5.3.2	Materials and Methods	176
5.3.3	Results.....	177
5.3.4	Discussion	179
5.4	Ribosome-Sec61 translocon: The role of small structured domains in cotranslational secretion.....	180
5.4.1	Introduction	180
5.4.2	Material and Methods	182
5.4.3	Results and Discussion.....	182
5.5	ROS1: New drugs for treatment of small cell lung cancers.....	184
5.5.1	Introduction	184
5.5.2	Materials and Methods	185
5.5.3	Results and Discussion.....	186
5.6	Conclusion.....	187
6	Summary and conclusions	197
6.1	Histidine protonation in the propeptides of furin and PC1/3 is responsible for pH-mediated activation of cognate catalytic domains	198
6.2	Histidine enrichment is an evolutionary mechanism for propeptides of proteases to become pH-sensitive.....	199

6.3	Multiple histidine residues are involved in shaping the pH-dependent energy landscape of PRO ^{FUR}	200
6.4	Differences in pH-activation are due to modulation of the pKa-value of a conserved histidine residue	201
6.5	Combination of proteolysis, CID, and ETD fragmentation allows for measurement of residue-specific deuterium uptake in side-chains	201
6.6	Computational tools allow rapid testing and generation of biological hypotheses	202
6.7	Future directions	204
A1	Software for analysis of histidine hydrogen-deuterium exchange.....	207
A1.1	Introduction	207
A1.2	Overall design	208
A1.3	Individual Applications	209
A1.3.1	Fragment_analyzer	209
A1.3.2	Peptide_analyzer.....	211
A1.3.3	Exchange_quantifier	213
A1.4	Discussion	216
A2	List of publications.....	227
A2.1	Publication included in this thesis.....	227
A2.2	Publications not included in this thesis	228

List of Figures

Figure 1-1: Structure and Mechanism of Subtilases.....	52
Figure 1-2: Mechanism of subtilase activation.....	53
Figure 1-3: Structures of subtilase propeptides	54
Figure 1-4: Thermodynamic model of two-site titration allosteric regulation and example titration curves.....	55
Figure 1-5: Structures of low and high pH states of different proteins.....	56
Figure 1-6: Mechanism of histidine hydrogen-deuterium exchange	57
Figure 1-7: Examples for Hidden Markov model sequence profiles.....	58
Figure 2-1: Propeptides are more divergent than cognate catalytic domains.....	79
Figure 2-2: Histidines are enriched in propeptides of eukaryotic, but not prokaryotic, subtilases.....	81
Figure 2-3: Distribution of $[AA]_{Pro}$ and $[AA]_{Cat}$ for all 20 amino acids in eukaryotic and prokaryotic subtilases.....	82
Figure 2-4: Distribution of $\Delta[AA]$ for all 20 amino acids in eukaryotic and prokaryotic subtilases.....	83
Figure 2-5: Histidine enrichment exists only in propeptide domains of the Cathepsin L family, while it is also present in the occluding loop of the Cathepsin B family.....	85
Figure 2-6: The cytosolic caspase family shows no histidine bias in propeptides.	87

Figure 3-1: Comparison of sequences, structures, evolution, and composition biases of propeptides in prokaryotic and eukaryotic subtilases.	124
Figure 3-2: pH-dependent structure and function of propeptides.	125
Figure 3-3: pH-dependent structural dynamics of prokaryotic and eukaryotic propeptides.	127
Figure 3-4: His ₆₉ Leu-PRO ^{FUR} is more structured than WT-PRO ^{FUR}	128
Figure 3-5: His ₆₉ Leu-PRO ^{FUR} is more stable than WT-PRO ^{FUR} to pH-induced unfolding.	129
Figure 3-6 Molecular dynamics simulations of His ₆₉ Leu PRO ^{FUR}	131
Figure 3-7: Near-UV titration of WT and His ₆₉ Leu PRO ^{FUR}	132
Figure 3-8: Far-UV circular dichroism spectra of single-point variants of PRO ^{FUR}	133
Figure 3-9: Potential mechanisms of furin activation	134
Figure 4-1: Identification of ions for HD exchange measurements for individual histidines using the whole-protein ETD/CID approach.....	159
Figure 4-2: Identification of ions for HD exchange measurements for individual histidines using the pepsin/ETD approach.	160
Figure 4-3: pH-dependent HD exchange rate of PRO ^{FUR}	161
Figure 4-4: HD exchange in residues other than histidine.....	162
Figure 4-5: pH dependent HD exchange in PRO ^{PC1}	163
Figure 4-6: Structural interpretation of measured pK _{a,app} and maximal exchange rates	164
Figure 5-1: Ab initio computational model of LdNT1.1	189

Figure 5-2: Alignment of LdNT1.1 and the <i>e. coli</i> fucose transporter.....	190
Figure 5-3: Homology model of outward-open state of LdNT1.1 and mutations that affect activity.....	191
Figure 5-4: Crosslinks based on homology model	192
Figure 5-5: UbcH5 binds preferentially to monoubiquitinated Otub1 through backside UbcH5-Ub interaction.	194
Figure 5-6: Frequency and identity of N-terminal domains in secretory versus cytosolic and nuclear proteins.....	195
Figure 5-7: Homology models of ROS1 bound to crizotinib and foretinib	196
Figure A1-1: Parameter file for Fragment_analyzer application	219
Figure A1-2: Screenshots of Fragment_analyzer application	220
Figure A1-3: Parameter file for Peptide_analyzer application	221
Figure A1-4: Screenshots of Peptide_analyzer application.....	223
Figure A1-5: Parameter files for Exchange_quantifier application.....	224
Figure A1-6: Screenshots of Exchange_quantifier application.....	226

List of Tables

Table 2.1: Results of Mann-Whitney tests to evaluate differences in distribution of $\Delta[\text{AA}]$ between eukaryotes and prokaryotes.....	76
Table 2.2: List of human proteins with histidine enrichment in their propeptides	77
Table 4.1: Parameters of histidine HD exchange rate fits for PRO^{FUR} at different pH values using equation 2	157
Table 4.2: Parameters of histidine HD exchange rate fits for PRO^{PC1} at different pH values using equation 2	158

Acknowledgements

First, I would like to thank my advisor, Dr. Ujwal Shinde, for his guidance, support, and friendship. His excitement for research and rigorous thinking has been inspiring.

Furthermore, I would like to thank Dr. Larry David and Dr. Eric Gouaux for their mentorship and access to their laboratories.

I owe much gratitude to my research advisory committee: Dr. Larry David, Dr. Caroline Enns, Dr. David Farrens, and Dr. Eric Gouaux, for their constant support and dedication to my training.

I would like to thank Danielle Williamson for companionship and help as a fellow member of the laboratory. I also like to thank Stephanie Dillon for teaching me cell culture and introducing me to the laboratory. Furthermore, I would like to thank Glenn Goldston and Audrey Brown for help they provided during my PhD.

I would like to thank fellow researchers in the Biochemistry department for many helpful discussions and help along the way. Especially, I would like to thank Amber Jones Brunette for introducing me to American Football as well as Yoga.

I would like to thank my family for their unquestioning support and love from thousands of miles apart.

Finally, I would like to thank Nathalie Javidi-Sharifi. Her love and support has been the most important thing of all.

Abstract

This thesis describes the combination of biochemical, biophysical, and computational tools to investigate the mechanism of pH-mediated activation of proprotein convertases. Proprotein convertases are a family of human serine endoproteases that are involved in processing of hormones, enzymes, and receptors. In order to spatiotemporally control this processing in distinct organelles of the secretory pathway proprotein convertases have evolved to sense the pH of the correct organelle to mediate activation.

In this thesis I demonstrate (i) an enrichment of histidine residues in the propeptides of eukaryotic, but not prokaryotic, homologs of proprotein convertases and in propeptides of another eukaryotic pH activated protease family, (ii) that the propeptides of proprotein convertases sense pH and partially unfold at the pH of activation, (iii) and that proprotein convertases fine-tune the pH of their activation by using multiple titratable sites and adjusting the pKa values of a conserved histidine.

Furthermore, this thesis describes other examples where the combination of computational biology with experimental methods allows a deeper understanding of the biology of proteins relevant in multiple diseases, ranging from protozoan parasites to cancer.

Introduction

This thesis describes how combinations of experimental and computational approaches can be used to understand how proteins function. I will begin by introducing the biology of a protease family called proprotein convertases, which is the focus of the majority of this thesis. I will then introduce the concept of biological pH-sensors, which are protein whose biological function is modulated by the titration of side-chains within physiologically relevant pH ranges. A strong emphasis is put on the experimental and theoretical tools that are available to study these phenomena. Finally, I will describe algorithms that allow us to understand proteins at the level of sequence and structure. In the course of this introduction I want to convey how synergy between experiments and computation allow us to overcome many challenges that hinder investigation of protein function.

1.1 Biology of proprotein convertases and their propeptides

Proprotein convertases are serine endoproteases found in all metazoans. They are part of the larger family of subtilases, which are found in all domains of life [1]. Since many aspects of function and activation are shared by subtilases in eukaryotes, prokaryotes, and archaea I will begin by describing the commonalities of subtilases, mainly their catalytic mechanism and activation.

Subsequently, I will describe the unique features of proprotein convertases and finish by giving a review of the mechanisms by which propeptides modulate structure, function, folding, and dynamics of cognate proteins.

1.1.1 Subtilases: Phylogenetic distribution, mechanism and activation

Subtilases are defined by their sequence similarity with bacterial subtilisin. They share a common fold, named the subtilase fold, and an arrangement of three residues, called the catalytic triad that, as the name suggests, is responsible for the catalysis of endoproteolytic digestion of proteins (Figure 1-1). Despite their commonality in chemical function, hydrolysis of peptide bonds, they differ vastly in their biological roles. While bacterial and archaeal homologs are mostly scavenger enzymes that digest extracellular proteins into peptide fragments for uptake into the host cell [2], many eukaryotic homologs are involved in the activation, not degradation, of proteins. The most prominent example is the proprotein convertase family, which is described in detail later. Other examples include plant subtilases, which appear to be critical in signal cascades triggered by pathogen interaction [3], and a subtilase in the malaria parasite *plasmodium falciparum* that is critical for egress of the parasite from the host cell [4]. The astonishing variability in biological roles is strikingly highlighted by the human homolog PCSK9, which despite a fully intact active site does not cleave any substrates but appears to function by binding to a surface receptor leading to receptor internalization [5]. This variability is caused by multiple factors, such as their sequence specificity, regulation of their activity within the cell and specific protein-protein interaction.

The active site of subtilases consists of three residues, a serine, histidine, and aspartate. Figure 1-1 A shows the arrangement of the catalytic triad within the subtilase fold as found in the crystal structure of Tk-subtilisin [6]. Based on this structure and other studies the following catalytic mechanism can be proposed [7] (Figure 1-1 B). The Michaelis-complex is formed by binding of the substrate backbone to the active site. Residues situated at the N-terminus of the cleavage site (P1, P2, etc.) have to be accommodated by the protease to allow for optimal orientation of the peptide bond. Specifically, P1 and P2 must fit into pockets, whose chemical environment is decisive in sequence specificity [8]–[11]. In most subtilases the residues C-terminal to the cleavage site (P1',P2', etc.) are not directly involved in specificity, although a recent crystal structure of subtilisin-like protease 1 (SUB1) of *plasmodium falciparum* suggested some specificity for these residues [12]. The hydroxyl group of the active site serine then attacks the carbonyl-carbon of the substrate backbone, with the C-terminal part of the substrate acting as the leaving group, resulting in an acylenzyme complex. The hydrogen bound to the serine hydroxyl is buffered by the active site histidine, whose electronegativity is increased by the active site aspartate. Nucleophilic attack by a water molecule allows for release of the N-terminus of the substrate, again with the active site histidine involved in proton buffering. This step happens with relatively slow kinetics compared to the C-terminal part since the residues P1 and P2 form substantial interactions with the enzyme.

This prolonged interaction is important in the activation mechanism of many subtilases. Like most proteases subtilases are expressed as zymogens, meaning

they are initially inactive and require post-translational modification for activation. In the case of subtilases the required modification is removal of an N-terminal propeptide [13], which is usually ~80 residues long, but can be substantially longer in some homologs. This propeptide is required for the folding of the protease and therefore also called intramolecular chaperone (IMC)[14]. The role of the propeptide as a chaperone is described in more detail later in this chapter (Section 1.1.3). After folding, the peptide bond between propeptide and protease occupies the protease active site and is primed for cleavage (Figure 1-2 B) [6]. Due to this optimal placement cleavage occurs rapidly and spontaneously in an intramolecular reaction [15]. The C-terminus of the propeptide, however, remains in the active site due to interactions formed by P1, P2, and a large interface between the beta-sheet of the propeptide and helices of the protease [6], [16] (Figure 1-2 B). This complex is inactive as the propeptide C-terminus blocks access of substrate. Final activation requires proteolytic degradation of the propeptide. In bacterial subtilisin this is a stochastic process *in vitro*, as activation requires a random dissociation event which leads to a chain-reaction, where one active protease molecule can proteolyze propeptides from other inactive complexes [17]. *In vivo*, where active protease already exists in the extracellular environment, inhibition of catalytic activity by the propeptide is likely a mechanism to delay activation until the protease reaches the extracellular environment. In other cases, such as the proprotein convertases, this last step requires an additional biochemical cue to trigger the activation [18] (Figure 1-2).

1.1.2 Proprotein convertases: Tissue distribution and functions

Proprotein convertases are the subtilase homologs found in metazoans. Their name derives from their primary function, namely converting proteins from their initial proforms into their mature forms by endoproteolysis. In humans there are nine members: PC1/3, PC2, Furin, PC4, PC5/6, PACE4, PC7, SKI-1/S1P, and PCSK9 [19]. However, only the first seven belong to the homology group of proprotein convertases/kexin, while SKI-1 belongs to the pyrolysins homology group and PCSK9 belongs to the Proteinase K family [1]. The first seven members share a common sequence specificity, cleaving after one or two basic residues, which can be rationalized by the high negative charge in the P1 and P2 pockets of their protease domains [8], [9]. As described earlier, PCSK9, at the current state of knowledge, does not cleave any substrates other than itself during activation, since the propeptide never dissociates [20]. While the proprotein convertases are very similar in their cleavage specificity, they differ in subcellular and tissue localization leading to different sets of substrates even though some overlap exists.

PC1/3 and PC2 are the classic neuroendocrine convertases, expressed exclusively in neuroendocrine tissue [21], [22]. They are maximally active in mature secretory granules [23] where they process a wide array of neuropeptides and peptide hormones, such as insulin [24], glucagon [25], [26], and POMC derived peptides [27]. While many of these substrates are specific to either PC1/3 or PC2, some peptides such as proinsulin are processed by both proteases and in single knockouts of either enzyme the other one can at least partially

compensate. Therefore, single knockout mice are viable, although they show metabolic diseases, while the double knockout is embryonic lethal in mice [28]. In patients mutations in PC1/3 have been associated with obesity [29], which might be connected to the role of PC1/3 in processing insulin and glucagon.

PC4 is uniquely expressed in the germ line and PC4 knockout in mice leads to infertility in males [30]. This is believed to be due to the role of PC4 in activating metalloproteases of the ADAM family at the cell surface of sperm [31]. PC4 seems to also play roles in the ovaries and the placenta, but its function there is redundant with other proprotein convertases leading to no knockout phenotype [32].

PC5, PACE4, PC7, and furin are all nearly ubiquitously expressed and show redundancies both *in vivo* and *in vitro*. Furin and PC7 are active in the TGN, cell surface, and endosomes [33], [34], while PC5 and PACE4 seem to be active mainly at the cell surface [35], [36]. PC5, PC7, and furin have a C-terminal transmembrane helix, which tethers them to the plasma membrane, but furin and PC5 can be shed into the extracellular matrix [37]. PACE4 has no C-terminal transmembrane helix, although some isoforms might be membrane-associated due to a C-terminal hydrophobic cluster [38]. Despite these apparent overlaps in specificity, expression and cellular localization, knockouts in mice have distinct phenotypes. PC5 knockout mice die at birth, due to a lack of kidneys and other developmental defects [39]. This seems to be due to PC5's unique role in processing growth/differentiation factor 11 (GDF11), as GDF11 knockout mice share the phenotype and GDF11 seems to be cleaved by PC5 with the highest

efficacy. PACE4 knockout is lethal in about 25% of mice embryos due to defects in cardiac development and bone morphogenesis [40]. This might be connected to the role of PACE4 in processing the axis determinants Nodal and Lefty. PC7 knockout mice are viable and healthy suggesting high redundancy with other proprotein convertases [41].

Furin knockout mice die at embryonic day 11 because of cardiac ventral closure defects [42]. Furin has the widest array of identified substrates ranging from TGF β -like growth factors, such as Lefty and BMP10 [43], the insulin receptor [44], adhesion molecules [45], [46], metalloproteases [47], and proton pump V-ATPase subunits [48]. Additionally, many pathogens rely on furin to process their proteins. Examples include gp160 of HIV [49] and the anthrax toxin [50]. Even though processing of these substrates is observed in tissue-specific knockouts, it is often severely impaired suggesting that furin is the major constitutive processing enzyme in the secretory pathway [51]. Probably due to this central role in many processing events no disease-causing mutations in furin are known, due to strong embryonic lethality of such mutations. However, furin plays central roles in many diseases. Besides the involvement of processing pathogenic proteins as described above, furin is often overexpressed in cancer [52] and atherosclerotic lesions [53]. In both cases, the role of furin in activating matrix metalloproteases is thought to contribute to cancer metastasis and atherosclerotic lesion rupture, respectively. Therefore inhibition of furin could be a therapeutic strategy against anthrax [54], HIV [55], and cancer [56]. However, practical use of furin inhibitors is hindered by its plethora of side effects. Clearly,

strategies that inhibit furin in a more specific manner, i. e. within specific cells or organelles, are needed. An example of such a specific inhibitor is the furin propeptide.

The activation of furin closely resembles the activation of subtilisin described earlier (Figure 1-2 A). A signal peptide targets furin for cotranslational insertion into the endoplasmic reticulum (ER), where it folds and readily cleaves the propeptide after Arg₁₀₇ of the ⁷²RTKR₇₅ motif at the C-terminus of the propeptide, which represents a strong furin cleavage motif [57]. After this cleavage the propeptide stays associated with the protease domain of furin, functioning as a temporary inhibitor with an apparent affinity of ~10 nM (Figure 3-5). Only after the complex has trafficked to the early Trans-Golgi network (TGN) the slightly acidic pH (6.5) triggers an internal cleavage at residue Arg₇₂ [18], [58]. This is most likely due to protonation of histidine residues in the propeptides of furin, since mutation of a conserved histidine, His₆₉, to leucine blocks the second cleavage upon acidification [59]. The second cleavage results in active and mature furin, which is then trafficked to the cell surface where some fraction stays associated with the cell-surface and cycles back to the TGN, while another fraction is shed into the extracellular space. Therefore, furin activity is present in throughout most of the secretory pathway, from the early TGN to the cell surface. One rationalization of this activation pathway is that the inhibition of furin by the propeptide in the endoplasmic reticulum may protect proteins with exposed dibasic motifs during the folding process that are not intended to be cleaved.

Other paralogs of the proprotein convertase family are also activated by removal of the propeptide, but the second internal cleavage occurs within different compartments. For example PC1/3 rapidly cleaves its propeptide in the endoplasmic reticulum, but it remains associated with the protease domain. Furthermore PC1/3 requires processing at the C-terminus of the protease domain, which is auto-catalyzed *in vitro*, but *in vivo* only occurs after the protein complex trafficks into the mature secretory vesicles, which have a pH of about 5.5. This suggests that the second cleavage of the PC1/3 propeptide requires a more acidic pH [23]. Interestingly, unlike PC1/3 and furin, PC2 does not undergo primary processing in the ER, but only after trafficking to secretory granules. It also requires the chaperone 7B2 in addition to its propeptide to fold properly and the C-terminus of 7B2 also acts as an inhibitor [60], [61]. The details of this mechanism are still unclear, but, similar to PC1/3 and furin, activation of PC2 in secretory granules seems to be triggered by low pH [62]. In the case of PC7, the propeptide dissociates readily from the protease domain in humans, while the rat propeptide can be immunoprecipitated with the protease [34]. Apparently, no second cleavage is required and intact propeptide is secreted into the media [34]. The activation of PC4 is not well established, but recombinant expression of PC4 in mammalian cell results in active enzyme that can be recovered from the cell [63]. PC5/6 undergoes primary processing in the ER, but secondary processing does not occur during trafficking through the secretory pathway, resulting in membrane bound PC5/6:propeptide complex that undergoes final activation at the plasma membrane, probably through interaction with heparin sulfate

proteoglycans. Since PACE4 behaves similar to PC5/6 it probably uses a similar mechanism [36].

1.1.3 Propeptides: Powerful modulators of protein function

Even though propeptides are degraded and are not part of the active enzymes, they are essential for the biogenesis of subtilases and modulate their cognate protease in many ways. Having to maintain the precise arrangement of the catalytic triad, the protease domains appear to evolve slowly when compared with their rapidly evolving cognate propeptides. Even though all propeptides of subtilases whose structure have been determined share a common fold (Figure 1-3), their sequences have diverged substantially leaving almost no recognizable sequence similarity between different families. Hence, despite similar enzymatic activity of subtilases, much of the diversity in their biological roles may be due to differential regulation by their propeptides. Propeptides are not unique to subtilases, or even proteases in general, but are found in many proteins, such as hydrolases, cell adhesion molecules and growth factors. About 700 of the 20,000 (~ 3.5%) reviewed human proteins in the UniProt database are annotated with a propeptide.

In many cases propeptides are essential for folding cognate proteins and are therefore called intramolecular chaperones. In contrast to molecular chaperones such as GroEL or Hsp90, these propeptides act as true foldases [14]. That means that they accelerate the rate of folding instead of just preventing misfolding. In some proteins, including many subtilases, this is necessary due to the way by which their tertiary structure is stabilized. In contrast to most

globular proteins, the native state of these proteins has a higher free energy than both the unfolded and a partially folded molten globular state [14], [64]. These remain in their correctly folded conformation because the rate of unfolding is extremely slow. Since the equilibrium favors the unfolded state and the equilibrium constant is equal to the ratio of the folding and unfolding rate this necessitates that the rate of folding must be even slower. In order to reach the folded state these proteins therefore require a catalyst for the folding reaction and an interaction partner that initially shifts the equilibrium towards the folded state. The propeptide can fulfill both these roles. Since the propeptide is degraded after folding the protease is trapped in the folded state. This role of the propeptide is not unique to subtilases, but also found in alpha-lytic protease [64], lipases [65], and cathepsins [66]. While this concept, called kinetic stability, has been shown for bacterial subtilisin, there is no direct evidence that the eukaryotic subtilases are also kinetically stable, just the fact that they require the propeptides for folding [67], [68]. Kinetic stability has been demonstrated for the eukaryotic cathepsins, a lysosomal protease family that is amenable to *in vitro* folding studies [66].

A fascinating finding is that modulation of the folding pathway by the propeptides can lead to differences in the activity of the mature protein. For example, point mutations in the propeptides of subtilisin E have to been shown to change kinetic parameters of subtilisin E [69] and cathepsin E shows differences in catalytic efficiency and inhibitor specificity after being folded by the propeptides of cathepsin D [70]. Modification of substrate specificity by

mutagenesis of the propeptide has also been demonstrated for subtilisin natto kinase [71], carboxypeptidase Y [72], and *Rhizopus oryzae* lipase [73]. Since the propeptide is no longer part of the active enzyme, these results suggest that the nature of the folded state is not independent of the folding pathway, a phenomenon that has been called “folding memory”. However, more research is needed to understand the mechanism of this phenomenon and to characterize the differences of the folded state on a structural level.

Besides folding of the cognate protein, an important role of propeptides is inactivation of the immature protein. Initial synthesis of enzymes as inactive precursors, called zymogens, that require activation by proteolysis is especially common in proteases, but also found in other enzymes such as hydrolases [74], [75]. Besides enzymes, other proteins are kept in inactive forms by their propeptides, such as integrins [45] or the later described anthrax protective antigen [50] (Section 1.2.3). However, the mechanism by which the propeptide is removed to activate zymogens and other precursors varies from protein to protein. In many zymogens, the proform of the enzyme is not yet correctly folded for activity and a conformational change towards the active structure is triggered by removal of the propeptide. One such example is trypsin, where removal of the short N-terminal propeptide allows for formation of a salt-bridge between the new amino-terminus and Asp₁₉₄, which triggers a conformational change that forms the substrate binding site [76]. Since in this case the proform is not capable of cleavage, an external protease is needed for activation. In the case of trypsin this enzyme is enterokinase [77]. This mechanism allows for stringent

protection of premature activation, which in this case limits trypsin activity to the intestine. In other enzymes, such as the subtilases, the enzyme folds into its active structure and is capable of cleaving its own propeptides. In these cases the propeptides also acts as a noncovalent inhibitor that remains bound even after cleavage from the enzyme. As described for bacterial subtilases and proprotein convertases (Section 1.1.1 and 1.1.2), final activation in this case requires additional degradation of the propeptides. This mechanism, in contrast to the one described earlier, allows activation to become independent of another protease, and thereby allows the integration of other signaling factors, such as pH or influx of calcium ions.

In several proteins propeptides can mediate protein-protein interactions that prime them to be in proximity with other proteins after activation or facilitate their trafficking. Deciphering the effect of propeptides on protein sorting is not trivial because removal of N-terminal sequences often removes the signal peptide that targets proteins for secretion. However, even if only the propeptide is removed without removing the signal peptide many proteins show trafficking defects. Bacterial proteins accumulate inside the cell and eukaryotic proteins accumulate in the ER. This is probably not a direct consequence of the requirement for propeptides for trafficking, but is likely due to misfolding of the cognate protein. In some cases direct involvement of the propeptides in trafficking has been established. For example the signals for targeting carboxypeptidase Y and proteinase A into the yeast vacuole are contained in the propeptides [78], [79]. Another example of the diverse roles propeptides can play

in sorting proteins is human cathepsin B. While cathepsin B has a signal peptide that under normal circumstances targets cathepsin B to the ER and ultimately the lysosome, some transcript variants lack the exon that contains the canonical start codon [80]. Under these conditions translation starts at Met₅₂, which results in removal of the signal peptide and parts of the propeptides. This exposes an amphipathic helix in the propeptide, which targets procathepsin B to mitochondria, where it triggers apoptosis [81]. Roles for propeptides in mediating protein-protein interaction are prominent in metazoan caspases. Specifically the initiator caspases that trigger the activation of the caspase cascades possess interaction domains in their propeptides that tether them to the larger protein complexes. In the case of caspases 8 and 9, death effector domains in their propeptides interact with the death-inducing signaling complex, while caspases 1, 4, and 5 encode caspase recruitment domains in their propeptides that recruit them to inflammasomes [82].

In summary, these examples described here show how propeptides are essential for correct biological function. Although they are not part of mature enzymes, propeptides can drastically modulate protein function. They often appear to be involved in sensing biological signals to mediate temporal and spatial regulation of their cognate proteins.

1.2 Biological pH-sensors and ways to study them

Addition or removal of a proton is the smallest possible biochemical reaction. Since almost all biological processes happen in aqueous solution, with a nearly unlimited supply of protons, the effective concentration or activity of protons,

most commonly expressed as the negative decadic logarithm (pH), is an important characteristic of a biological compartments. Since proteins possess many titratable chemical groups their properties depend heavily on pH. Even though protons are the smallest chemical group, they confer one positive charge unit upon binding, thereby affecting electrostatics, and are able to convert hydrogen-bond donors and acceptors.

In contrast to other chemical modifications of proteins, such as phosphorylation, acetylation, or deamidation, protonation is extremely fast and completely reversible. Due to this fact the protonation of a site as a function of pH can be modeled just using a thermodynamic constant, without having to consider kinetics. The constant used is pKa, which is defined as the pH at which a titratable site is 50% protonated, both in time- and ensemble-average, which due to the rapid kinetics of protonation can be considered identical. While these fast kinetics, together with the rapid diffusion of protons, make pH a useful regulatory mechanism for the cell, the rapid rates make experimental determination of protonation state difficult.

The titratable sites in proteins are their amino- and carboxy-termini as well as the side-chains of multiple residues. The carboxyl groups of aspartate and glutamate both have acidic pKa values of about 4.4. The imidazole group of histidine is the only side-chain with a close to neutral pKa of about 6.5. Cysteine and tyrosine have slightly alkaline pKa values of 8.5 and 10.0, respectively, with both being uncharged at pH 7.0. Lysine and arginine have pKa values of 10.0 and 12.0, respectively and are positively charged at neutral pH. Additionally,

phosphate groups bound to serine or tyrosine residues are titratable as well with pKa values close to 7.0 [83], [84].

In the following section, I will first describe how cells regulate the pH of their compartments, which is critical for correct function of proteins. Subsequently, I will describe the underlying thermodynamic theory of allostery between protein structure and protonation. I will then review several examples of proteins that actively exploit the change in properties of titratable sites to use pH for regulation of their biological function. Finally, I will describe methods by which pH-sensors in proteins can be studied.

1.2.1 Cellular control of pH

Under normal physiological conditions human cells maintain a cytosolic pH of 7.2 and an extracellular pH of 7.4. The cytosolic pH is maintained by buffer systems, such as phosphate and $\text{CO}_2/\text{HCO}_3^-$, which has the advantage of CO_2 being able to diffuse across membranes, thereby drastically increasing the buffering capacity, and by active regulation using pH-regulated proton pumps. The cell can raise its intracellular pH, to counteract an acid load, using V-type H^+ pumps which use ATP as energy source, Na-H exchangers that use the intrinsic Na gradient of the cells or Na- HCO_3^- -cotransporters, which also use the intrinsic Na gradient. Lowering the pH is achieved by passive diffusion of H^+ and HCO_3^- , due to the intrinsic membrane voltage, and by Cl- HCO_3^- -exchangers [85].

Since intracellular pH is tightly controlled, small changes in pH can have drastic effects. For example decrease of the intracellular pH by 0.3-0.4 pH units triggers apoptosis [86], while small increases of 0.2-0.3 pH units promote cell

proliferation [87]. Changes in intracellular pH have also been shown to be critical in cell migration [88]. The role of pH as a dynamic regulator of cell function has only recently been appreciated, because measurement of cytosolic pH is not trivial. While microelectrodes can be manufactured that are small enough to be used on cells, their introduction causes damage to the cell membrane. Initial approaches that aimed to overcome this problem were measurements of the passive diffusion of radiolabeled weak acids and bases across the membrane or to use NMR to measure the spectra of ^{31}P . More recently, the most convenient way to measure intracellular pH is through the use of pH-dependent fluorescent dyes. While chemical fluorescent dyes can be used, strongly pH-dependent variants of GFP have been developed, which are easy to introduce and target to specific organelles [89].

While cytosolic pH is regulated closely around 7.2, organelles can have drastically different pH values. Most drastically, lysosomes are strongly acidic with a pH of 4.8 to 5.0. This provides an optimal environment for the hydrolytic enzymes found in lysosomes, which are organelles that are involved in the degradation of biomolecules. The requirement for acidic pH of these hydrolytic enzymes provides protection for the cell should these enzymes leak into the cytosol. In addition, most other cellular compartments display different pH values compared to the cytosol. While the ER has a similar pH to the cytosol, the downstream compartments of the secretory pathway become increasingly acidic. The pH of the Golgi network is about 6.4, while secretory granules acidify further until they reach a pH of about 5.4. The endocytotic pathway also employs

a pH gradient, with early endosomes being slightly acidic with a pH of 6.2 and late endosomes having a pH of about 5.2. The pH of these compartments is maintained by a flux equilibrium of H⁺ leak over their membrane and the action of ATP-dependent V-type H⁺-pumps [90].

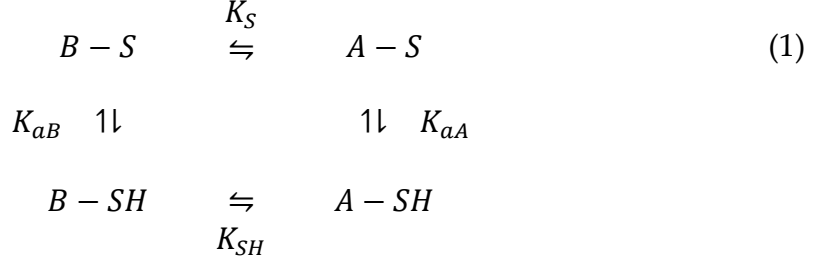
In summary, pH is used by the cell both as a temporal regulator, as seen in apoptosis, and as a spatial regulator, as observed in the secretory pathway. Understanding how proteins sense these changes in pH is therefore critical for our understanding how their function is regulated within the cell.

1.2.2 The thermodynamics of pH sensing

Before describing concrete examples of how proteins use the pH of the cell to regulate their activity, I will discuss the thermodynamic theory behind pH-dependent conformational changes. I will begin with the simplest possible case of a two-state model with a single titratable site before describing more complicated cases with multiple titratable sites and cooperativity between them. These fundamental considerations about the thermodynamic theory of protonation has been established by Charles Tanford [91].

Two-state model with a single titratable site

The simplest possible case is a protein with two conformational states, one higher populated at basic pH, B, and one more populated at acidic pH, A. Furthermore, we assume that despite the presence of multiple titratable sites only one influences the equilibrium between the two states and acts as a pH-sensor. This site either exists in a deprotonated state S or the protonated state SH. From these assumptions we can construct this simple thermodynamic model:



Since B is the state preferentially assumed at high pH we can assume that $\left[K_S = \frac{[A-S]}{[B-S]} \right] < 1$ and $\left[K_{SH} = \frac{[A-SH]}{[B-SH]} \right] > 1$. The proton dissociation constants K_{aB} and K_{aA} are defined as follows:

$$K_{aB} = \frac{[B-S][H^+]}{[B-SH]} \quad K_{aA} = \frac{[A-S][H^+]}{[A-SH]}
\tag{2}$$

Since they describe the titration equilibrium under the assumption that the protein remains in one structural state, I will refer to these constants in the future as “microscopic” constants, or in the negative decadic logarithmic form as “microscopic” pKa values. Since the equilibrium constants are connected by a thermodynamic cycle it follows that

$$\frac{K_{aB}}{K_{aA}} = \frac{K_S}{K_{SH}}
\tag{3}$$

Since K_S and K_{SH} are not equal K_{aB} and K_{aA} must also be unequal. This means that the pKa of the titratable site must be different in the two conformational states, a basic principle that is often used to identify allosteric titratable site in proteins, either by computational methods or in cases where the protein can be locked in the two states by factors other than pH.

However, in most cases these equilibrium constants are not readily accessible by experiments. In most cases one can only measure the effective dissociation of site S averaged over both conformational states. We can write the fraction of unprotonated site f_S as follows:

$$f_S = \frac{[A - S] + [B - S]}{[A - S] + [B - S] + [A - SH] + [B - SH]} \quad (4)$$

Similarly we can write the fraction of conformation B, f_B , as:

$$f_B = \frac{[B]}{[A] + [B]} \quad (5)$$

However, in most cases we cannot measure the absolute fraction f_B , but only a biophysical value that depends on f_B , such as circular dichroism or fluorescence, as a function of pH. Since we don't know if the value we reach at very low or very high pH represents a complete shift towards one site of the equilibrium (i.e. $K_S = 0$ and $K_{SH} = \infty$), it is more appropriate to define the apparent extent of conversion y_B :

$$y_B = \frac{f_B - \frac{K_S}{1 + K_S}}{\frac{K_S}{1 + K_S} + \frac{K_{SH}}{1 + K_{SH}}} \quad (6)$$

Using equations 2, 3, 6, and 4 one can derive that both y_B and f_S are equal and can be described as a Hendersson-Hasselbalch like equation with an apparent constant K_a^*

$$f_S = y_B = \frac{K_a^*/[H^+]}{1 + K_a^*/[H^+]} \quad (7)$$

With the apparent equilibrium constant K_a^* defined as

$$K_a^* = \frac{K_{SH}(1 + K_S)}{K_S(1 + K_{SH})} K_{aB} = \frac{1 + K_S}{1 + K_{SH}} K_{aA} \quad (8)$$

Therefore in this simple case one would expect any spectroscopic or similar measurements that depend on f_B to fit to a Henderson-Hasselbalch equation as a function of pH, with a midpoint that is equal to the apparent pK_a^* of the titratable site. Since this is the equilibrium constant that described the actual protonation of the titratable site I will refer to this value from here on as

“macroscopic” pKa. The macroscopic pKa will naturally be an average between the two microscopic pKa values, weighted by the difference between K_S and K_{SH}

Two state-model with multiple titrating sites

This model can be extended to two or more titrating sites. In order to account for two titrating sites one can introduce another titrating site, T, and the model shown in equation 1 is extended accordingly (Figure 1-4). This scheme is parameterized by 12 equilibrium constants: four that describe the equilibrium between the A and B state at each titration state, K_{ST} , K_{SHT} , K_{STH} , and K_{SHTH} , and eight that describe the protonation equilibria. Two equilibria describe the protonation of individual sites in both conformational states, depending on whether the site is protonated first or second (See Figure 1-4 A). The titration curves for site S and T, f_S and f_T , as well as the degree of apparent conversion, y_B , can be described mathematically, under the assumption that in the acidic state the equilibrium constants for each site is independent of the other site, since the protein loses structure ($K_{aAS1} = K_{aAS2} = K_{aAS}$ and $K_{aAT1} = K_{aAT2} = K_{aAT}$).

Theoretical modeling of several example cases with arbitrarily chosen parameters shows that this system can have multiple complex behaviors (Figure 1-4 B-E). The titration curves of S and T are no longer necessarily equal to the apparent degree of conversion, the slopes of the curves can become shallower or steeper, and can even show multiple steps.

For example if the two titrating sites have very different pKa values, they titrate essentially independently, but the degree of structural conversion shows a two-step titration (Figure 1-4 B). If the pKa values are sufficiently similar,

cooperativity effects can emerge. If the titration of a single site has only little effect on the overall equilibrium ($K_{ST} \cong K_{SHT} \cong K_{STH}$), negative cooperative effects emerge (Figure 1-4 C). In this case the titration curves show a two-step behavior as only one of them can be protonated, while the structural conversion is shifted towards more acidic pH values. In practice this can emerge if two titratable sites are in close proximity and the protonation of one site is stabilized by the other, due to hydrogen bonding.

The two previous examples are given under the assumption that protonation leads to almost complete conversion of state B to state A ($K_{ST} \cong 0$ and $K_{SHTH} \cong \infty$). If this is not the case, the two sites can titrate very similar to a one-proton mechanism, while the overall structure conversion is acid-shifted and has a steeper slope (Figure 1-4 D). Finally the titrating sites can show positive cooperativity, leading to much steeper slope, both for the titration curves of S and T and the structural conversion (Figure 1-4 E). Under these conditions ($K_{ST} \cong 0$, $K_{SHTH} \cong \infty$, and $K_{aAS} \cong K_{aAT}$) the titration curves can be described as a function of the square of the proton concentration:

$$f_S = F_T = \gamma_B = \frac{K_a^*/[H^+]^2}{1 + K_a^*/[H^+]^2} \quad (9)$$

This can be easily extended to multiple sites and written equivalent to the Hill equation:

$$f_S = F_T = \gamma_B = \frac{K_a^*/[H^+]^n}{1 + K_a^*/[H^+]^n} \quad (10)$$

where n is the number of titratable sites. This behavior is probably very easily explained from a structural perspective. If protonation of two sites contributes to

destabilization of a protein, positive cooperativity will almost certainly emerge. Since protonation at one site destabilizes the protein and destabilization of the protein favors protonation (see equation 3) this results in positive cooperativity. Therefore by placing multiple titrating sites at different positions that favor either the protonated or unprotonated form, protein can easily evolve positive cooperative pH-dependence.

In summary one can appreciate that by coupling the titration of multiple sites to the equilibrium between two structural states, multiple complex titration behaviors can emerge, which allows a protein to fine-tune its response to changes in pH. In particular, positive cooperativity allows for responses that occur over small changes in pH, like protein encounter in cells.

1.2.3 Examples of pH-sensing proteins

Since the cell goes to considerable lengths to ensure precise control of the pH in various compartments, it is not surprising that proteins use differences between organelles and in some cases minor changes of the pH within a compartment to regulate their function. While all proteins have intrinsically pH-dependent properties due to their numerous titratable groups, in many proteins specific titratable groups are strategically placed in order to achieve desired changes in structure and function as a consequence of changes in pH. The propeptides of proprotein convertases are of course the example that is the subject of this thesis. In the following sections I will discuss several examples of other biological pH-sensors in order to demonstrate how during evolution the same principle has been used on numerous occasions.

Hemoglobin

Hemoglobin, the protein responsible for the carriage of oxygen from respiratory organs to all other tissues, is optimized in several ways to allow rapid uptake of oxygen in the lung and release it in tissues. One of these mechanisms exploits the fact that solubilized CO₂ will react with water to form bicarbonate and protons, a process that is catalyzed by the enzyme carbonic anhydrase. Therefore blood that has high carbon dioxide levels is more acidic than blood with low carbon dioxide levels. Hemoglobin's affinity for oxygen is lower at acidic pH than at neutral pH, which is commonly referred to as the Bohr effect. This leads to rapid release of oxygen in peripheral CO₂ rich blood and uptake of oxygen in CO₂-depleted blood in the lungs [92].

Inspection of the crystal structure of human Hemoglobin indicated that His_{β146} forms a salt bridge with Asp_{β94} in the low oxygen-affinity T state. This salt bridge is broken in the high oxygen-affinity R state. Therefore, higher proton concentration would shift the equilibrium towards the protonated state of His_{146b} and therefore towards the low-affinity T-state (Figure 1-5a) [93], [94]. Consistent with the thermodynamic theory described in Section 1.3.2 pKa measurement of histidine side-chains in the T and R state showed that the pKa of His_{β146} changes from 6.5 in the R state to 8.0 in the T state [95]. While this difference was substantially larger than observed for other histidine residues, additional titratable sites probably contribute to the Bohr effect, since removal of His_{β146} only reduces but does not abolish the Bohr effect [96]. While other histidine residues show much smaller differences in pKa between T and R state,

Berenbrink estimates that they contribute about 40% of the total Bohr effect [97]. Additionally, the amino group of the alpha-chain N-terminus has been shown to bind chloride in the R state and not in the T state, and therefore could explain the chloride dependence of the Bohr effect. Interestingly, the number of titratable sites, as estimated by the histidine content from sequence data seems to correlate well with experimentally determined buffer values between different vertebrate species. This indicates that during evolution the sensitivity of Hemoglobin towards protons is fine-tuned by the number of titratable sites [97].

Anthrax protective antigen

The pH gradient of the secretory and endocytotic pathway is not only used by endogenous proteins, but also exploited by pathogens. Many viral and bacterial pathogens use the acidification of the endosomal compartment to trigger mechanisms for host-cell penetration. Also, many viral capsids mature during their transit through the secretory pathway and exploit the pH-gradient to regulate their maturation steps, as exemplified by the dengue virus [98]. Here I want to describe the mechanism of the anthrax toxin as an example, since it exploits the pH gradient of the endocytotic pathway in several ways and is also dependent upon furin.

Anthrax toxin is secreted by *Bacillus anthracis*, the causative agent of anthrax. It consists of three proteins: Lethal Factor (LF), Edema Factor (EF), and Protective Antigen (PA) [99]. LF and EF are enzymes that are inserted into the cytosol, where they inactivate MAP kinases and generate cAMP. PA is responsible for delivery of LF and EF into the cytosol. PA is bound by receptors on the cell in a

83 kDa form, which is subsequently activated by endoproteolysis to a 63kDa form by furin[50]. This allows self-assembly into a heptameric ring that can bind up to three molecules of LF and/or EF. This complex is then endocytosed and the following decrease in endosome pH triggers formation of a pore by PA, which allows entry of EF and LF into the cytosol.

The hypothesis that low pH triggers the formation of a transmembrane pore by the PA63 prepore complex is supported by multiple experiments. Purified PA63 forms non-selective cation channels in liposomes, with a rate constant that is dependent on the pH of the compartment to which PA63 was added [100].

Furthermore, PA63 can permeabilize the plasma membrane and translocate LF and EF across the plasma membrane if the extracellular pH is lowered, with an optimum at pH 5.5 [101]. Less is known about the mechanism by which PA63 senses pH. Crystal structures of the prepore complex suggested that residues around the 2 β 2-2 β 3 loop that forms the transmembrane pore are responsible [102], however mutation of these residues changed but did not abolish pH-sensitivity [103]. A recent cryo-EM structure of the mature pore suggests movement of the 2 β 10-2 β 11 loop is the trigger for pore formation and proposes the 2 β 10-2 β 11 loop as the pH-sensor [104]. Interestingly, this loop contains no histidine residues, but aspartate residues. These aspartate residues in this loop experience a drastically different environment in prepore and pore structure (Figure 1-5b). While the intrinsic pKa of aspartate is not close to the relevant pH-range, equation 8 demonstrates that coupling of the protonation event to a conformational transition can change the effective pKa.

The proton gradient between the endosome compartment and the cytosol is also the driving force for translocation of EF and LF across the pore [105]. While proton cotransporters for metabolites normally use an alternating access mechanism, where simultaneous binding of protons and metabolite is required for isomerization of the transporter, the mechanism seems to be different for the anthrax toxin. The current model is a “brownian-ratchet”, where spontaneous movement of EF and LF are stabilized by deprotonation of acidic residues in the cytosol. The narrowest part of the pore is highly hydrophobic and therefore probably impermeable to deprotonated aspartate and glutamate [104]. Since protonation of these residues is more likely in the acidic endosome compartment this creates a driving force for translocation.

Nitrophorin 4

Nitrophorin 4 is found in the secretory glands of the kissing bug *Rhodnius prolixus* and is responsible for the delivery of nitrous oxide into the bitten tissue of a victim to initiate vasodilation. Selective uptake of NO in the salivary glands and release in the victim is supported by the difference in the pH, which is 5.5 in salivary glands and 7.5 in the victim’s tissue [106]. While the bond strength between the heme group of nitrophorin 4 and NO is pH dependent this change in NO affinity is also caused by a conformational change from a closed state at low pH to an open state at neutral pH. While this mechanism is similar to hemoglobin this conformational state is not influenced by the concentration of NO, providing no orthologous way to modify the equilibrium between open and closed state. So while crystal structures of the open and low state are available

[107], [108], only the effective macroscopic pKa of titratable sites could be determined by observing shifts in the pH-dependence after site-directed mutagenesis [109], with no means to study the difference between pKa between open and closed state experimentally. However, recent advances in computational prediction of pKa-values, described in more detail in Section 1.3.4.2, allow prediction of pKa values based on structures, which are in this case are available for both states (Figure 1-5C).

Prediction of pKa values in the open and closed state of all titratable side-chains showed that the pKa of Asp₃₀ was 8.6 in the closed conformation and 5.2 in the open conformation [110]. Calculation of the pH-dependent free energy difference between the closed and open state using replica-exchange methods predicted a macroscopic pKa of 6.5 for Asp₃₀, which is in agreement with experimental results [111]. Interestingly, even though the pH of the structural transition is closer to the intrinsic pKa of histidine, an acidic residue functions as the pH-sensor. However, in order to shift the effective pKa of an acidic residue into the neutral range, the microscopic pKa in the low pH structure must be strongly shifted to alkaline values. This requires the presence of strong tertiary structure at low pH and therefore is not possible for systems that lose structure at acidic pH, such as the propeptides of acid-activated proteases.

1.2.4 Approaches to study pH sensing proteins

The small size of a proton and the rapid kinetics of protonation and deprotonation presents a unique challenge to gaining experimental insight into pH sensing mechanisms. In this section, I will review the most prevalent

methods used to investigate pH sensing in proteins and describe the hydrogen-deuterium exchange method for histidine pKa determination that I employed in chapter 4 of this thesis.

Titration of proteins

The most basic method to study pH sensing in proteins is measurement of their structural or functional characteristics as a function of pH. Most often the first indication of the function of a protein as a pH sensor, is if these characteristics can affect biological function and happen at physiological pH. In each case the nature of these experiments is different based on the biology of the protein in question. Examples include the measurement of (i) hemoglobin oxygen affinity at different pH values [92], (ii) anthrax toxin translocation at different pH values [112], (iii) nitric oxide dissociation from nitrophorin 4 as a function of pH [106], and (iv) furin activation as a function of pH [18].

More direct information about the relationship between protonation and function can be gained by determining the structure of proteins at different pH values. In cases where the structure of both the acidic and the basic state is well defined, X-ray crystallography can be used to determine the structure of the two different states. This has been achieved for both hemoglobin and nitrophorin 4, which is one of the major reasons why their pH regulation is so well understood. An exciting development is the emergence of cryo-electron microscopy (cryo-EM) as a method to determine structure at atomic-level resolution. Since pH heavily influences crystal formation it may at times be impossible to obtain crystal at a specific pH. However, cryo-EM does not require crystallization of

protein. The power of this approach is demonstrated by the low-pH triggered cryo-EM structure of the mature anthrax protective antigen pore structure [104], which had been impossible to crystallize.

An alternative to high-resolution structural methods is the use of different forms of UV-Vis spectroscopy to analyze the structures of a protein as a function of pH. In some proteins, where the functional groups have characteristic absorbance one can just measure changes in the extinction coefficient or the absorption maxima as a function of pH. For example, the binding of nitric oxide to nitrophorin 4 can be monitored by changes in the absorption maximum of the heme group [106], while the absorption of the retinal ligand of rhodopsin has been used to study the pH dependence of photo-intermediates [113]. In cases where there are no specific absorbing groups the quantum yield and wavelength-maxima of the intrinsic fluorescence of tryptophan, phenylalanine, and tyrosine residues can be exploited as their fluorescence parameters depend on their local environment. Hence, by observing the change in fluorescence properties as a function of pH one can gather information about pH dependent structural changes. This has been successfully employed to measure differences in pH-dependence of hemoglobins from different species [114]. However, one must take care to exclude effects that are due to titration of the fluorophore and not due to changes in protein structure [115]. Another label-free alternative is the use of circular dichroism spectroscopy, which measures differences in absorption between left- and right-circularly polarized light, due to the asymmetrical local environment of absorbing groups. For example the peptide bond absorbs

between 190 and 240 nm and circular dichroism in this range (Far-UV CD) can give information about the secondary structure of a protein. A major advantage of UV-Vis methods is that the ease of these measurements allows titration of protein in small pH-increments and by using small pH-electrodes and automated pumps one can readily titrate the protein directly in a cuvette within a spectrometer. The use of UV-Vis spectroscopy is therefore the easiest way to obtain the apparent extent of conversion, γ_B , as a function of pH.

Raman and infrared spectroscopy

Spectroscopic methods can also be used to obtain direct information about the protonation status of residues. Protonation of a residue changes the available vibration modes, which can be detected either by infrared spectroscopy or by Raman spectroscopy. Infrared spectroscopy measures absorption of photons with an energy equal to the energy differences of various vibration modes [116]. Raman spectroscopy measures slight changes in the wavelength of scattered photons with higher energy than the vibration modes [117]. Besides differences in the underlying measurement technology, they differ in the nature of vibrational modes that can be detected. The practical hurdle that one has to overcome in both techniques is to isolate the spectral features of the particular residue whose protonation status one wants to analyze from the plethora of other vibration states in a protein molecule.

Infrared spectroscopy has been successfully used to determine protonation status of aspartate and glutamate residues in bacteriorhodopsin [118]. Raman spectroscopy has been successfully used to assess the protonation

state of histidines [119]. In praxis Raman UV-spectra of proteins are often dominated by the vibration modes of the aromatic residues tyrosine and tryptophan. In order to isolate the vibration feature of a residue of interest one can use a difference spectrum of the native protein and the protein where the residue of interest has been mutated. Another approach is to use isotopic labeling. By performing the Raman experiment in deuterated buffer, the exchange of protons bound to the nitrogen atoms of the imidazole ring cause the emergence of a unique nitrogen-deuterium vibration mode that can be used to assess histidine protonation [120]. More recently, deuteration of the C2 atom in the imidazole sidechain has been used to specifically measure the protonation status of histidine in proteins [121].

NMR

Nuclear magnetic resonance (NMR) spectroscopy is probably the most powerful tool to study pH sensing, because it can provide simultaneous information about the titration state of specific residues and protein structure. Even though NMR allows direct detection of individual protons, in most cases NMR is not used to directly observe the proton bound to titratable side-chains, because these protons readily exchange with the solvent and solvent protons must be suppressed in most NMR experiments. Instead NMR experiments rely on observing shifts of peaks derived from the titratable side-chains, while the protein is being titrated. These shifts as a function of pH can then be fitted by a Henderson-Hasselbalch equation to determine the pKa of a residue [122]. One caveat of these experiments is that conformational changes due to changes in pH can also lead to

shifts in these peaks, which may obscure the measurement of true pKa values. Furthermore, it requires that peaks are observable in the complete titration range, which can be problematic if the conformational changes induce loss of structure, which leads to peak overlap, due to similar chemical environments and to peak broadening, because the chemical groups experience dynamic changes in their chemical environment. Additionally, NMR experiments require large amounts of purified, isotope labeled protein, which can be difficult to obtain and are limited by the size of proteins.

New approaches have been developed that allow direct measurement of the equilibrium between protonated and unprotonated forms of a histidine side-chain at a specific pH [123]. This is a very exciting development for proteins that lose structure as a function of pH, such as the propeptides of proprotein convertases. It also allows the determination of microscopic pKa values, instead of the macroscopic pKa value (see Section 1.3.2.1), which allows deeper insight into the thermodynamics of pH sensing. However, these approaches still require large amounts of isotope labeled protein and are in some cases very sensitive to systematic error caused by incomplete labeling. Furthermore these experiments have specific requirements for protein size, because of interference with protein tumbling [123].

Computational prediction

Since determination of pKa values of individual residues is not straightforward, the rapidly increasing database of protein structures makes predicting pKa values from structural information a very intriguing prospect. There are three

principal methods to predict the pKa value of a titratable side-chain: (i) Solving of the Poisson-Boltzmann equation, (ii) molecular dynamics based methods, and (iii) empirical predictions.

The Poisson-Boltzmann equation (PBE) describes the electric potential caused by a collection of point charges in an ion-screened solvent [124]. pKa prediction approaches based on the PBE assume that major factor in perturbing the acid-dissociation constant of a titratable site within a protein is the difference in electric potential compared to the solvent [125]. Since the PBE for an irregular object like a protein cannot be solved analytically, these approaches divide the conformational space into finite grids and solve the PBE at each grid point numerically, dependent on a set of point charges derived from high-resolution structures for atom coordinates and force-field parameters for partial charges. Besides high dependence on accurate force field parameters these approaches suffer from the fact that they do not intrinsically take into account the conformational rearrangements upon protonation and the fact that the dielectric constant of the protein interior is poorly defined.

While molecular dynamics normally cannot simulate chemical reactions (see Section 1.4.2.1), an approach called constant pH molecular dynamics (cpHMD) makes an exception for the protonation [126], [127]. cpHMD introduces a new coordinate, λ , for each titratable site that describes whether the site is protonated and unprotonated. The potential function for this coordinate is derived from pH-dependent parameters optimized to reproduce pKa values in model components. Since the simulation will attempt to move down a potential

gradient it will populate the titration state that makes more favorable interactions with the rest of the protein. By performing the simulation at different pH values, i.e. different parameters for the lambda coordinate, one can computationally titrate the protein and then directly observe the equilibrium between protonated and unprotonated forms during the simulation. While these methods explicitly take conformational changes into account, it is hard to gauge for how long simulations need to be performed in order to fully explore the conformational space for each pH. Furthermore, these methods mostly rely on implicit solvent approaches to model the solvent, thereby potentially missing some physical solvent effects.

Empirical methods describe the protein environment of a titratable site by a variety of measures, such as number of hydrogen bonding partners, hydrophobicity of neighboring residues, or distance from other charged sites. From these measures a scoring function is derived, that is optimized using a set of experimentally determined pKa values [128]. Empirical methods offer easy use and are computationally very efficient, while still performing as well as more sophisticated PBE approaches. However, empirical methods often fail in cases that are not well represented by the test set they were parameterized against. Recent benchmarking of various pKa prediction methods demonstrated that none of these methods offer perfect predictions and errors in predictions can easily be more than one pH unit [129]. However, they are potentially robust enough to identify side-chains whose pKa values are strongly distorted. Furthermore, since these prediction methods assume a rigid structure or even in

the case of cpHMD can only model minor structural changes, such approaches can provide microscopic pKa values. This can be very useful to predict pH sensors when structures of two different functional states are known, as shown in the case of nitrophorin 4.

Hydrogen-deuterium exchange

Even protons in groups with extreme pKa values will occasionally dissociate and then be replaced by protons from the solvent environment. While these groups will only spend a small amount of time in an unprotonated state, the rate at which this occurs can be measured by replacing the protons of the solvent with a different isotope, most often deuterium. Since the most common mechanism of protons abstraction is catalyzed by a hydroxonium ion, the rate constants of the exchange from hydrogen to deuterium are strongly pH dependent with the slowest rates at acidic pH (~2.6). The rate of exchange also depends on the solvent accessibility of the particular group and the free energy barrier of proton dissociation. This energy barrier is small if hydrogen is bound to a highly electronegative atom. As a consequence, hydrogen bound to nitrogen, oxygen, and sulfur can rapidly exchange within the sub-millisecond timescale. An exception is the hydrogen of the peptide-bond amide group. At neutral pH these hydrogens exchange with a half-life on the order of seconds. However, this exchange rate can be strongly reduced, either by protection of the amide from solvent or the participation of the amide in hydrogen bonds within the secondary structure of proteins. This is often exploited to gain information about protein structure [130].

Protons bound to carbon, however, due to the low electronegativity of carbon, essentially do not exchange with the solvent. An exception is the hydrogen bound to the C2 carbon of the imidazole ring. In the protonated form of the imidazole ring the deprotonated C2 is stabilized, thereby lowering the energy barrier of hydrogen exchange (Figure 1-6 A)[131], [132]. This leads to an exchange rate with a half-life of several days, which although orders of magnitude slower than backbone amides can be measured experimentally.

Since the exchange depends on the fraction of the histidine side-chain in the protonated form and on the concentration of the hydroxonium ion [131], [132], the rate constant for exchange k_{ex} can be written as:

$$k_{ex} = k_2 * [OH^-] * \frac{[HisH^+]}{[HisH^+] + [His]} \quad (11)$$

where k_2 is the second-order rate constant of the rate-limiting proton abstraction.

This formula can be rewritten as:

$$k_{ex} = \frac{k_2 * [OH^-]}{1 + \frac{[His]}{[HisH^+]}} \quad (12)$$

By multiplying with $\frac{[H^+]}{[H^+]}$ and substituting in the dissociation constant of water $K_W = [H^+][OH^-]$ and the equilibrium constant of the histidine acid-base reaction $K_a = \frac{[His][H^+]}{[HisH^+]}$ we can write the exchange rate as a function of proton concentration:

$$k_{ex} = \frac{k_2 * K_W}{[H^+] + K_a} \quad (13)$$

Therefore, the exchange rate approaches 0 under acidic conditions, where the proton concentration is very high and has a maximum value $k_{max} = \frac{k_2 K_W}{K_a}$ at basic conditions where $K_a \gg [H^+]$. By introducing k_{max} this simple Hendersson-Hasselbalch equation can be derived:

$$\log\left(\frac{k_{max} - k_{ex}}{k_{ex}}\right) = pK_a - pH \quad (14)$$

By determining the exchange rate of the C2-proton as a function of pH one can derive the pKa and k_{max} of a histidine side-chain. Figure 1-6 B shows an example titration curve of k_{ex} and how pKa and k_{max} can be derived. One underlying assumption is that k_2 is independent of pH, which is only true if the solvent accessibility of the histidine side-chain is independent of pH, which in case of pH dependent conformational changes is uncertain. However, this weakness applies similarly to other techniques such as NMR titration, where changes in peaks can be caused by protonation and/or conformational changes.

There are multiple ways to measure the exchange rate of the C2-proton. Early studies used tritium, instead of deuterium, as the exchange label. This allows quantification of uptake by scintillation counting. However, first the protein must be digested into peptides, which have to be separated by chromatography and then be identified. While uptake can also be quantified using NMR this requires high concentration of protein and assignment of peaks, which can be a time-intensive process. Recent progress in mass spectrometry has made it fairly straightforward to identify peptide fragments and high resolution instruments also allow quantification of deuterium uptake with high precision [133]. The major advantage of using mass spectrometry is that it requires very little protein and can be applied to proteins with large molecular weight, which need not be in a highly purified form. Moreover, since information about the pKa is essentially encoded chemically in deuterium uptake it is not necessary for the

protein to remain stable in the whole pH range, as it would be if protonation would be measured directly by NMR.

1.3 Computational approaches used in this thesis

Computational algorithms that exploit the large amount of biological data stored in databases coupled with our physical understanding of proteins are now enabling us to analyze biological phenomena that are difficult to study using experimental methods. Most importantly computational approaches can lead to novel hypotheses, which can then be tested by experimental methods. To illustrate this, Chapter 5 of this thesis will describe multiple studies were I collaborated with several laboratories to overcome roadblocks in experimental approaches using computational techniques. In the following sections, I will describe the algorithms used to investigate the pH-sensing mechanism of proprotein convertases and to test or propose hypotheses in collaborative projects.

1.3.1 Sequence based approaches

Homology detection and sequence databases

Theoretically all the information for a proteins function is contained in its sequence [134]. Since protein sequences can be easily described by a computer as a string of symbols that represent each amino acid and given that genomic sequencing together with efficient prediction of open reading frames makes protein sequences almost trivial to obtain, it is highly desirable to be able to predict properties of a protein from its sequence. While it is possible to predict

certain properties, such as the mass or the secondary structure [135], using only the sequence information, the prediction of more complicated properties, such as enzyme mechanisms or ligand binding, requires knowledge of the tertiary structure of a protein. Although advances in protein structure prediction have made this idea more feasible (see Section 1.3.2.3), an easier approach to obtain information from the sequence is to compare it to other homologous proteins.

We think of two sequences as homologous if they share a common ancestor. Since the evolutionary history of a sequence is not directly known the most common way to detect homology is to compare their sequences. If two sequences are more similar than we would expect by chance we infer that they are homologous[136]. It is important to note that due to the astronomical number of theoretically possible sequences and the fact that homologous sequences arose by natural mutation, while having to preserve function, it is very unlikely to make false positive assignments (to assume two sequences are homologous even though they are not). However, false negatives (two sequences are homologous even though we can detect no sequence similarity) are probably fairly common, given that proteins with similar structure and function often share no detectable sequence similarity (as in the case of subtilase propeptides) [137].

There are multiple algorithms that can be used to infer similarity of two sequences. Since during evolution insertions and deletion are created, comparison of two sequences requires their alignment, which can be performed using the Smith-Waterman [138] or Needleman-Wunsch algorithm [139] to obtain local or global alignments, respectively . However, it is impractical to use

these algorithms to calculate similarity of a query sequence to all sequences in a database. More efficient heuristic algorithms, such as BLAST or FASTA [140], [141], have been developed to enable routine search of single sequences against large databases. Crucial to these algorithms is the calculation of an expectation value that gives an estimate for the probability that a detected similarity occurs purely by chance. This expectation value is critical to avoid false positive errors during a database search.

A well maintained database of protein sequences is as important for homology detection as is an efficient search algorithm. The enormous increase in available sequence data has necessitated the creation of a robust database system to make these sequences rapidly and easily accessible. All published nucleotide sequences are deposited into primary DNA sequence databases, maintained by the International Nucleotide Sequence Database Collaboration (INSDC), consisting of the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ) [142]. Based on automatic annotation of coding regions protein sequence databases are created from the deposited DNA sequences. In particular the UniProt project maintains two protein databases: the first, TrEMBL, consists of all available non-redundant protein sequences; the second, SwissProt, is created by manual annotation and review of these sequences and is therefore able to provide rich annotation with domain regions, critical residues, post-translational modification, and links to other databases [143].

This rich annotation makes homology searches against the UniProt database particularly useful. Homology between two protein sequences implies that they evolved from a common ancestor and therefore share some degree of conservation in structure and function. Therefore, structural and functional annotations of homologous protein can probably be transferred to a query sequence with unknown function. Even more useful than aligning just two sequences is the alignment of several sequences in a multiple sequence alignment (MSA). In a MSA every row contains a sequence, while every column contains the residues of all sequences at a particular position. This allows the calculation of conservation scores for each position [144], which can be very helpful to understand protein function under the assumption that residues with important functions are more conserved than residues with no particular function. Furthermore, the coupling between residues at two different position (i.e. does the amino acid at one position influence which amino acid is found at another position) can be used to predict whether two residues are close in the tertiary structure and to find networks of functionally important residues [145], [146]. By employing statistical models of protein evolution one can use MSAs to calculate the most likely sequence of an evolutionary precursor to these sequences. This has proven to be a useful tool in understanding how proteins may have evolved [147]. Another use of MSAs is the calculation of sequence profiles, where the MSA is represented by the relative occurrence of each amino acid at every position in the MSA. Such sequence profiles can drastically increase the statistical power of a homology search, by reducing the false negative rate, and iterations

of BLAST searches that create MSA profiles is used by the PSI-BLAST algorithm to find distant homologs of a query sequence [141].

Hidden Markov models

While MSAs and the generation of statistical profiles from these alignments are powerful tools for homology detection, they have problems with the treatment of insertions and deletions. Since an insertion within one protein will introduce an additional column with a gap character in every other sequence, large alignments of diverse sequences tend to become very long, often an order of magnitude longer than the average sequence length. Therefore, a mathematical representation that can handle insertions and deletion more gracefully during the alignment process is needed.

Hidden Markov models (HMM) were developed by Leonard E. Baum in 1966 [148]. A Markov model describes a system as random transitions between discrete states, with probabilities that depend only on the previous state. Every state can emit different observables with probabilities that also only depend on that state and not on previously visited states. Cases where the underlying model is unknown and only observations of the model outcome are known are called hidden Markov models. Homologous proteins can be represented by a hidden Markov model (HMM) sequence profile, where the model represents the protein fold and requirements for amino acids at each position and individual protein sequences are emissions of this model. The most commonly used model is shown in Figure 1-7 A [149], where every position of the profile is represented by a match or a deletion state. Additionally insertion states are included between

match states. Match and insertion states can emit amino acids. A sequence can be represented as a path through this profile model as indicated in Figure 1-7 B. The Viterbi algorithm allows to calculate the path through the model that has the highest likelihood to produce this sequence [150]. Inversely, the Baum-Welch algorithm can be used to calculate the set of parameters for a model that have the highest likelihood to emit a set of sequences [151]. These algorithms can be used for homology detection in databases in three ways: (i) An initial BLAST search can be used to generate a preliminary MSA of the query sequence, which is then used to generate a HMM sequence profile using the Baum-Welch algorithm. Every sequence in the database is then tested by the Viterbi algorithm if it is an emission of this HMM-profile. (ii) Alternatively, HMM sequence profiles for every sequence in the database can be calculated beforehand. The query sequence is then tested against each of these models. (iii) The highest sensitivity is achieved by doing both, generating HMM sequence profiles for the search sequence and the database, as is done by the HHsearch server [152].

In addition to increasing the sensitivity of homology searches, HMM sequence profiles can also be used for automated domain annotations. For example, several databases maintain HMM profiles for domains that can be used to annotate domains found in database deposited sequences by searching each sequence for statistically significant local alignments against the HMM profile database. These databases differ in the ways the HMM profiles are generated. Pfam contains manually curated HMM profiles, which are generated using seed alignments that contains representatives for this domain that are chosen by a

human [153]. Superfamily [154] and Gene3D [155] are based on the structural classification of domains found in the PDB database as curated by the SCOP [156] and CATH [157] databases, respectively.

1.3.2 Structure based approaches

While sequences are very straightforward to represent computationally, structures are substantially more complex. Structures are most commonly represented by the Cartesian coordinates of every atom within an arbitrary reference frame. Many properties of a protein, such as the electrostatic potential [158], pore radii of ion channels [159], or presence of binding pockets [160] can be calculated from this description for any given protein. However, it is often more useful if the available conformational space of the protein can be explored to understand how its structure changes in response to stimuli. This can be used to predict binding of ligands or other proteins and to predict the structure of the protein from its sequence. In the following sections, I will initially illustrate two major classes of algorithms, Molecular Dynamics and Monte-Carlo, which are used to generate conformational ensembles and subsequently describe how these methods can be used to predict the structure of a protein from its sequence.

Molecular dynamics

In a molecular dynamics simulation every atom is characterized by its location in space and by its velocity. The simulation progresses by a fixed time step after which the coordinates evolve according to the velocities of the atoms and the velocities change according to the forces acting on each atom. The forces that act on each atom are approximated by fairly simple equations that are derived from

classical Newtonian physics, but are parameterized to model the quantum phenomena that govern the motions of atoms[161]. For example, the forces created by chemical bonds are simulated using a formula derived from Hooke's law, analogous to spheres connected by springs. The bond and dihedral angles are modeled using similar formulas. Nonbonded forces are divided into a van-der-Waals term, which heavily penalizes atoms that are closer than the sum of their van-der-Waals radii and mildly favors their proximity, and an electrostatic term that assumes a specific point charge for every atom [162]. Since the time step that can be used to advance the simulation is very short (typically 2fs) and the calculation of the nonbonded forces principally requires calculating a term for every possible pair of atoms, the simulation are computationally very expensive and limited in their timescale and number of simulated atoms. While approaches to alleviate these constraints, such as coarse-graining the atomic structure, representing the solvent by an intrinsic force-field term instead of individual solvent molecules, or steering the simulation towards the desired structure, can facilitate more extensive sampling, they all come at the cost of reduced accuracy [163].

The usefulness of molecular dynamics simulations has been most dramatically impacted by the rapid increase in computational power and the development of specialized hardware for molecular dynamics simulation [164]. Using these approaches the available timescales have been extended into the μs or even ms domain [165], which is long enough to simulate the binding of ligands [166] or folding of small proteins . Additionally, the maximal size of

simulated systems has been increased from traditionally small protein domains in a small solvent box to, most recently, a complete HIV virion [167].

Monte-Carlo sampling

Similar to molecular dynamics simulations, Monte-Carlo sampling is performed by iterations of structural changes. However, the changes are not calculated according to forces, but are chosen randomly. The random step is then either accepted or rejected based on an energy evaluation of the resulting conformation. The most commonly used criterion for accepting a random move is the Metropolis criterion, where the move is accepted if it results in a lower energy and in case of an energy increase, the move is randomly accepted based on the Boltzmann equation [168]. Compared with molecular dynamics, the Monte-Carlo algorithm provides multiple advantages such as (i) more available flexibility in the move step, (ii) introduction of larger perturbations instead of small local movements of atoms, (iii) representation of the protein by dihedral angles instead of Cartesian coordinates and (iv) the function that evaluates the energy of the system does not need to be differentiable. All these advantages allow more rapid sampling of the conformational space. The major disadvantages of Monte-Carlo sampling are that the resulting trajectories are not biophysically meaningful and that the solvent cannot be explicitly modeled, because the large movements of a protein would require re-equilibration of the solvent molecules. The fact that Monte-Carlo sampling allows any kind of random perturbation is not only an advantage, but also a disadvantage, because the kind of perturbations must be carefully chosen and tested.

One of the most common software packages for application of Monte-Carlo simulations is the ROSETTA package that was developed by David Baker's group [169]. ROSETTA combines implementation of the Monte-Carlo algorithm with many possible types of perturbation, energy functions with increasing degrees of accuracy, and an internal representation of the atomic structure that is optimized for rapid application of moves and energy evaluation. Most importantly, these modules have been combined into specialized protocols to facilitate many types of predictions, such as structure prediction [170], protein design [171] or protein-protein docking [172].

Structure prediction

Anfinsen's hypothesis that the sequence of a protein determines its structure [134], implies that it should be possible to predict the structure of a protein from its sequence. Ideally one could just start with the atomic model of the extended peptide chain and by applying our knowledge of atomic motion, at least in appropriate simplifications as in molecular dynamics simulations, one should be able to simulate the folding of the protein. Indeed recent advances in computational power have enabled the routine folding of small proteins in molecular dynamics simulations [173]. However, with commonly available computational power, this approach is still impractical for most large proteins.

The advances in computational structure predictions can best be judged by the Critical Assessment of Structure Prediction (CASP) competition wherein the attempts of different groups to predict the conformations of recently solved, but yet unpublished, structures are judged in an unbiased fashion [174]. Over the

last few years, the two most successful programs that try to predict using *ab initio* approaches are I-TASSER [175] and ROSETTA [170]. Both of these methods use a similar approach, wherein initial models are built by assembling the proteins from short fragments with structures sampled from the PDB. A Monte-Carlo algorithm is used to optimize the arrangements of these fragments into a structure, using a fairly coarse-grained energy function. The best scoring models from this approach are then subjected to further refinement using more detailed energy functions. Overall it is important to note that their accuracy is still limited to small proteins (<150 residues) and neither of these methods can predict correct structures consistently [176].

The most accurate way to predict structures is to rely on the observation that structures evolve slower than sequences. Analysis of the PDB shows that most proteins pairs that have more than 40% sequence identity have highly similar structures. Even pairs with sequence identity as low as 20% are often highly similar [177]. In case of even lower sequence identity it becomes harder to accurately assume that two sequences are homologous, but it is not uncommon for two proteins with no detectable sequence homology to show similar structures. Homology modeling was first demonstrated by Andrej Sali, who developed the program MODELLER [178]. Homology modeling with MODELLER is performed by first searching the PDB database for homologous proteins by using an algorithm like BLAST, or for higher sensitivity a HMM based algorithm like HHsearch [152]. If a suitable homolog is found an alignment is built between the sequence of interest and the template protein. This

alignment is then used to derive constraints between main chain atoms from the template structure that are then used to build a structure of the protein of interest. The main advantages of homology modeling are the high accuracy if a template with high sequence identity is available and the fact that the accuracy of the model can be estimated from the sequence identity [179].

1.4 Outline of thesis

The major theme of this thesis is how computational and experimental approaches can be combined to gain deeper insight into protein function. Most of this thesis focuses on how propeptides of proprotein convertases sense organellar pH to mediate activation. In the second chapter I will describe how bioinformatics analysis of protease sequences enabled me to support the hypothesis that there is a functional requirement for histidine residues in the propeptides of proprotein convertases, which most likely is due to the role of histidine as a titratable residue. This work also hinted towards at the usage of multiple titration sites and demonstrated that another protease family, the cysteine cathepsins, rely on a similar mechanism. In the third chapter I describe how experimental approaches, mostly pH dependent circular dichroism spectroscopy, combined with molecular dynamics simulations, showed that the propeptides of proprotein convertases are sufficient to mediate pH sensitivity and to elucidate how the His⁶⁹Leu mutation in the furin propeptides blocks activation. The fourth chapter describes how I employed a histidine hydrogen-deuterium exchange mass spectroscopy technique to measure the histidine pKa values in proprotein convertase propeptides. This work demonstrated that pH of

activation is fine-tuned by modulating the energy landscape of pH-dependent allostery, indicated by a pKa shift of the conserved pH-sensing histidine. The final chapter describes collaborative efforts to understand protein function in a variety of diseases, ranging from cancer to leishmaniasis, where I contributed computational approaches to test or create hypotheses.

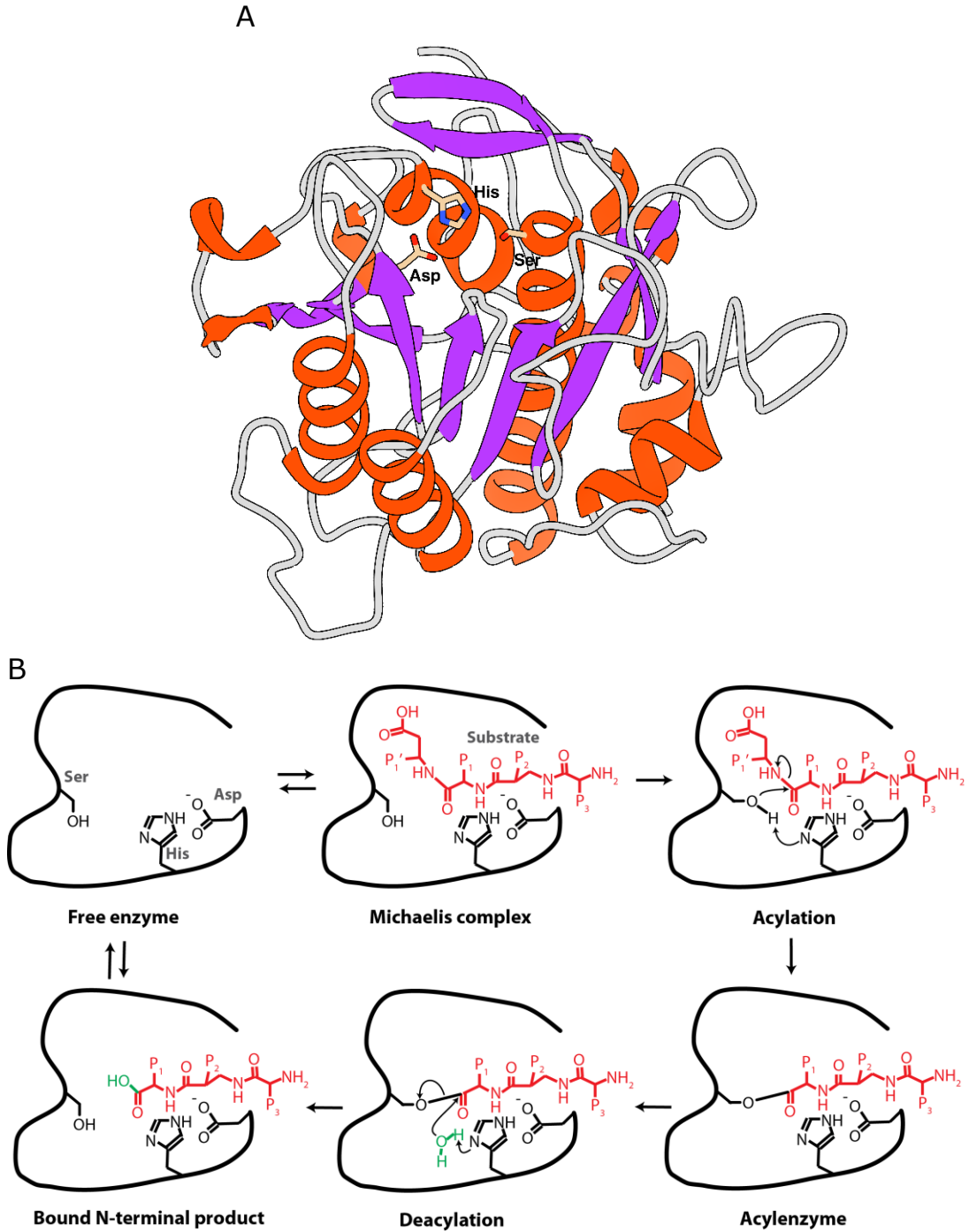
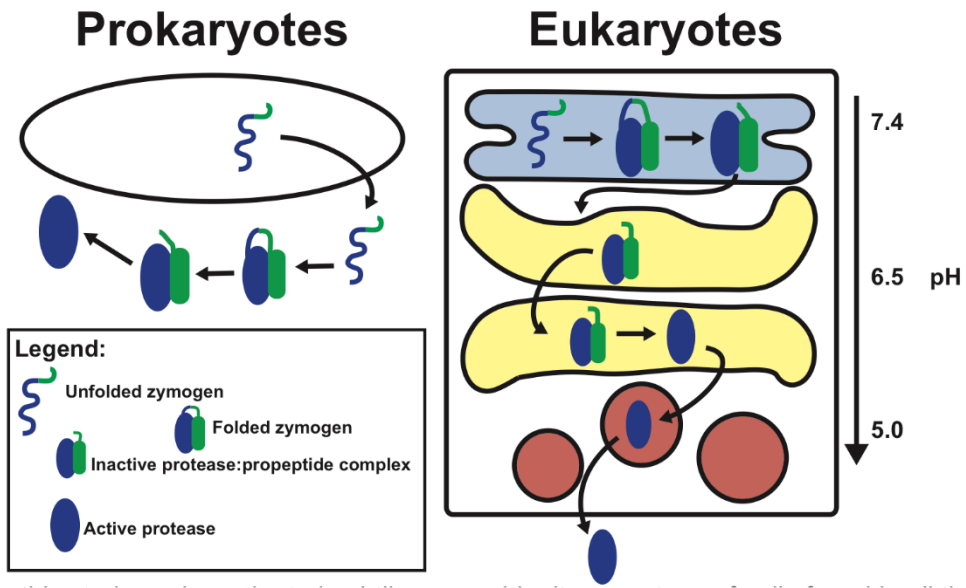


Figure 1-1: Structure and Mechanism of Subtilases.

(A) Ribbon representation of the subtilase fold. The residues of the catalytic triad are shown in stick representation (B) Proposed mechanism of subtilase catalytic cycle

A



B

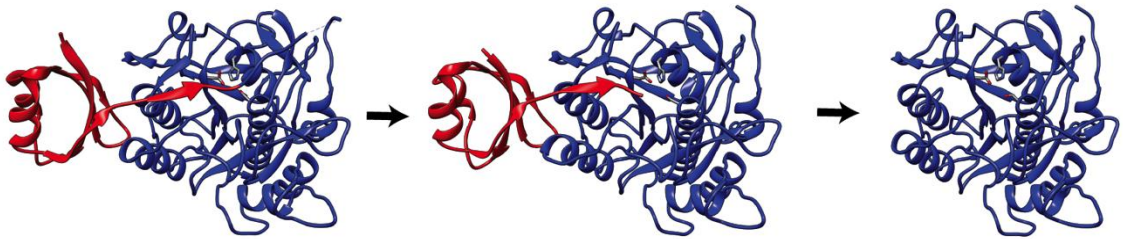


Figure 1-2: Mechanism of subtilase activation

(A) Schematic of the activation mechanism of subtilases in prokaryotes and eukaryotes (B) Structures of activation intermediates of Tk-Subtilisin

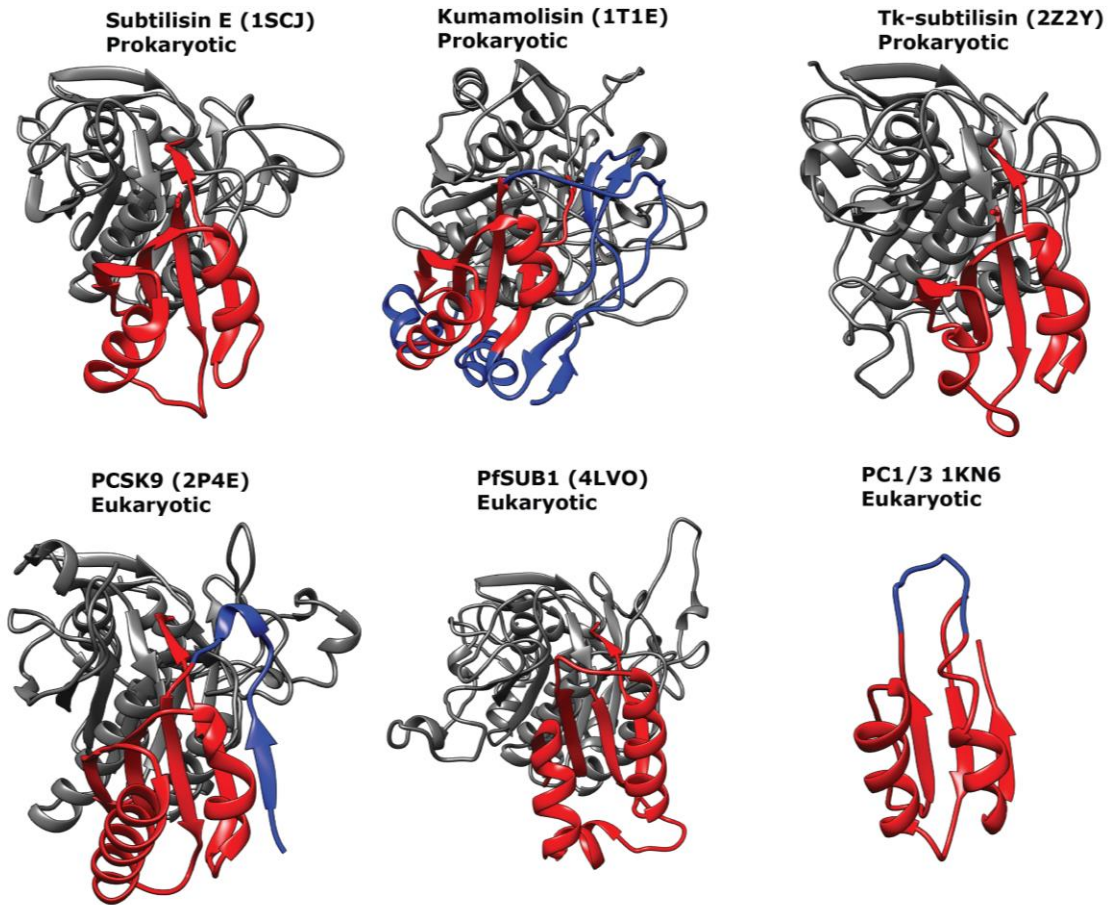


Figure 1-3: Structures of subtilase propeptides

Crystal structures of several propeptide:subtilase complexes and NMR solution structure of the PC1/3 propeptide. Proteases are shown as grey cartoon, while the basic fold of the propeptide is shown as red cartoon. Protein specific insertion are shown in blue. All propeptides share the same basic fold, with an anti-parallel beta-sheet and two alpha-helices. In all cases the C-terminus is located in the active site with the helix-distal site of the beta-sheet contacting the protease.

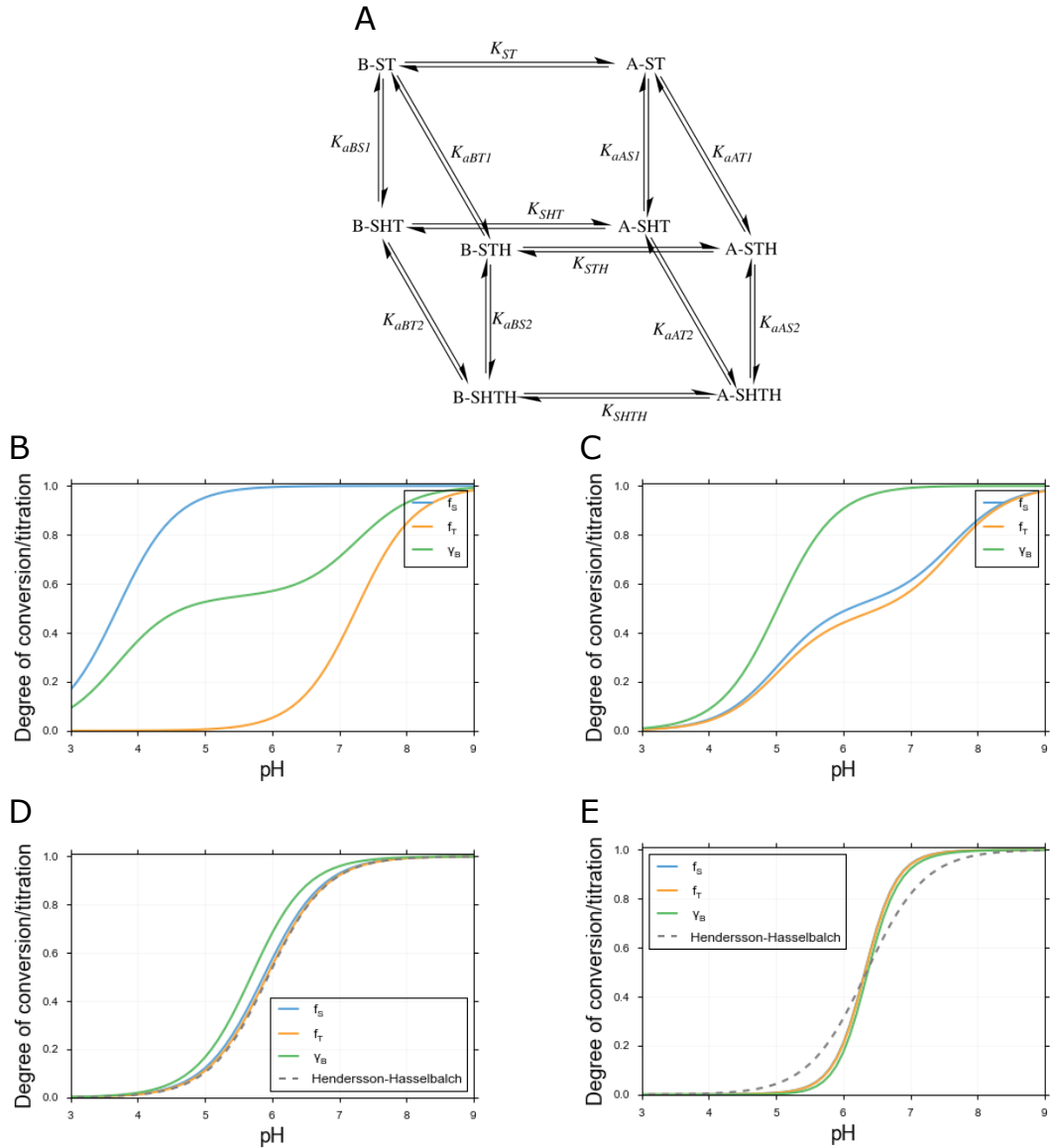


Figure 1-4: Thermodynamic model of two-site titration allosteric regulation and example titration curves.

Thermodynamic model of two-site allosteric equilibrium between state A and B (B) Theoretical titration curves where $pK_{aS} = 4$, $pK_{aT} = 8$, $K_{ST} = 0.1$, $K_{SHT} = 1$, $K_{STH} = 0.2$, and $K_{SHTH} = 1000$, providing an example for independent titration of S and T (C) Theoretical titration curves where $pK_{aS} = 7.3$, $pK_{aT} = 7.35$, $K_{ST} = 0.01$, $K_{SHT} = 0.0101$, $K_{STH} = 0.0101$, and $K_{SHTH} = 1000$, providing an example for negative cooperativity between S and T (D) Theoretical titration curves where $pK_{aS} = 7.3$, $pK_{aT} = 7.35$, $K_{ST} = 0.08$, $K_{SHT} = 0.25$, $K_{STH} = 0.25$, and $K_{SHTH} = 1.25$, providing an example for a behavior where titration does not completely convert B to A. A dashed line shows a Henderson-Hasselbalch titration curve as expected for a single site. (E) Theoretical titration curves where $pK_{aS} = 7.3$, $pK_{aT} = 7.35$, $K_{ST} = 0.01$, $K_{SHT} = 500$, $K_{STH} = 500$, and $K_{SHTH} = 1000$, providing an example for positive cooperativity between S and T. A dashed line shows a Henderson-Hasselbalch titration curve as expected for a single site.

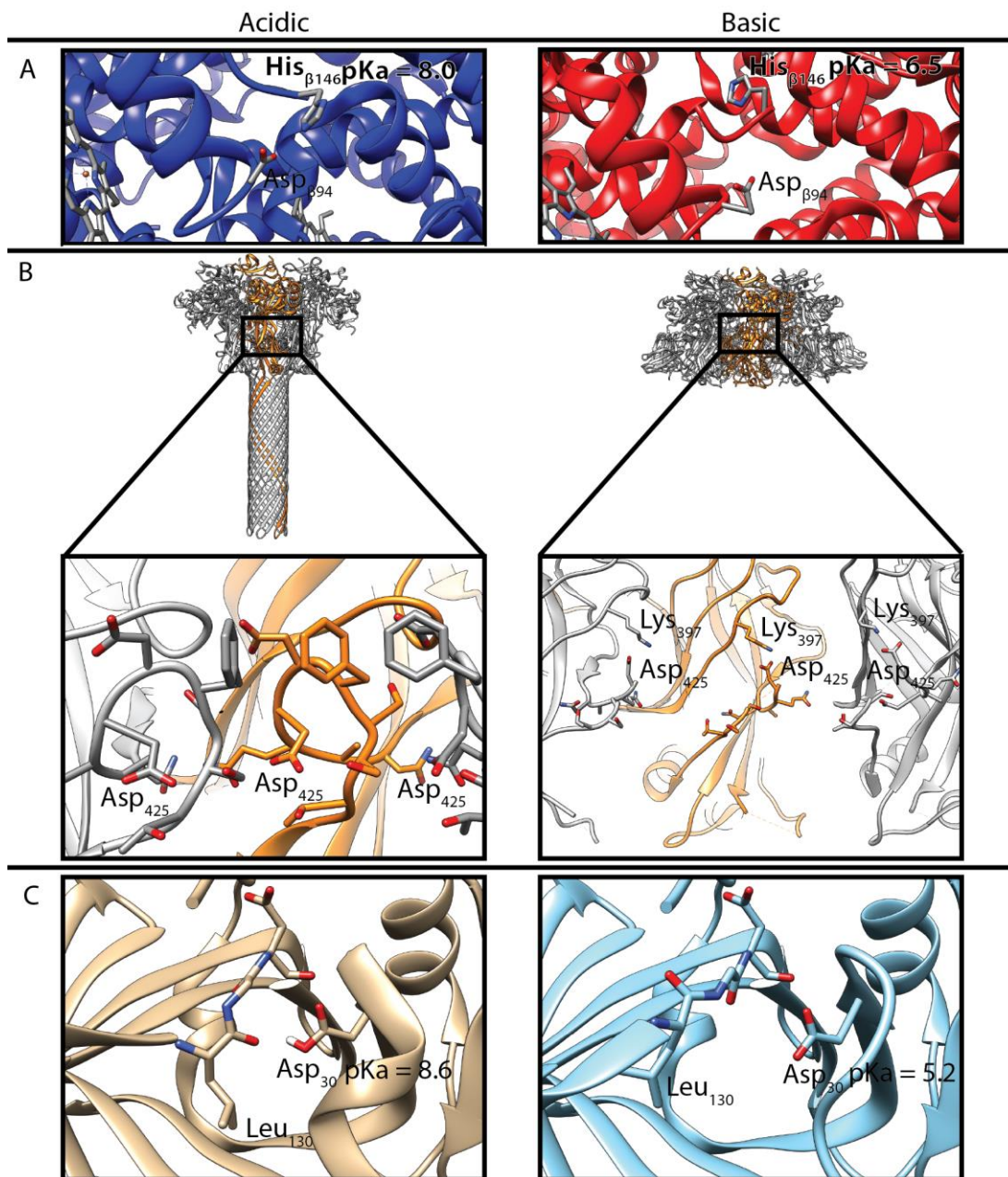


Figure 1-5: Structures of low and high pH states of different proteins.

(A) Crystal structures of deoxyhemoglobin and oxyhemoglobin. The close proximity of His₁₄₆ and Asp₉₄ in deoxyhemoglobin shifts the pKa to 8.0. Therefore protonation of His₁₄₆ shifts the equilibrium towards deoxyhemoglobin.

(B) Cryo-EM structure of mature anthrax protective antigen and crystal structure of the prepore from of anthrax protective antigen. Boxes indicate the different environments of Asp₄₂₅ in the two structural states.

(C) Crystal structures of closed and open form of Nitrophorin 4. In the closed form the pKa of Asp₃₀ is predicted to be 8.6, presumably due to hydrogen-bonding with the peptide backbone around Leu₁₃₀

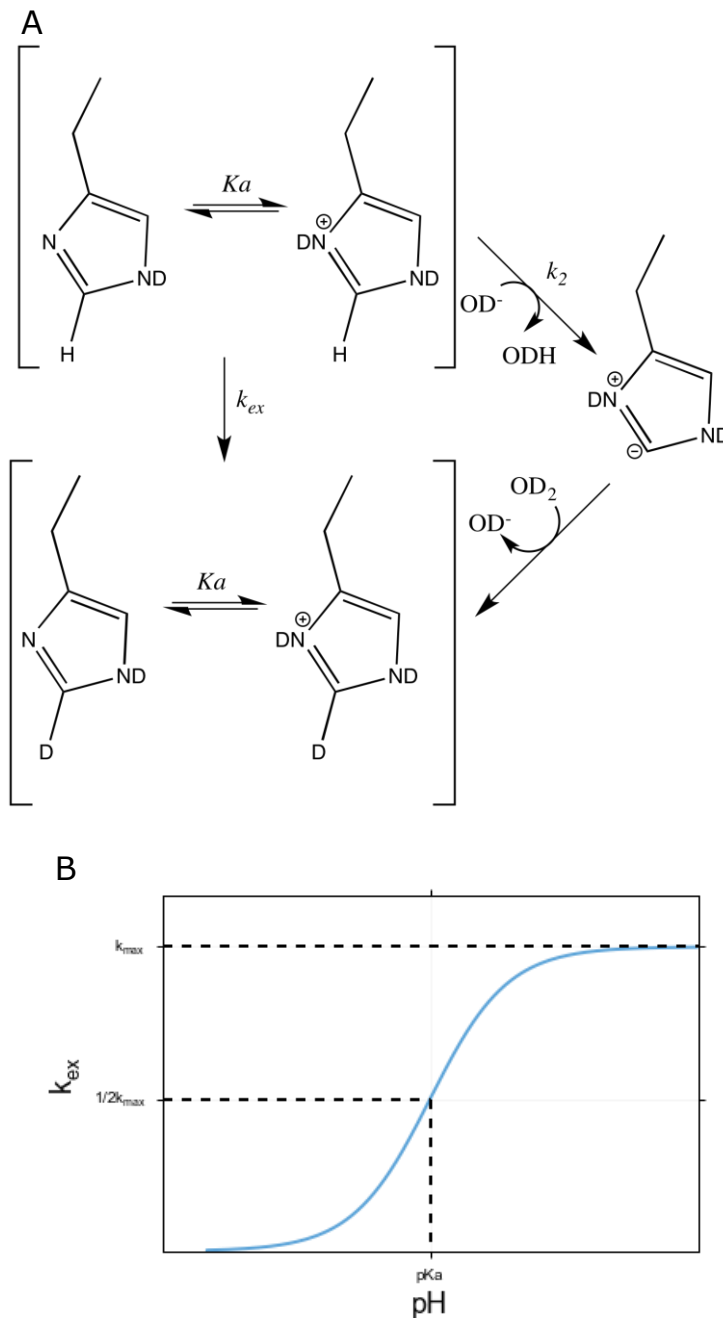


Figure 1-6: Mechanism of histidine hydrogen-deuterium exchange

(A) Mechanism of HD-exchange in the imidazole side-chain. The effective exchange rate k_{ex} contains information about the equilibrium K_a because hydrogen abstraction can only occur in the protonated state with the rate-limiting step being base-catalyzed (k_2).

(B) Example plot of measured k_{ex} as a function of pH. The pH of half-maximal exchange rate corresponds to the pK_a of the side-chain.

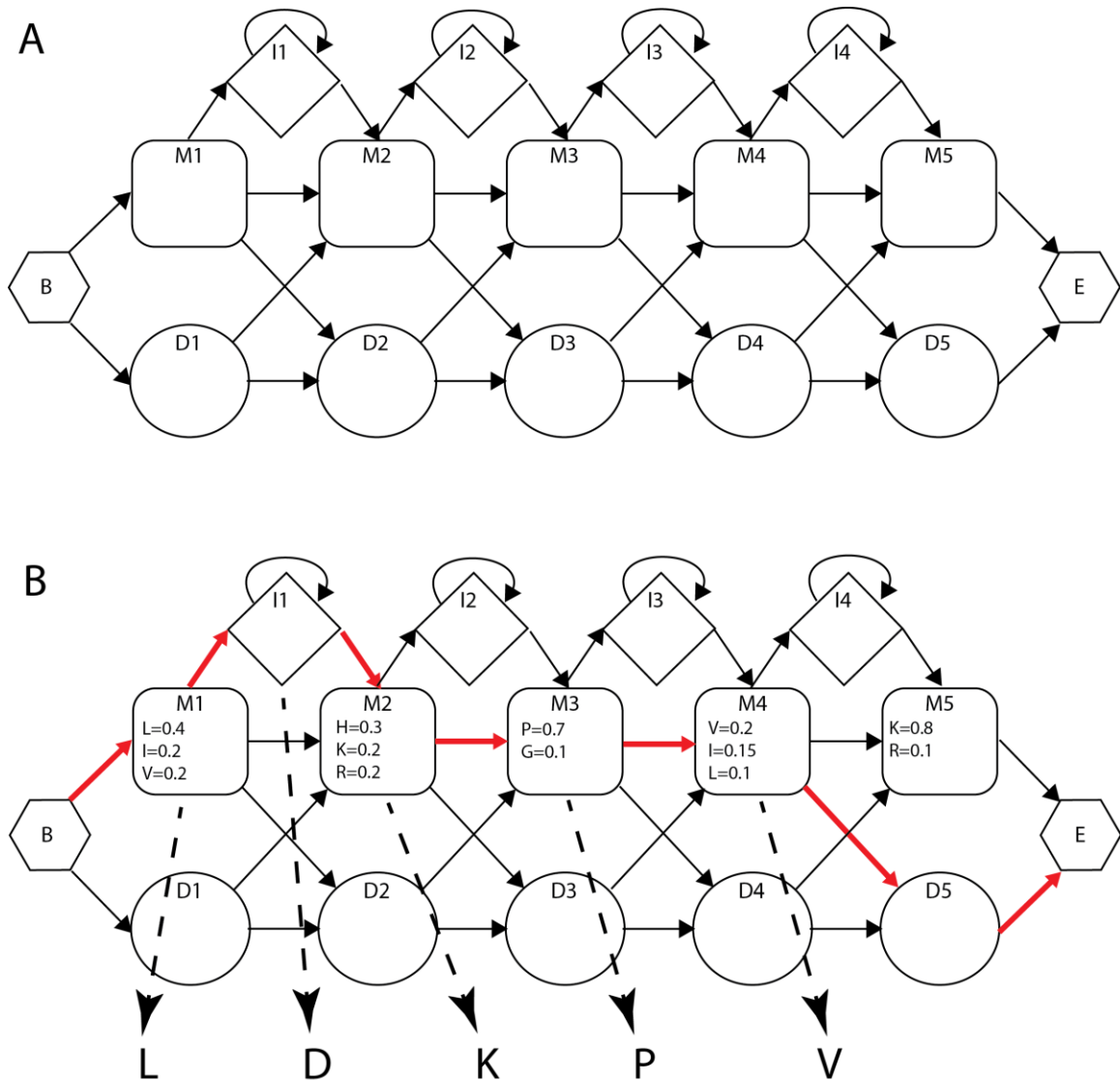


Figure 1-7: Examples for Hidden Markov model sequence profiles

(A) Example how HMM sequence profiles are constructed. Every conserved position is represented by a match state (M) that emits amino acids with site-specific probabilities. Sequences can have deletions at these positions, because the model can transit through deletion states (D) that do not emit amino acids. Insertions are introduced by insertions states (I). The probabilities of insertion and deletions are site specific and modeled by the probability of each match state to transit to the D or I state.

(B) Example of how the sequence LDKPV can be emitted by a path through an example HMM sequence profile. The path through the HMM is shown in red arrows. The Viterbi algorithm produces the most probable path through an HMM profile given the profile parameters and a sequence.

Propeptides of eukaryotic proteases encode histidines to exploit organelle pH for regulation.

Johannes Elferich, Danielle M. Williamson, Bala Krishnamoorthy and Ujwal Shinde

This manuscript was published by the FASEB journal on April 12th 2013 in volume 27 number 8 pages 2939-2945.

2.1 Abstract

Eukaryotic cells maintain strict control over protein secretion, in part by utilizing the pH-gradient maintained within their secretory pathway. How eukaryotic proteins evolved from prokaryotic orthologs to exploit the pH-gradient for biological function remains a fundamental question in cell biology. Our laboratory previously demonstrated that protein domains located within precursor proteins, propeptides, encode histidine-driven pH-sensors to regulate organelle-specific activation of the eukaryotic proteases furin and proprotein convertase-1/3. Similar findings have been reported in other unrelated protease families. By analyzing over 10,000 unique proteases within evolutionarily unrelated families, we show that eukaryotic propeptides are enriched in histidines when compared to prokaryotic orthologs. On this basis, we hypothesize that eukaryotic proteins evolved to enrich histidines within their propeptides to exploit the tightly controlled pH-gradient of the secretory pathway, thereby regulating activation within specific organelles. Enrichment of histidines in propeptides may therefore be used to predict the presence of pH-sensors in other proteases or even protease substrates.

2.2 Introduction

Eukaryotes are descendants of distinct prokaryotic cells that united symbiotically and evolved complex cellular compartments called organelles [180]. Secretory and endocytic organelles maintain a precise pH-gradient from the endoplasmic reticulum (ER; pH~7.2) to secretory vesicles (pH~5.5) and provide unique environmental conditions essential for optimal protein structure and function

[181]. How eukaryotic proteins differ from prokaryotic orthologs, to be able to exploit the pH-gradient for biological function, is a fundamental question in cell biology, and represents a major challenge to our understanding of protein trafficking, protein evolution, and organelle complexity. Comparing secreted eukaryotic proteases with bacterial orthologs will help to decipher the theoretical underpinnings that enable proteins to exploit the pH-gradient within the secretory pathway, and provide general principles for the relationship between structure, dynamics and function of biomolecules.

Proteases hydrolyze peptide bonds, and likely arose during evolution as simple catabolic enzymes to generate amino acid nutrients for primitive organisms [182]. Their ubiquitous distribution and presence of orthologs in prokaryotes, eukaryotes, and archaea, makes proteases ideal models for analyzing selective pressures that drove adaptation of eukaryotic proteins to complex organelle trafficking. Since uncontrolled proteolysis can have catastrophic consequences, cells have evolved at least two distinct mechanisms to maintain tight spatiotemporal control of protease activity. The first involves co-evolution of specific endogenous inhibitors, typically within compartments distinct from those containing active enzymes. The second mechanism involves proteases being synthesized as inactive precursors called zymogens, which become active by limited intra- or intermolecular proteolysis. In some cases the two regulatory mechanisms are combined; N-terminal propeptides co-evolved to chaperone folding of cognate catalytic-domains and act as potent temporary inhibitors after cleavage from the catalytic-domain [2].

Here we hypothesize that one of the ways proteases adapted to eukaryotic organelle systems was by encoding histidine-based pH-sensors in their N-terminal propeptides. When encoded at suitable positions within domains, the unique pKa of histidine side-chains (~6.5) can alter electrostatic interactions and modify conformations through changes in pH within the secretory pathway. These changes can then either affect the inhibitory potential of propeptides, or can increase their susceptibility to proteolysis.

2.3 Material and Methods

2.3.1 Conservation Analysis

Analysis of conserved residues was performed using ConSurf with standard settings [144]. The crystal structure of the propeptide:subtilisin E complex (PDB: 1SCJ) was used as input for analyzing bacterial subtilases, while a homology model for the catalytic domain of PC1 based on the crystal structure of furin (PDB: 1P8J) and an NMR solution structure of the PC1 propeptide (PDB: 1KN6) docked onto the catalytic domain using the subtilisin structure as a reference, was used for eukaryotic subtilases. Results were analyzed and plotted using the UCSF Chimera package [183].

2.3.2 Data acquisition

The BioMart interface of the InterPro database [184] was used to download UniProt sequence identifiers, start and stop positions, and taxonomy identifiers of annotations from the entries PF00082, PF00112, and PF00656 of the PFAM database for subtilases, cathepsins, and caspases, respectively [143], [153].

Protein sequences were downloaded from the UniProt database. Sequences with two annotated catalytic domains or those marked as deprecated in the UniProt database were discarded. Phylogeny was downloaded from the PFAM database, and taxonomy was obtained from the NCBI Taxonomy homepage.

2.3.3 Tree construction

NCBI taxonomy based trees were constructed using taxonomy identifiers as input for the iTol Tree generator [185] and adding each protein as a node of their species. Trees were plotted using the ‘ape’ package written in R statistical computing language [186], [187].

2.3.4 Statistical testing

A non-parametric Mann-Whitney test was performed to assess differences in the distribution of $\Delta[AA]$ between prokaryotes and eukaryotes using the R statistical computing language. The effect size was calculated as U/mn , by dividing the test statistic U by the product of the two sample sizes [188].

2.3.5 Sliding window analysis

For each sequence the number of histidines, $\#His(i,k)$, in a window of length k starting at position i , ranging from 1 to $n-k+1$ were counted, where n is the length of the sequence. To account for different sequence lengths, the starting sequence positions were normalized as follows:

$$\#His_{Norm}(i,k) = \#His\left(\frac{i}{\tilde{n}} * n, k\right) \quad (1)$$

Where, \tilde{n} is the median sequence length and the term $\frac{i}{\tilde{n}} * n$ was rounded to the nearest integer. For each position i , the $\#His_{Norm}(i,k)$ values were averaged and

then divided by k to obtain the average histidine content, $\#His(i)$, at that position. This method assumes that differences in length due to insertion and deletions are evenly distributed within the protein. Using a multiple sequence alignment (MSA) for normalization would potentially account better for the position of insertions and deletions, but the number of sequences and the low quality of the alignment especially in the propeptide region rendered this approach impractical.

2.4 pH-sensors in the propeptides of subtilases

Subtilases, a ubiquitous super-family of serine proteases, represents an ideal group of homologs to analyze protein adaptation to eukaryotic organelles. Bacterial subtilases are mostly secreted and undergo *pH-independent* activation, while eukaryotic subtilases undergo *pH-dependent* activation, usually in specific organelles [2], [18]. Bacterial subtilisin and mammalian proprotein convertase (PC) sub-families constitute the most extensively studied subtilases [17], [58]. Despite evolutionary divergence, proteins in these subfamilies display common folds with conserved catalytic triads, and are almost always expressed as zymogens, with N-terminal and occasionally C-terminal propeptide extensions. Similar to bacterial subtilases, propeptides of PCs assist folding and require two ordered steps of proteolytic cleavage for activation [2]. Our understanding of PC-activation is based on studies of pro-furin, a constitutively expressed PC-homolog; the first cleavage occurs at a consensus site $RTKR_{107} \downarrow$ after protein folding in the ER and results in a non-covalent propeptide:protease complex. Activation requires an additional cleavage at the internal site ${}^{69}HRGVTKR_{75} \downarrow$ in

the furin propeptide, which only occurs when furin trafficks into the trans-Golgi network/endosomal system (TGN) [58]. Other PCs are activated in a similar manner, but within different organelles (8). Studies establish that (i) a conserved histidine residue (His₆₉) in the propeptide of furin acts as a pH-sensor [189], (ii) the pH of the TGN is sufficient to trigger the second, activating cleavage of furin [18], (iii) propeptide-domains in PCs contain sufficient information to mediated pH-dependent activation of cognate proteases and undergo conformational changes that correlate with their respective pH of activation [190] (Chapter 3). Prokaryotic propeptides are stable over the pH range. Amino acid composition analysis shows an increased histidine content in propeptides of furin and PC1, compared to the average content in the UniProt database, while the bacterial propeptides show no such bias [191].

To identify conserved sequence elements unique to either prokaryotic or eukaryotic subtilases, we performed an evolutionary conservation analysis using the ConSurf server [144]. Analysis of prokaryotic subtilisin and eukaryotic proprotein convertase families was initiated using sequences of Subtilisin E and Proprotein Convertase 1/3 (PC1/3), respectively. The resulting conservation scores were mapped on the X-ray structure of the propeptide:Subtilisin E complex (PDB: 1SCJ) and on a homology model of the propeptide:PC1/3 complex (based on PDB: 1P8J and 1KN6), respectively (Figure 2-1). Catalytic domains of eukaryotic and prokaryotic subtilases depict a highly conserved core. On the contrary, propeptides demonstrate less sequence conservation, with the dibasic

cleavage motif at the C-terminus of eukaryotic propeptides representing the only conserved region.

Since His₆₉ was demonstrated to function as a pH sensor in furin [189], we analyzed whether histidine residues demonstrate any sequence conservation within propeptides. Although we could not identify absolutely conserved histidines in propeptides of eukaryotic subtilases, several positions in our alignment contain a histidine in a substantial fraction of sequences, especially at the position corresponding to His₆₉ in furin (53.3% of sequences). In contrast, prokaryotic subtilases, which do not traverse the secretory pathway, appear to encode less histidines within their propeptides. When catalytic sequences are compared, we find strictly conserved histidines within prokaryotic and eukaryotic sequences, and studies indicate that they play essential roles in catalysis or protein stability [192]. Hence, biased enrichment for histidines appears localized within propeptides of eukaryotic subtilases. This is consistent with the hypothesis proposed in this chapter that encoding of histidines in the propeptides allows for sensing organellar pH to direct activation. Such a broad hypothesis is difficult to test experimentally, as it would require biochemical studies on a large number of proteins. However, amino acid composition can be easily calculated from large amounts of available sequence data, and one would expect a consistent bias for histidine content in propeptides of eukaryotic, but not prokaryotic proteins.

To examine whether residue specific biases exist within eukaryotic propeptides, we computed the abundance of individual residues in propeptides

and catalytic domains in 6,533 unique subtilases from the PFAM database entry PF00082 [153]. We calculated for each sequence the difference in histidine content ($\Delta[\text{His}]$) between the propeptide ($[\text{His}]_{\text{Pro}}$) and the catalytic-domain ($[\text{His}]_{\text{Cat}}$). Positive $\Delta[\text{His}]$ values indicate abundance of histidines in propeptides, negative values signify abundance in proteases, while near zero values imply equal distribution. While $\Delta[\text{His}]$ values in individual proteins may be subject to random fluctuations, the absence of any functional requirements would result in a distribution centered around zero. If histidine residues in propeptides are required for the experimentally observed function of sensing organelle specific pH, they would be selected during evolution, and one would expect statistical bias for positive $\Delta[\text{His}]$ only within eukaryotic subtilases and near zero or negative $\Delta[\text{His}]$ for prokaryotes.

For initial assessments, we plotted $\Delta[\text{His}]$ on a phylogenetic tree generated by the PFAM database (Figure 2-2A), which is consistent with homology groups defined earlier [193], with the largest clades representing subtilisin, kexin, proteinase K, and pyrolisin, as well as the later characterized sedolisin family [194]. While the subtilisin family is exclusive to prokaryotes, the remaining four families contain eukaryotic and prokaryotic proteins, suggesting these families diverged before speciation. Interestingly, three of the four families display predominantly positive $\Delta[\text{His}]$ in eukaryotes, but not in prokaryotes. Only sedolisins show positive $\Delta[\text{His}]$ values in both prokaryotes and eukaryotes. The distributions of $[\text{His}]_{\text{Pro}}$ and $[\text{His}]_{\text{Cat}}$ (Figure 2-2B) establish that catalytic-domains in subtilases display a distribution centered on 2%, with eukaryotes

having slightly higher $[\text{His}]_{\text{Cat}}$ values than prokaryotes, as expected from the average histidine content in the UniProt database (2.3%). While, the distribution of $[\text{His}]_{\text{Pro}}$ within propeptide-domains in prokaryotes shifts towards lower values, with several lacking histidines, the $[\text{His}]_{\text{Pro}}$ distribution in eukaryotic propeptides is shifted to higher values, much greater than in catalytic-domains. This bias is clearly evident by the $\Delta[\text{His}]$ distributions (Figure 2-2C) with median values of -0.56% and 1.5% in prokaryotes and eukaryotes, respectively. This difference in distribution between prokaryotes and eukaryotes is unique to histidine and not observed for any other amino acid (Figure 2-3 and Figure 2-4).

To quantify the significance of the difference in distribution between species, we employed a non-parametric Mann-Whitney test (Table 1) across all twenty amino acids. The test resulted p-values <0.05 for several amino acids indicating statistically significant differences in $\Delta[\text{AA}]$ distributions between eukaryotes and prokaryotes. Since large sample sizes can result in statistically significant p-values even for tiny differences, a more meaningful, sample size independent, measure of the difference in distribution can be obtained using effect sizes (U/mn) [188]. These values vary between 0.0 and 1.0, and estimate the probability that a random sample of $\Delta[\text{AA}]$ in eukaryotes is larger than a random sample of $\Delta[\text{AA}]$ in prokaryotes. Equal distribution of $\Delta[\text{AA}]$ in both species would result in an effect size of 0.5. As seen in Figure 2-2D, histidine shows the highest deviation from 0.5, suggesting this bias is not by pure chance. Only cysteine deviates substantially (more than 0.15 units) from 0.5, which is likely due to higher frequency of disulfide bonds in eukaryotes than prokaryotes. The

fact that deviation from 0.5 in the effect size for histidine is considerably greater than that observed for cysteine suggests a biological significance for a histidine bias.

Since errors in database annotation and differences in length between propeptides and catalytic-domains may result in a false-positive bias, we developed a test that is independent of the start annotation in the PFAM database. We calculated histidine content in a 20-residue sliding window from the beginning of the sequence to the end of catalytic-domain for all sequences. After normalization to sequence length, we averaged the resulting histidine content profiles for eukaryotic and prokaryotic proteins. Eukaryotic proteins alone show an increase in histidine content in the first 100 residues corresponding to the propeptide (Figure 2-2E), while both species have increased histidine content at positions 200-250, likely due to the active-site histidine, along with a small increase at the C-terminus due to a conserved histidine in the catalytic-domain.

To decipher correlations that may exist between the histidine bias and experimental evidence of pH-dependent activation, we analyzed histidine contents in propeptides and catalytic-domains of individual proteins (Figure 2-2F). Most bacterial proteins display similar histidine content within propeptides and catalytic-domains (approximately 2%), with only Kumamolisin and Xanthomonolisin displaying histidine enrichment (>4%) in cognate propeptides. Consistent with our hypothesis, both proteins undergo activation at acidic pH *in vitro* [195], [196], which is not surprising because their hosts display

optimum growth under acidic conditions. Since intracellular pH within cells is maintained near neutral, pH sensing is an ideal mechanism for discerning intracellular and extracellular environments. Both proteins belong to the sedolisin family, which has evolved to function under acidic conditions [194], explaining the histidine bias in propeptides in eukaryotes as well as prokaryotes. Eukaryotic propeptides display histidine contents > 4%, excepting Proteinase K and SKI-1. Expression of proteinase K in *E. coli* produces active protease [197], and SKI-1 loses its propeptide in the ER [198], suggesting activation occurs at neutral pH, relaxing the necessity for histidines.

2.5 pH-sensors in the propeptides of cathepsins

To investigate whether our hypothesis applies to other pH-activated, propeptide-dependent proteases, we analyzed histidine content in cathepsins, a large family of lysosomal cysteine peptidases [199], which similar to subtilases, can activate at acidic pH. Due to these parallels, we hypothesized that eukaryotic cathepsins should show a similar bias for histidine in their propeptides.

We plotted the phylogenetic tree for cathepsin sequences (PFAM family PF00112) along with their $\Delta[\text{His}]$ values (Figure 2-5A) in a manner identical to subtilases.

The two major well-studied cysteine cathepsin subfamilies are the cathepsin L-like (CatL-like) and the cathepsin B-like (CatB-like) families, both of which activate at low pH [199]–[201]. The CatL-like family includes human cathepsins L, V, H, K, and S and the CatB-like family includes human cathepsin B.

Additionally, humans encode five more cathepsin genes, which we do not discuss here due to their inability to autoactivate (cathepsin C and O) or unusual

long or short propeptides (cathepsin F and X, respectively) [199]. The CatL-like family shows positive $\Delta[\text{His}]$ values, the CatB-like family does not. Moreover, the distributions of $[\text{His}]_{\text{Pro}}$ and $[\text{His}]_{\text{Cat}}$ in the CatL-like family mimics eukaryotic subtilases (Figure 2-5B), while the CatB-like family displays increased $[\text{His}]_{\text{Pro}}$ and $[\text{His}]_{\text{Cat}}$ values, leading to near-zero $\Delta[\text{His}]$ values (Figure 2-5C). Prokaryotic cathepsins show similar distributions as prokaryotic subtilases. The small number of prokaryotic sequences precludes a statistical comparison between species with robustness similar to subtilases.

We next validated the increased histidine content in the CatL-like family propeptides and mapped the specific location of increased histidine content in the CatB-like sequences using the sliding window analysis (Figure 2-5D). Prokaryotic cathepsins showed low histidine content throughout the sequence, with one peak between residues 250 and 300, which is due to the catalytic histidine. Consistent with our hypothesis, an increase in histidine content exists within the first 100 residues of CatL-like sequences. Interestingly, CatB-like sequences show a moderate increase in histidine content within the first 100 residues compared to prokaryotes, along with a second peak corresponding to the occluding loop within the catalytic-domain (Figure 2-5D and E). A comparison of the crystal structures of CatL and CatB (Figure 2-5E) shows that while the catalytic-domains of the two families are similar. However, compared to CatL, the CatB propeptide is truncated while its occluding loop in the catalytic-domain is longer and form direct contacts with its propeptide by extending into the region occupied by the CatL propeptide in a complex with its

cognate protease. Notably, histidines within the occluding loop of CatB occupy similar spatial locations as histidine residues within the CatL propeptide (Figure 2-5F). This suggests that the pH-sensing capability in CatB is encoded not only within the propeptide, but also in the occluding loop within the catalytic-domain, which is consistent with experimental data demonstrating that the occluding loop interacts with the propeptide in a pH-dependent manner, and histidine to alanine substitutions within the occluding loop blocks activation [202]. Moving pH-sensitivity from propeptides onto catalytic-domains provides evolutionary advantages to CatB-like family by enabling members to switch between endo- and exopeptidases in a pH dependent manner [203]. In summary, histidine distribution in cathepsins is consistent with our hypothesis, although subtle variations can exist within individual propeptide-dependent protease families.

2.6 The cytosolic caspase family encodes no pH-sensors in their propeptides

Our hypothesis assumes that eukaryotic proteases require histidines in their propeptides to sense the pH of the secretory pathway. Therefore cytosolic proteases, such as the caspase-family, would be expected to show no histidine bias within their propeptides. Caspases are responsible for initiating apoptosis within eukaryotic cells [204] and are expressed as inactive zymogens that are activated by proteolytic processing similar to subtilases and cathepsins.

We plotted the phylogenetic tree for caspase sequences (PFAM family PF00656) along with their $\Delta[\text{His}]$ values (Figure 2-6A) in a manner identical to subtilases

and cathepsins. The phylogenetic tree demonstrates that caspase homologs are found in metazoans, fungi, and plants. We excluded metacaspases (homologs in fungi and plants) from our analysis because their propeptides contain histidine residues that are involved in zinc binding [205]. Metazoan caspases demonstrate increased $[\text{His}]_{\text{Cat}}$ values, while the $[\text{His}]_{\text{Pro}}$ values are similar to that of prokaryotic caspases (Figure 2-6B). Consistently, $\Delta[\text{His}]$ values were slightly smaller for eukaryotic proteins (Figure 2-6C). The sliding window analysis of prokaryotic and eukaryotic caspases shows that there is no substantial histidine enrichment in the N-terminal residues (Figure 2-6D). Overall these results are consistent with the assumption that functional requirement of histidines in propeptides is unique to proteases that need to sense pH to direct their activation.

2.7 Implications for other proteins and disease

Since histidine enrichment correlates with pH-mediated activation in subtilases and cathepsins, we propose that it can be used to predict proteins that use a similar mechanism for activation. A list of all human proteins with annotated propeptides in the UniProt database, which have more histidines in their propeptides than expected, assuming a probability for histidine of 2.3%, (Table 2.2) includes 52 proteins that are either secreted or targeted to the secretory or endocytotic pathway. While this bias can be random or caused by other factors, such as zinc binding sites, which could explain why metalloproteases such as ADAM and matrixmetalloprotease family members are frequent in the list, we propose that proteins with high histidine content in their propeptides use the pH

of the secretory pathway to regulate their activation. This not necessarily applies exclusively to proteases, as other proteins can also be inhibited by their propeptides. One example is Bone morphogenic protein 4 (BMP4), which has a histidine content of 6.23% (Table 2.2) suggesting a propeptide mediated pH-sensing mechanism. Indeed, sequential processing of the propeptide of BMP4 by furin is pH-dependent and a histidine residue in the propeptide (His251) has been implicated as pH-sensor [206]

Our hypothesis suggests a prominent role of the pH gradient in the secretory pathway in orchestrating proteolytic processing of secreted proteins. Any disturbances in this gradient could lead to dysregulation of protease activity, which in PCs and cathepsins can have adverse effects, and are associated with diseases like cancer, atherosclerosis, and Dent's disease [206][207]. Since all these diseases are also associated with changes in cytosolic pH [208], [209], studies that address whether the secretory pH-gradient is also affected are needed to address the question of whether pH-dysregulation plays a role in disturbing regulation of the secretory pathway.

After a review of sequences in three evolutionary unrelated protease families we find a correlation of increased histidine content in propeptides with the requirement to sense pH. But does a correlation imply causality? Histidines play multiple unique roles in proteins because they can (i) function as proton exchangers in enzyme catalysis, (ii) form complexes with soft metals, (iii) provide unique hydrogen bonding geometry, and (iv) alter protein structure and interactions in a pH-dependent manner. Since propeptides are not part of the

active site that mediates proteolysis, and because propeptides analyzed in this study do not bind metal ions, one can exclude the first two roles. It is also unlikely that propeptides in eukaryotes have different hydrogen bonding requirements than their prokaryotic orthologs, thus endorsing their roles as pH-sensors as the most likely explanation for the observed histidine bias.

Table 2.1: Results of Mann-Whitney tests to evaluate differences in distribution of $\Delta[AA]$ between eukaryotes and prokaryotes.

Residue	Eukaryotes	Prokaryotes	U	p	U/mn
A	-2.01	-0.29	3484667	6.3×10^{-56}	0.38
V	-0.03	-0.13	4692108	1.4×10^{-1}	0.51
L	1.27	1.53	4494450	1.8×10^{-1}	0.49
I	-0.29	-0.61	4845335	2.4×10^{-4}	0.53
M	-0.32	-0.44	4781449	5.7×10^{-3}	0.52
F	0.53	0.06	5019341	7.3×10^{-10}	0.55
Y	-0.27	-1.00	5564109	3.6×10^{-44}	0.61
W	-0.46	-0.92	5529692	3.2×10^{-41}	0.60
S	-0.47	-0.18	4329195	2.2×10^{-4}	0.47
T	-1.36	-0.07	3616488	9.2×10^{-44}	0.39
N	-1.63	-1.86	4339854	3.9×10^{-4}	0.47
Q	1.11	1.49	4198344	2.6×10^{-8}	0.46
C	-1.25	-0.28	2881834	7.5×10^{-132}	0.31
G	-5.2	-5.3	4576112	8.7×10^{-1}	0.50
P	-0.67	0.04	3852582	8.5×10^{-26}	0.42
D	-0.25	-1.52	5644738	1.9×10^{-51}	0.62
E	2.85	2.51	4864907	7.7×10^{-5}	0.53
H	1.53	-0.56	7048731	1.6×10^{-270}	0.77
K	1.45	1.58	4356812	9.6×10^{-4}	0.47
R	1.86	0.95	5376156	2.2×10^{-29}	0.59

For each amino acid the following numbers are reported: Median of $\Delta[AA]$ for eukaryotes and prokaryotes, test statistic of the Mann-Whitney test, the resulting p-value, the effect size U/mn. Sample sizes were 2156 and 4256 for eukaryotes and prokaryotes, respectively.

Table 2.2: List of human proteins with histidine enrichment in their propeptides

UniProt identifier	Name	Length propeptide	Histidine content	P(X>=k)	Cellular location	Location
P12821	Angiotensin-converting enzyme	74	6.76%	2.80E-02	Secreted, Cell membrane	C-terminus
O14672	ADAM10	194	7.73%	5.08E-05	Cell membrane, Endomembrane	N-terminus
O75078	ADAM11	202	4.46%	4.54E-02	Cell membrane	N-terminus
Q9Y3Q7	ADAM18	168	4.76%	4.15E-02	Cell membrane	N-terminus
Q9P0K1	ADAM22	197	6.09%	2.22E-03	Cell membrane	N-terminus
O75077	ADAM23	227	4.85%	1.69E-02	Cell membrane	N-terminus
Q9UKF2	ADAM30	171	4.68%	4.52E-02	Cell membrane	N-terminus
Q9BZ11	ADAM33	174	5.17%	2.01E-02	Cell membrane	N-terminus
O15204	ADAM-like protein decysin-1	175	6.29%	2.61E-03	Secreted	N-terminus
Q9H324	ADAM-TS10	208	5.29%	9.33E-03	Secreted	N-terminus
P58397	ADAM-TS12	215	6.05%	1.59E-03	Secreted	N-terminus
Q8TE57	ADAM-TS16	255	5.88%	9.60E-04	Secreted	N-terminus
Q8TE60	ADAM-TS18	237	5.91%	1.34E-03	Secreted	N-terminus
P59510	ADAM-TS20	232	4.74%	1.96E-02	Secreted	N-terminus
Q9UKP5	ADAM-TS6	223	8.52%	1.31E-06	Secreted	N-terminus
Q9P2N4	ADAM-TS9	269	4.09%	4.88E-02	Secreted	N-terminus
O95972	Bone morphogenetic protein 15	249	5.22%	5.55E-03	Secreted	N-terminus
P12643	Bone morphogenetic protein 2	259	5.79%	1.12E-03	Secreted	N-terminus
P12644	Bone morphogenetic protein 4	273	6.23%	2.38E-04	Secreted	N-terminus
P18075	Bone morphogenetic protein 7	263	5.70%	1.30E-03	Secreted	N-terminus
P55287	Cadherin-11	31	12.90%	5.36E-03	Cell membrane	N-terminus
Q13634	Cadherin-18	29	17.24%	4.82E-04	Cell membrane	N-terminus
P12830	Cadherin-1	132	6.06%	1.17E-02	Cell membrane, Endosome, Golgi apparatus	N-terminus
P14091	Cathepsin E	34	8.82%	4.29E-02	Endosome	N-terminus
P09668	Cathepsin H	85	8.24%	3.52E-03	Lysosome	N-terminus
P43235	Cathepsin K	99	6.06%	2.71E-02	Lysosome	N-terminus
P25774	Cathepsin S	98	8.16%	1.97E-03	Lysosome	N-terminus
Q6YHK3	CD109 antigen	25	12.00%	1.92E-02	Cell membrane	C-terminus
P0CG37	Cryptic protein	65	7.69%	1.70E-02	Cell membrane	C-terminus
Q14126	Desmoglein-2	26	11.54%	2.13E-02	Cell membrane	N-terminus
P12259	Coagulation factor V	836	3.59%	1.27E-02	Secreted	Internal
P02765	Alpha-2-HS-glycoprotein	40	12.50%	2.17E-03	Secreted	Internal
P09958	Furin	83	6.02%	4.28E-02	Golgi apparatus, Cell membrane	N-terminus

O60383	Growth/differentiation factor 9	295	4.41%	2.04E-02	Secreted	N-terminus
P07686	Beta-hexosaminidase subunit beta	79	6.33%	3.58E-02	Lysosome	N-terminus
P55103	Inhibin beta C chain	218	4.59%	3.07E-02	Secreted	N-terminus
P58166	Inhibin beta E chain	217	5.07%	1.25E-02	Secreted	N-terminus
P51460	Insulin-like 3	47	14.89%	9.55E-05	Secreted	Internal
P19827	ITI heavy chain H1	246	4.47%	2.84E-02	Secreted	C-terminus
Q99538	Legumain	110	9.09%	2.40E-04	Lysosome	C-terminus
P09848	Lactase-phlorizin hydrolase	847	3.78%	5.15E-03	Apical cell membrane	N-terminus
P10253	Lysosomal alpha-glucosidase	42	9.52%	1.56E-02	Lysosome	N-terminus
P14151	L-selectin	10	20.00%	2.11E-02	Cell membrane	N-terminus
Q9NRE1	Matrix metalloproteinase -26	72	8.33%	6.34E-03	Secreted	N-terminus
P16519	PCSK2	84	8.33%	3.29E-03	Secretory vesicle	N-terminus
P29122	PCSK6	86	5.81%	4.86E-02	Secreted, Endoplasmic reticulum	N-terminus
P01127	PDGF subunit B	112	5.36%	4.53E-02	Secreted	N- and C-terminus
Q96B86	Repulsive guidance molecule A	147	5.44%	2.10E-02	Cell membrane	N- and C-terminus
P10600	Transforming growth factor beta-3	280	5.00%	5.96E-03	Secreted	N-terminus
Q9BZD6	Proline-rich Gla protein 4	32	9.38%	3.68E-02	Cell membrane	N-terminus
Q8N2E6	Prosalsin	163	5.52%	1.37E-02	Secreted	N-terminus
O43915	Vascular endothelial growth factor D	216	6.02%	1.66E-03	Secreted	N- and C-terminus

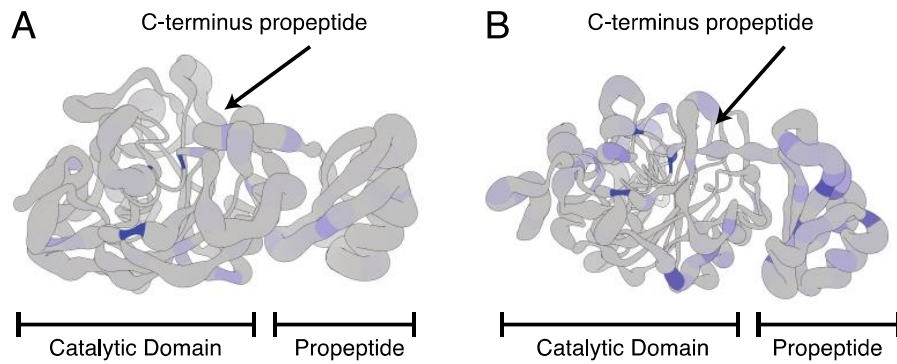


Figure 2-1: Propeptides are more divergent than cognate catalytic domains. Conservation scores mapped onto a ribbon presentation of (A) Subtilisin E and (B) PC1/3. Thick tubes represent high divergence at this position while thin tubes represent conservation. Color indicates percentage of sequences that encode a histidine residue at this position from 0% (grey) to 100% (blue)

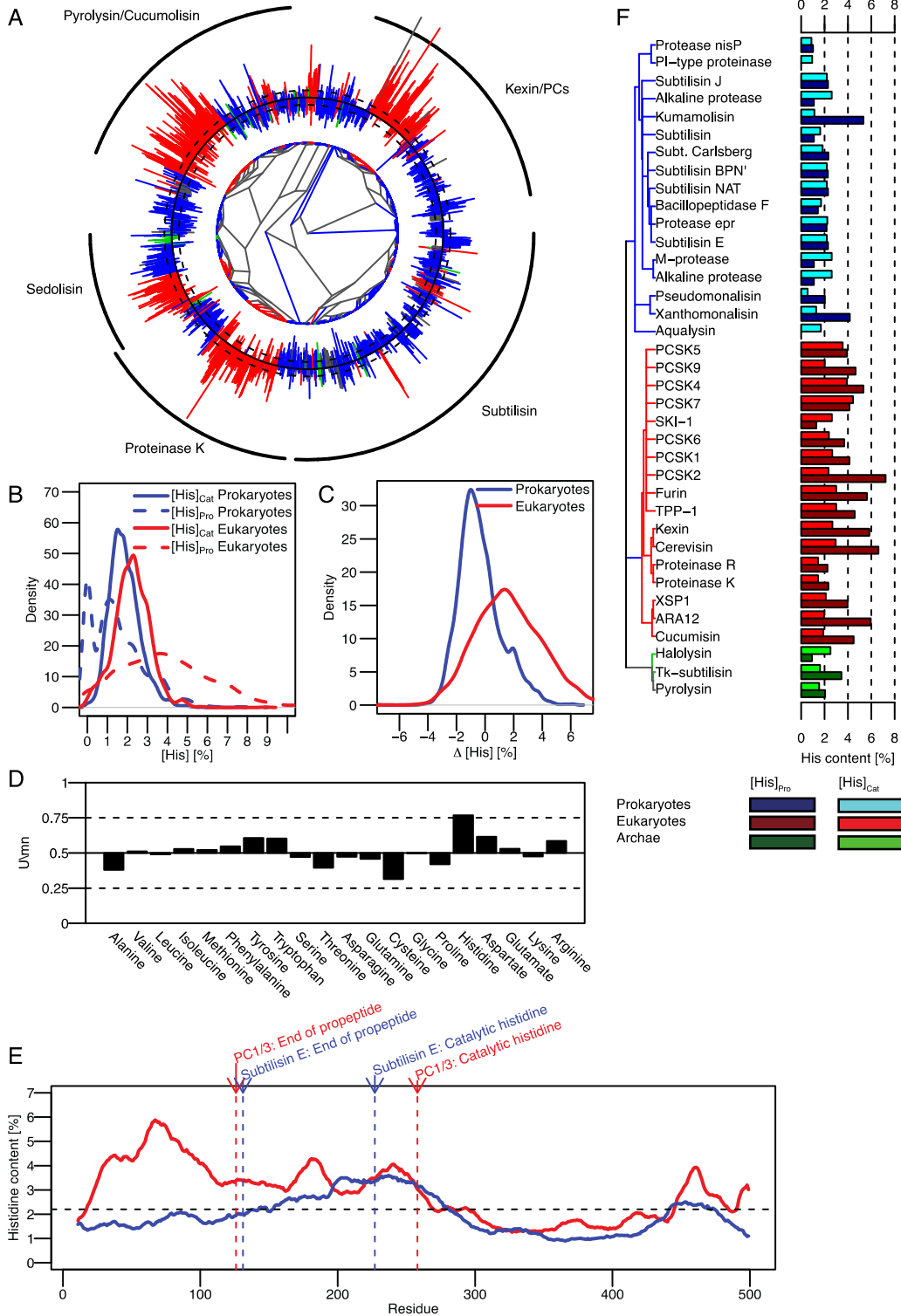


Figure 2-2: Histidines are enriched in propeptides of eukaryotic, but not prokaryotic, subtilases.

(A) Phylogenetic tree of subtilases from the PFAM database. Bars on the outside indicate the $\Delta[\text{His}]$ value of each sequence. A black circle represents 0%. Bars pointing outward and inward represent positive and negative $\Delta[\text{His}]$ values, respectively. Dashed circles outside and inside of the solid black circle represent $\Delta[\text{His}]$ values of $\pm 1\%$. Eukaryotic, prokaryotic, and archaean sequences are colored red, blue, and green, respectively. Black arcs on the outside mark the clades of major subtilase subfamilies. (B) Kernel density estimation of the distribution of $[\text{His}]_{\text{Pro}}$ and $[\text{His}]_{\text{Cat}}$ in prokaryotes and eukaryotes. (C) Kernel density estimation of the distribution of $\Delta[\text{His}]$ for prokaryotes and eukaryotes. (D) Effect size (U/mn) of the Mann-Whitney test for difference between the distributions shown in panel C performed for all 20 natural amino acids. (E) Sliding Window Analysis of average histidine content in eukaryotic and prokaryotic subtilases using a window of 20 residues. The black dashed line indicates average histidine content in the UniProt database. Arrows indicate relative position of annotations for the end of the propeptide domain and the catalytic histidine residue according to subtilisin E and PC1/3. (F) Bar graph showing $[\text{His}]_{\text{Pro}}$ and $[\text{His}]_{\text{Cat}}$ values for selected subtilases. Blue, red, and green shades represent prokaryotic, eukaryotic, and archaean sequences, respectively. Light shades indicate $[\text{His}]_{\text{Cat}}$ and dark shades indicate $[\text{His}]_{\text{Pro}}$.

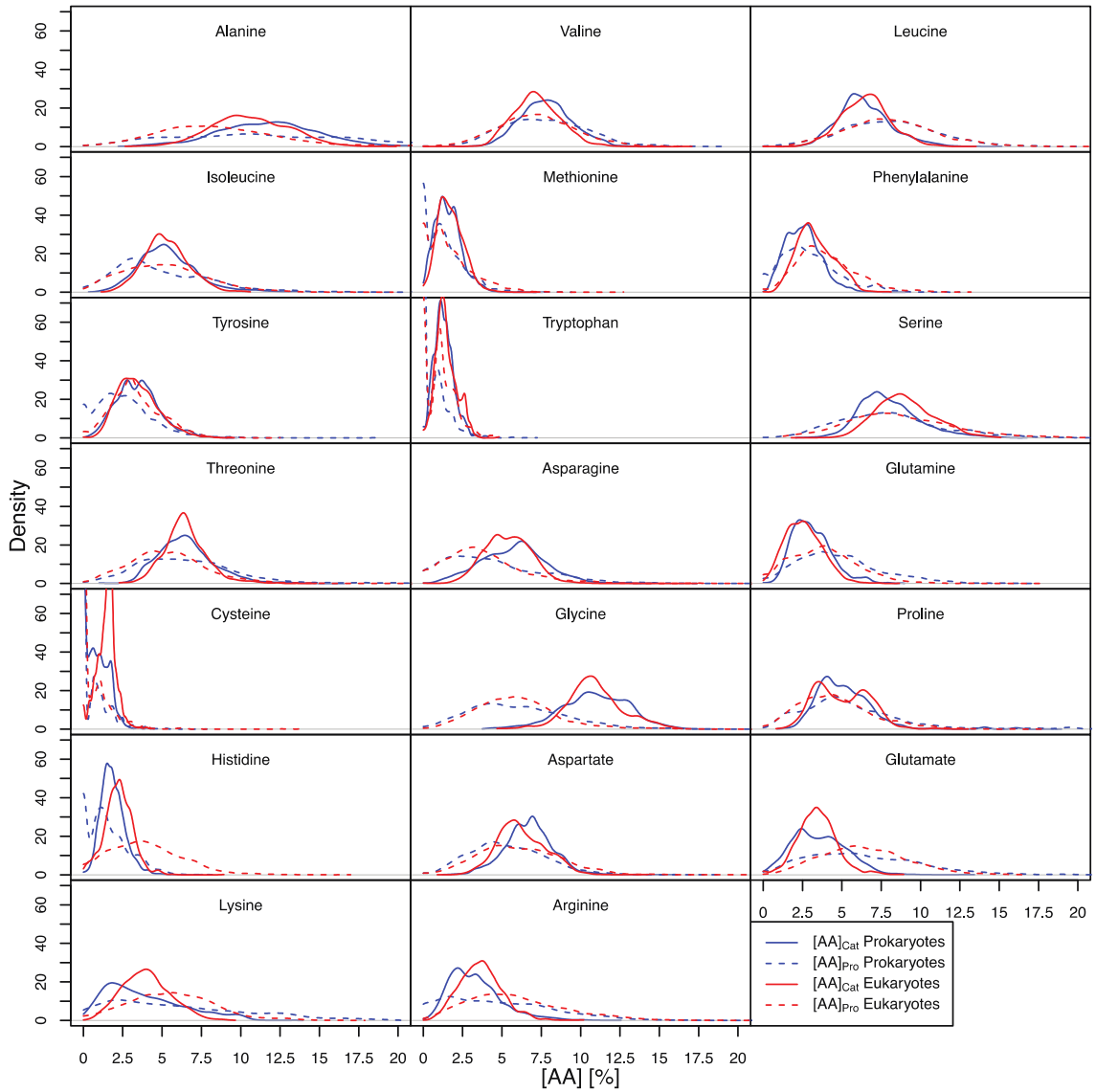


Figure 2-3: Distribution of [AA]_{Pro} and [AA]_{Cat} for all 20 amino acids in eukaryotic and prokaryotic subtilases

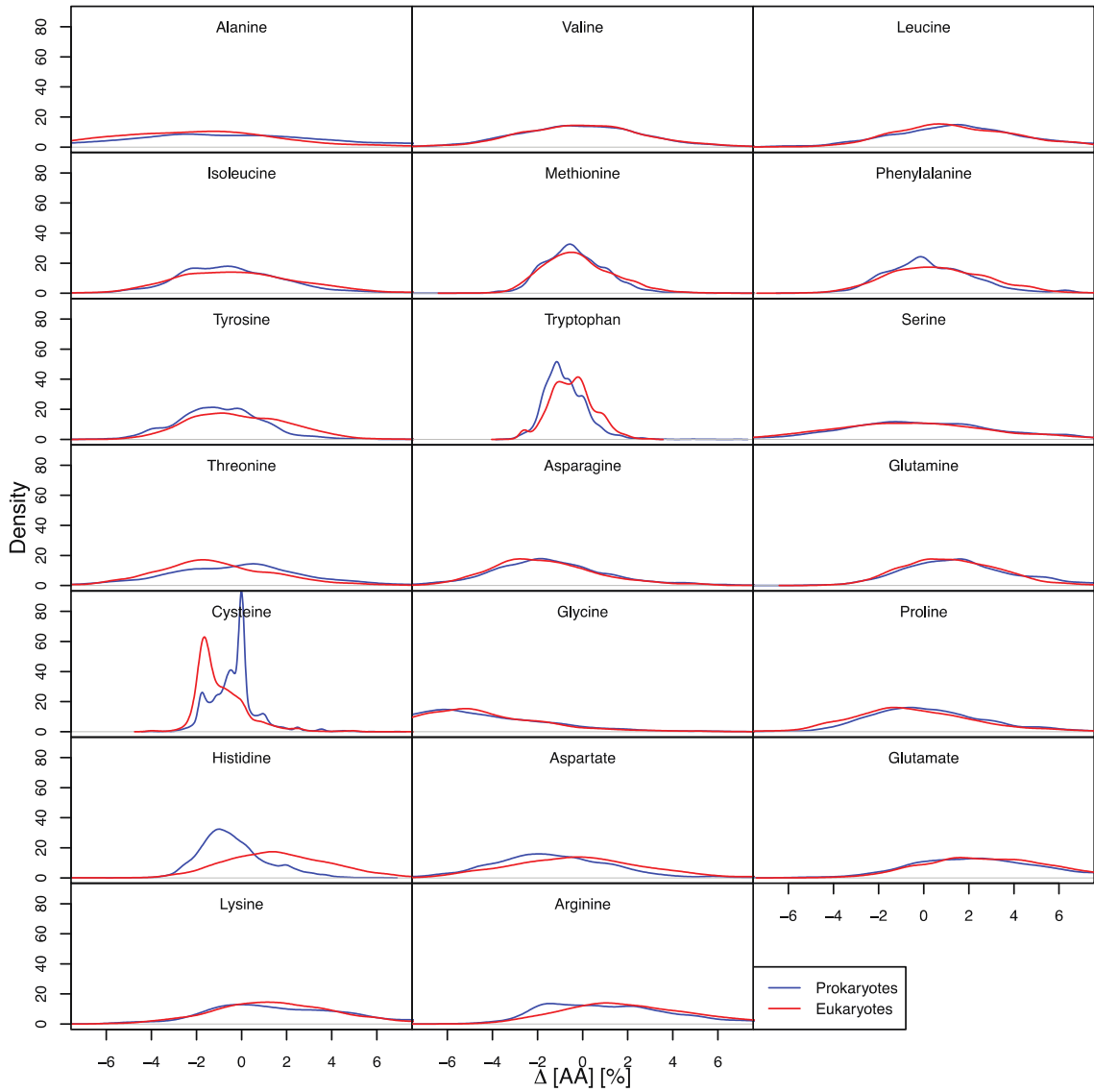


Figure 2-4: Distribution of $\Delta[AA]$ for all 20 amino acids in eukaryotic and prokaryotic subtilases

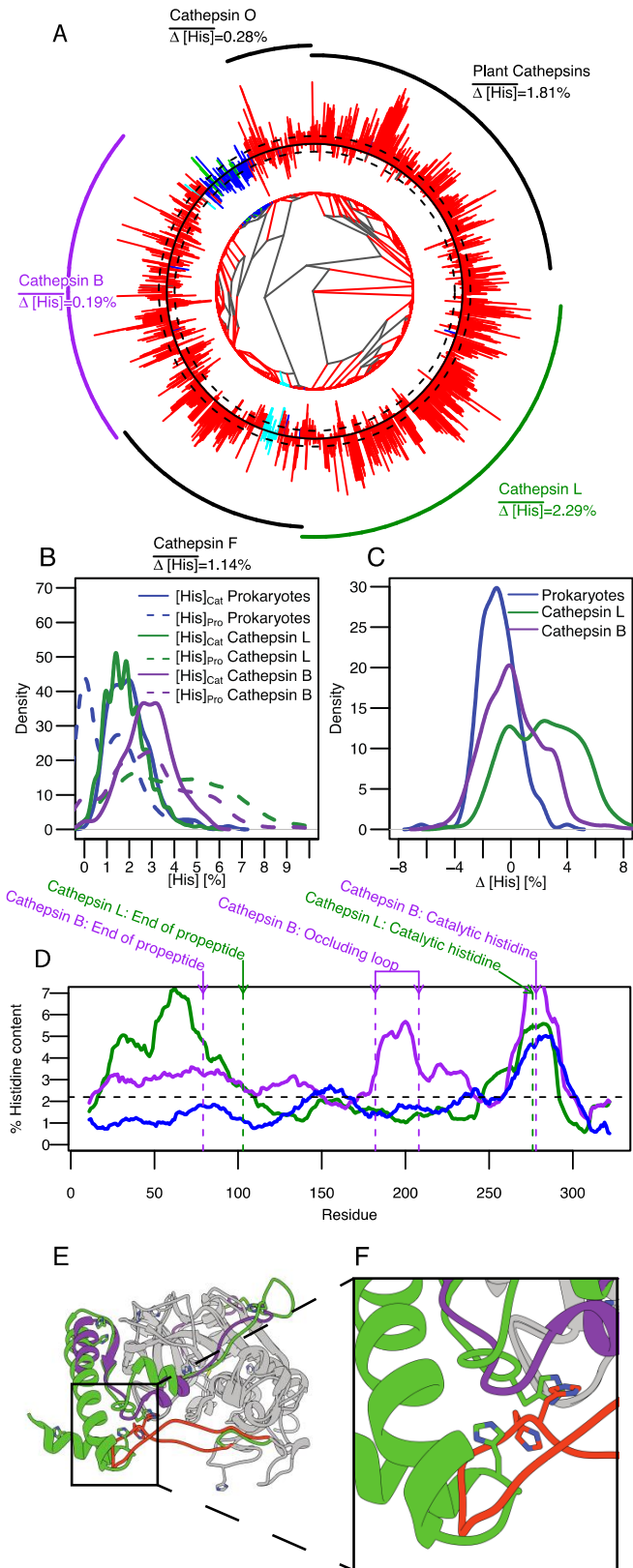


Figure 2-5: Histidine enrichment exists only in propeptide domains of the Cathepsin L family, while it is also present in the occluding loop of the Cathepsin B family.

(A) Phylogenetic tree of cathepsins from the PFAM database. Bars on the outside indicate the $\Delta[\text{His}]$ value of each sequence. A black circle represents 0%. Bars pointing outward and inward represent positive and negative $\Delta[\text{His}]$ values, respectively. Dashed circles outside and inside of the solid black circle represent $\Delta[\text{His}]$ values of $\pm 1\%$. Eukaryotic, prokaryotic, archaean, and viral sequences are colored red, blue, green, and cyan, respectively. Black arcs on the outside mark the clades of major cathepsin subfamilies, with the cathepsin L family shown in green and the cathepsin B family shown in purple. (B) Kernel density estimation of the distribution of $[\text{His}]_{\text{Pro}}$ and $[\text{His}]_{\text{Cat}}$ in cathepsin L and B families and in prokaryotes. (C) Kernel density estimation of the distribution of $\Delta[\text{His}]$ in cathepsin L and B families and in prokaryotes. (D) Sliding Window Analysis of average histidine content in cathepsin L and B families and in prokaryotes using a window of 20 residues. The black dashed line indicates average histidine content in the UniProt database. Arrows indicate relative position of annotations for the end of the propeptide domain and the catalytic histidine residue according to Cathepsin L and B, as well as the occluding loop in cathepsin B. (E) Structure superimposition of procathepsin L (PDB: 1BY8) and procathepsin B (PDB: 1MIR). The catalytic domains are shown in grey ribbon, while propeptides are shown in green and purple for cathepsin L and B, respectively. The occluding loop of cathepsin B is colored in orange and the corresponding loop in cathepsin L is colored green. The side-chains of histidine residues are depicted as stick representations. (F) A close up of interactions between the occluding loop and the propeptide. Colors are as above.

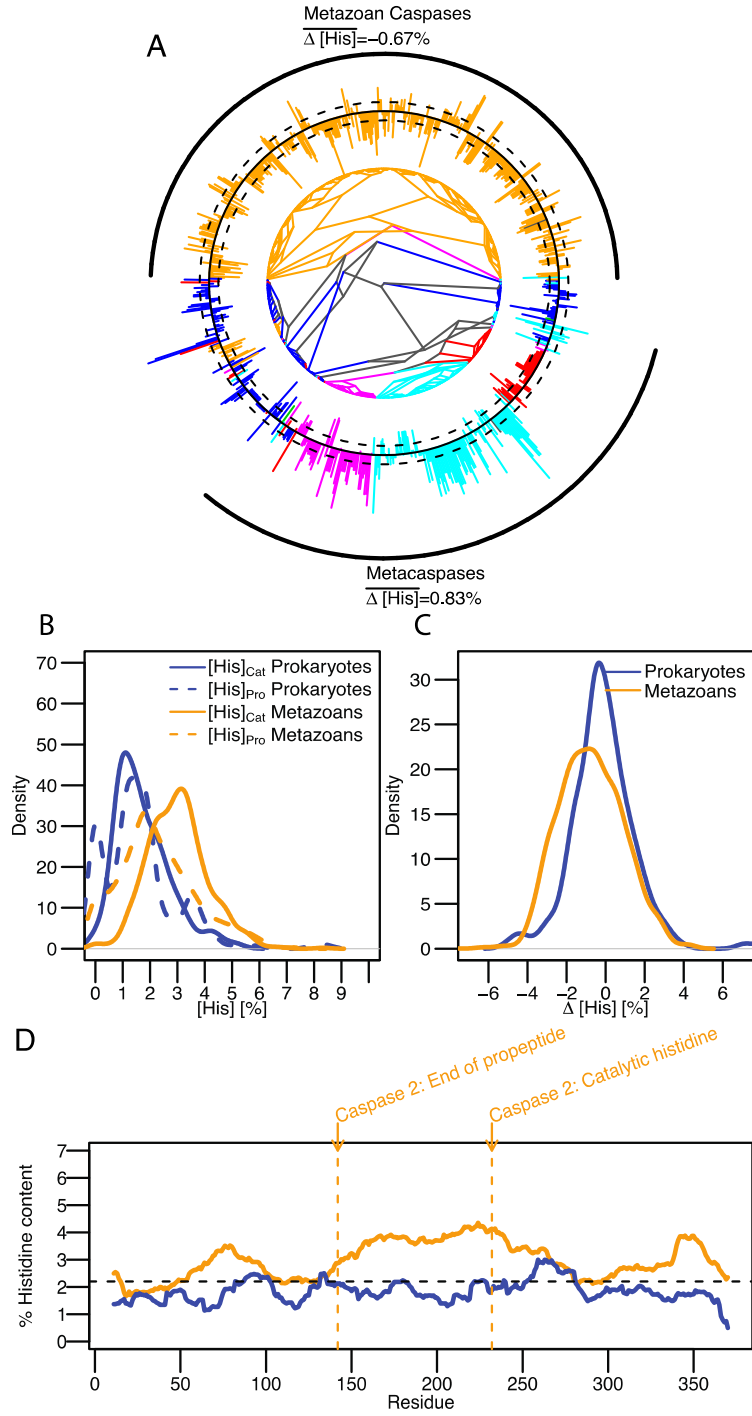


Figure 2-6: The cytosolic caspase family shows no histidine bias in propeptides.

(A) Phylogenetic tree of caspases from the PFAM database. Bars on the outside indicate the $\Delta[\text{His}]$ value of each sequence. A black circle represents 0%. Bars pointing outward and inward represent positive and negative $\Delta[\text{His}]$ values, respectively. Dashed circles outside and inside of the solid black circle represent $\Delta[\text{His}]$ values of $\pm 1\%$. Prokaryotic, metazoan, plant, fungal and other eukaryotic sequences are colored blue, yellow, cyan, purple and red, respectively. Black arcs on the outside depict the metazoan caspase and metacaspase families. (B) Kernel density estimation of the distribution of $[\text{His}]_{\text{Pro}}$ and $[\text{His}]_{\text{Cat}}$ in prokaryotes and metazoan shown in blue and yellow, respectively. (C) Kernel density estimation of the distribution of $\Delta[\text{His}]$ in metazoan and prokaryotic caspases shown in yellow and blue, respectively. (D) Sliding Window Analysis of average histidine content in metazoan and prokaryotic caspases using a window of 20 residues. Arrows indicate relative position of annotations for the end of the propeptide domain and the catalytic histidine residue according to Caspase 2.

3

The mechanism by which histidine protonation mediates activation

This chapter is based on two manuscripts. The manuscripts have been edited to remove overlaps in abstract, introduction and discussion. Some experimental data that was not central to the objective of this chapter have been omitted, but are mentioned and cited within the text when necessary. The data in Figures 3-2, 3-4, and 3-5 was generated by SLD and DMW.

Propeptides Are Sufficient to Regulate Organelle-Specific pH-Dependent Activation of Furin and Proprotein Convertase 1/3

Stephanie L. Dillon, Danielle M. Williamson, **Johannes Elferich**, David Radler, Rajendra Joshi, Gary Thomas, Ujwal Shinde

This manuscript was published by the Journal of Molecular Biology on October 12th 2013 in volume 423 issue 1 pages 47-62

The Mechanism by Which a Propeptide-encoded pH Sensor Regulates Spatiotemporal Activation of Furin

Danielle M. Williamson[†], **Johannes Elferich**[†], Parvathy Ramakrishnan¹, Gary Thomas² and Ujwal Shinde¹

[†]These authors contributed equally to this work

This manuscript was published by the Journal of Biological Chemistry on May 7th 2013 in volume 288 pages 19154-19165.

3.1 Abstract

The proprotein convertases (PCs) furin and proprotein convertase 1/3 (PC1) cleave substrates at dibasic residues along the eukaryotic secretory/endocytic pathway. PCs are evolutionarily related to bacterial subtilisin and are synthesized as zymogens. They contain N-terminal propeptides (PRO) that function as dedicated catalysts that facilitate folding and regulate activation of cognate proteases through multiple-ordered cleavages. Previous studies identified a histidine residue (His₆₉) that functions as a pH sensor in the propeptide of furin (PRO^{FUR}), which regulates furin activation at pH ~ 6.5 within the trans-Golgi network. Although this residue is conserved in the PC1 propeptide (PRO^{PC1}), PC1 nonetheless activates at pH ~5.5 within the dense core secretory granules. Here, we analyze the mechanism by which PRO^{FUR} regulates furin activation and examine why PRO^{FUR} and PRO^{PC1} differ in their pH-dependent activation. Sequence analyses establish that while both PRO^{FUR} and PRO^{PC1} are enriched in histidines when compared with cognate catalytic domains and prokaryotic orthologs, histidine content in PRO^{FUR} is ~2-fold greater than that in PRO^{PC1}, which may augment its pH sensitivity. Spectroscopy and molecular dynamics establish that histidine protonation significantly unfolds PRO^{FUR} when compared to PRO^{PC1} to enhance autoproteolysis. We further demonstrate that PRO^{FUR} and PRO^{PC1} are sufficient to confer organelle sensing on folding and activation of their cognate proteases. Swapping propeptides between furin and PC1 transfers pH-dependent protease activation in a propeptide-dictated manner *in vitro* and in cells. Structural analyses and binding

experiments comparing the wild-type PRO^{FUR} with a nonprotonatable His⁶⁹Leu mutant that blocks furin activation *in vivo* revealed protonation of His⁶⁹ reduces both the thermodynamic stability of the propeptide as well as its affinity for furin at pH 6.0. Structural modeling and molecular dynamic simulations suggests that His⁶⁹ does not directly contribute to the propeptide-enzyme interface but, rather, triggers movement of a loop region in the propeptide that modulates access to the cleavage site and, thus, allows for the tight pH regulation of furin activation. Our work establishes a mechanism by which His⁶⁹ functions as a pH sensor that regulates compartment-specific furin activation and provides insights into how other convertases and proteases may regulate their precise spatiotemporal activation.

3.2 Introduction

Subsequent to guiding protease domain folding, propeptide-dependent subtilases undergo ordered proteolytic cleavages within their propeptide domains. The first cleavage forms catalytically inactive propeptide:protease inhibition complexes wherein propeptides noncovalently bind to protease active sites, while subsequent cleavages activate proteases by facilitating propeptide dissociation, enabling the now unmasked catalytic domain to cleave substrates *in trans* [58], [59], [210], [211]. While these obligatory cleavages in prokaryotes are extracellular events that delay onset of protease activity until after protein export, they control secretory pathway compartment-specific activation of substrate-specific eukaryotic PCs [210]. Since eukaryotic PCs evolved from propeptide-dependent and not propeptide-independent prokaryotic subtilases

[212], it is tempting to speculate that propeptides confer functional advantages through speciation, namely, to regulate organelle-specific activation of secretory pathway proteases, a complexity absent in unicellular prokaryotes, but is essential to maintain physiological homeostasis within eukaryotic cells [180], [181], [213]. For example, the activation of furin is regulated in a pH-dependent manner as it transits the secretory pathway [58]. In the neutral pH in the endoplasmic reticulum (ER), the propeptide is cleaved to form a stoichiometric propeptide: furin inhibition complex. Upon reaching the early TGN (pH 6.5), the furin propeptide (PRO^{FUR}) undergoes a second cleavage, which removes the inhibitory propeptide and thus activates furin [58]. While PC1 transits the secretory pathway in much the same way, the PC1 propeptide (PRO^{PC1}) remains in a stoichiometric complex with the PC1 protease domain until it undergoes its activating second cleavage upon reaching the dense core secretory granules (DCSGs; pH 5.5). A study by Feliciangeli et al. demonstrated that in furin [59], mutating residue His₆₉ in the propeptide to a leucine can block activation of the complex in the TGN while allowing for correct folding, whereas a His₆₉Lys substitution results in accumulation of unprocessed furin precursor in the ER [59]. On this basis, they suggested that the His₆₉ in the propeptide not only is important for folding of furin but also is a vital pH sensor that regulates furin activation in the pH of the TGN. However, mechanisms by which the His₆₉ functions as a pH sensor in furin are unknown. Moreover, while the residue corresponding to His₆₉ (in furin) is strictly conserved within the PC family, PC1 and furin undergo their activating second cleavages at different pH values

within the TGN and DCSGs, respectively. This suggests that additional factors may play a role in regulating activation of the protease domains.

In this chapter, we demonstrate through various biophysical, biochemical, cell-based, and computational approaches that the PRO^{FUR} and PRO^{PC1} contain sufficient information to confer organelle sensing on the folding and activation of cognate proteases. Circular dichroism (CD) spectroscopy as a function of pH establishes that the pH-dependent stability of propeptide domains coincides with the optimum pH for compartment-specific activation. Monitored by ellipticity at 222 nm, PRO^{FUR} undergoes a transition in structure, the midpoint of which occurs at pH 6.5, while the midpoint in structural transition for PRO^{PC1} occurs at a lower pH (pH~5.5). Furthermore, swapping propeptides between eukaryotic paralogs - furin and PC1 - transfers pH-dependent protease activation in a propeptide-dictated manner in vitro and in cells. Our results suggest that PRO^{FUR} and PRO^{PC1} encode information essential for regulating compartment-specific activation of cognate proteases and that other residues in addition to the conserved pH sensor His₆₉ are necessary to enable subtle differentiation in pH-dependent activation between furin and PC1. Using molecular dynamics (MD) simulations, we also demonstrate that histidine protonation leads to conformational changes in PRO^{FUR} but not in PRO^{PC1}. Together, our results provide insights into the structural mechanisms by which propeptides can regulate the pH-dependent activation of their cognate PCs.

Although His₆₉ protonation is required for furin activation, the precise mechanism by which this pH sensor mediates activation has remained unclear.

Here, we use WT-PRO^{FUR} and the deprotonated state mimic His⁶⁹Leu-PRO^{FUR} to explore structure, stability, and pH-dependent binding, coupled with mathematical modeling and molecular dynamics, to understand how His⁶⁹ functions as a pH sensor. Taken together, our work explains the structural and mechanistic basis by which His⁶⁹ regulates compartment-specific furin activation and provides insight into how other PCs may regulate their own activation.

3.3 Materials and Methods

3.3.1 Expression and purification

Codon-optimized genes encoding human PRO^{FUR} and mouse PRO^{PC1} were synthesized from CELTEK genes, cloned into pET11a, and expressed in BL21(DE3) as previously described [214]. Inclusion bodies containing MAT^{FUR} and MAT^{PC1} were isolated and proteins were purified using reverse-phase chromatography. Enzymatically active MAT^{FUR} and MAT^{PC1} were obtained from recombinants expressing human VV:fur/f/ha/ Δ TCK and mouse VV:mPC1 in BSC40 cells as previously described [59]. Cos-7 cells were maintained in Dulbecco's modified Eagle's medium high glucose medium (HyClone) containing 10% fetal bovine serum and 1% penicillin-streptomycin. Cells were incubated at 37 ° C in a 5% CO₂ environment as previously described [59].

3.3.2 CD studies

CD measurements were performed on an AVIV model 215 CD spectrometer using a 1 mm path-length cell at 4 ° C as described earlier [17], [212]. Briefly, propeptide samples (4 mg/mL) stored in 6 M GdnHCl (to avoid side-chain

modifications commonly seen when samples are stored in urea) were diluted to a final concentration of 0.4 mg/mL) and were refolded using stepwise dialysis against 50 mM cacodylate buffer, pH 7.4, containing 150 mM KCl (Buffer A) and decreasing amounts of urea. The proteins were dialyzed twice in Buffer A without urea, against Buffer A in different pH values (5.0-7.0), and then subjected to ultracentrifugation in TLA-100 for 30 min to remove particulates. The CD spectra between 200 and 260 nm were averaged over three independent experiments and plotted as a change in ellipticity at 222 nm as a function of pH and plotted as $[\theta]$ molar ellipticity $\text{deg cm}^2 \text{dmol}^{-1}$. The PRO^{SUB-C} structure was obtained by a difference spectra between the cleaved PRO^{SUB}:S₂₂₁C-subtilisin complex and mature subtilisin as described earlier [14].

Near-UV CD measurements were performed in 1 cm path length cells at 4°C. Titrations were performed by adding small amount of acid or base using an automated syringe pump, mixing for 2 minutes with a small stir bar within the cuvette and direct automated measurement of the resulting pH with a small pH-electrode (Microelectrodes Inc.). After mixing the CD-signal at 285 nm was measured and recorded. After titration from pH 8 to pH 4, the titration solution was changed and the very same sample was titrated back to pH 8 to demonstrate reversibility of the structural change.

3.3.3 MD simulations

1SCJ [16], 1KN6 [215] and homology models of furin derived from 1KN6, and aqualysin derived from 1SCJ were used as Protein Data Bank models for PRO^{SUB}, PRO^{PC1}, PRO^{FUR}, and PRO^{AQU}, respectively. Homology models were built using

either SWISS-MODEL or MODELLER. All hydrogen and non-protein atoms were removed and hydrogen was added back using the autoPSF function in NAMD [216]. Structures were solvated in cubes with TIP3P explicit water using VMD, with a minimum distance of 12 Å to the edge. All simulations were carried out with periodic boundary conditions, particle mesh Ewald for long-range electrostatics, and a 12-Å cutoff for non-bonded interactions with the CHARMM22 force field using NAMD (version 2.5). Snapshots were saved every 10 ps using a time step of 1 fs. The system was equilibrated by first constraining the protein and minimizing solvent for 1000 steps using a conjugate gradient algorithm. The solvent was initially equilibrated for 100 ps and then fully constrained, and the protein was minimized for 500 steps. The entire system was subsequently minimized and used in the simulations. MD simulations require defining of a potential function or a force field that describes the ways through which particles in a simulation will interact. Force fields can be defined at many levels of physical accuracy and those used in MD simulations often embody a classical treatment of particle-particle interactions, which can reproduce structural and conformational changes, but usually cannot reproduce precise chemical reactions. Therefore, to simulate the pH-dependent protonation reactions, we have approximated the pH environment by predetermining the protonation state in the starting structure, an approach that has been extensively employed in the field of molecular dynamics. For pH 7, we used the HSD parameters for histidine residues, which represent an uncharged side-chain, with a proton bound to the nitrogen atom in the delta position. To simulate an

environment of pH 6, we used the HSP parameter, which represents a positively charged histidine with protons bound to both nitrogen atoms. For testing the robustness of our simulations, we took two different models of PRO^{FUR} and PRO^{PC1} and repeated the simulations as described above. An adjustment of the pH to exact values would require a prediction of the pK_a values of individual residues, which was not practical in the given study.

MD simulations for the His₆₉Leu variant were performed in a manner identically to WT PRO^{FUR} after mutagenesis of His₆₉ using VMD.

3.3.4 Amino acid content analysis

Protein sequences for human furin, mouse PC1, subtilisin from *Bacillus subtilis*, and aqualysin from *Thermus aquaticus* families were obtained from the 50% sequence identity clusters UniRef50_P09958, UniRef50_P29120, UniRef50_P00782, and UniRef50_P08594 in the UniRef database, respectively. Subsequences representing the propeptides and the protease domain were extracted using annotation from the InterPro database entries IPR009020 and IPR000209, respectively. Sequences that were not annotated by both entries were omitted. The amino acid content of both domains in all sequences was calculated and averaged for each domain and protein family. Contents of amino acids belonging to individual groups were added and divided by the sum of their content in the whole UniProt database (release 2011_12). The multiple sequence alignment of selected prokaryotic and eukaryotic subtilases was obtained using ClustalW and colored using GeneDoc.

3.3.5 Enzyme activity assays

For all assays, 113 μM furin substrate (Abz-RVKRGLA-Tyr[3-NO₂]) in dimethyl sulfoxide was incubated with 40 μL of secreted enzyme in 155 μL of 50 mM cacodylate buffer, pH 7.0, containing 1 mM CaCl₂ and 50 mM KCl. Cacodylate buffer was used in all experiments to maintain consistency throughout the analyses. The assays were conducted on a SpectraMax-M2 spectrofluorometer equipped with a 96-well plate reader. Excitation wavelength was set at 320 nm while emission wavelength was set at 425 nm. The given values are averages of triplicate assays. The activity was normalized by quantifying the relative amounts of proteins secreted in the media using ImageJ software.

3.3.6 Isolation of *in trans* propeptide:protease complexes

Since propeptides are potent competitive inhibitors of protease paralogs [214], PRO:MAT complexes *in trans* were generated by adding 10-fold excess of PRO^{FUR} and PRO^{PC1} (~ 2 nM) to MAT^{FUR} or MAT^{PC1} (~ 0.2 nM) in 50 mM cacodylate buffer, at different pH values (5.0 to 7.4) containing 150 mM KCl in a 96-well quartz plate. Complexes were incubated for 30 min at room temperature and the activities were assayed as described earlier [59]. The percent activity at each pH was calculated using the activity of uninhibited protease as a control.

3.4 Results

3.4.1 Eukaryotic propeptides harbor an internal cleavage site loop that is missing within their prokaryotic paralogs

To understand how eukaryotic propeptides can mediate compartment - specific activation of their cognate protease domains, we compared sequences and structures of prokaryotic propeptides - subtilisin (PRO^{SUB}) and aqualysin I (PRO^{AQU}) - with eukaryotic propeptides - PRO^{PC1} and PRO^{FUR} . While several laboratories have analyzed the sequences and structures of propeptides, no detailed comparison between the sequences and structures of the propeptides of prokaryotic and eukaryotic proteins has been conducted to date. PRO^{AQU} was selected because unlike its intrinsically unfolded prokaryotic homologue PRO^{SUB} , PRO^{AQU} adopts a well-defined structure and chaperones folding of its cognate protease domain. From the PC family members, we selected PRO^{PC1} and PRO^{FUR} because despite significant sequence and structural similarity with prokaryotic orthologs (Figure 3-1A and B), they activate in different organelles along the proton gradient of the secretory pathway, a complexity missing in prokaryotes. Furin is optimally active at pH 6.5, consistent with its role in cleaving proprotein substrates in the mildly acidic environment of the TGN/endosomal system. PC1 is optimally active at pH 5.5, consistent with its role in cleaving prohormone molecules in secretory granules.

Amino acids absent between residues 75 and 81 in PRO^{SUB} (red box; Figure 3-1A) coincide with organelle-specific cleavage sites within eukaryotes (red loop; Figure 3-1B). In prokaryotic subtilases, the secondary cleavage site is fairly

promiscuous and presumably occurs in the flexible region between β_1 and α_1 (Figure 3-1B). Additionally, there are significant differences in residues 100-107 within the propeptide domains between prokaryotic subtilisins and eukaryotic PCs. This C-terminal region harbors the primary cleavage site within propeptides and interacts with the substrate binding regions within cognate proteases to initiate activation. It is noteworthy that cellular substrates of PCs contain the consensus sequence [R/K]-X_n-[R/K] ↓, identical to the primary cleavage site within propeptides [2]. Given the promiscuous specificity of bacterial subtilases when compared to the stringent substrate specificity of eukaryotic PCs, the differences between residues 100 and 107 reflect the requirement of PCs to cleave at highly conserved dibasic residues. This region reflects the divergence of propeptides from prokaryotes and eukaryotes to function with more cleavage specificity, likely due to the difference in cellular environment, namely, the inclusion of membrane-bound organelles in eukaryotes [2].

Since the previous chapter demonstrated an overall enrichment of histidine residues in the propeptides of eukaryotic subtilases, we next analyzed the fold increase in amino acid residues within the propeptides and cognate proteases within the specific protein families of this study with the UniProt database as our baseline (Figure 3-1C). The individual amino acid content for each family of propeptides and proteases was calculated and averaged. The contents of amino acids belonging to individual groups were added and divided by the sum of their content in the whole UniProt database (release 2011_12) to

obtain the fold change as described in Material and Methods. Fold values greater than 1 (varying shades of red) indicate residue enrichment in propeptide domains within an individual group, values less than 1 (varying shades of green) indicate depletion of specific residues within propeptides, while a value of 1 (white) indicates no change. This graphical representation of the fold increase in specific groups of amino acid residues (Figure 3-1C) demonstrates that the His content in PRO^{FUR} and PRO^{PCI} from eukaryotes is significantly greater than their cognate catalytic domains and prokaryotic paralogs. Furthermore, protease domains of prokaryotes are biased towards acidic and basic residues as demonstrated by Inouye et al. [217], [218], which was hypothesized to enhance kinetic stability within their catalytic domains [212]. The average composition of proteins in the UniProt database establishes histidine (2.27%) as the third least abundant residue and is ~4-fold less than leucine (9.67%), the most abundant residue. While propeptide domains generally display a bias for charged and polar residues when compared to proteases [217], it is noteworthy that within subtilases, only PRO^{FUR} and PRO^{PCI} are rich in histidine content (Figure 3-1C) when compared with PRO^{SUB} and PRO^{AQU} and cognate catalytic domains.

3.4.2 CD spectroscopy demonstrates pH-dependent structural changes in eukaryotic propeptides

Since the pK_a (~ 6.0) of the imidazole side-chain of histidine is close to physiological pH, we next investigated whether small changes in proton concentration alter pH-dependent structural stability of propeptides in prokaryotes and eukaryotes. The secondary structures measured using CD

spectroscopy measured at pH 7.0 demonstrate that PRO^{FUR} and PRO^{PC1} adopt structures similar to PRO^{AQU} and PRO^{SUB-C} complexed to subtilisin (Figure 3-2a). Since isolated PRO^{SUB} is intrinsically unstructured [16], the PRO^{SUB-C} structure was obtained by a difference spectra between the cleaved PRO^{SUB}:Ser₂₂₁Cys-subtilisin complex and mature subtilisin as described earlier [14].

The pH-dependent structural stability of various propeptides was monitored by observing changes in negative ellipticity at 222nm as a function of pH (Figure 3-2b); as a representative example, we show the complete CD spectrum of PRO^{FUR} at the two ends of the pH range (pH 7.4 and pH 5.0) compared with a completely denatured PRO^{FUR} (Figure 3-2c). It is noteworthy that when the pH of the buffer is lowered from pH 7.4 to pH 5.0, PRO^{FUR} loses approximately 25% of its ellipticity at 222 nm when compared with the propeptide completely denatured in 8 M urea. Furthermore, changes in negative ellipticity at 222 nm as a function of pH (Figure 3-2b) suggest that the conformation of PRO^{FUR} tends to stabilize at approximately -2800deg cm² dmol⁻¹ under acidic conditions but does not reach the ellipticity of completely unfolded PRO^{FUR} (approximately -20 deg cm² dmol⁻¹). This suggests that the changes in pH do not result in complete unfolding and that PRO^{FUR} may adopt a partially folded molten-globule-like state similar to that observed using NMR spectroscopy under acidic conditions [219]. The NMR data also suggest that PRO^{PC1} and PRO^{FUR} do not aggregate in their isolated forms.

When conformational changes of the propeptides as a function of pH are compared, it is evident that PRO^{PC1} and PRO^{FUR} unfold at different pH values,

~5.5 and ~6.5, respectively (Figure 3-2B). Although the unfolding of PRO^{PC1} is not complete at pH 5.0, the structure of PRO^{PC1} at a pH below 5.0 was not analyzed because it is beyond the range of the buffering capacity of our system. While this prevents the accurate determination of the midpoint of unfolding transition in the case of PRO^{PC1}, changing buffer systems to accommodate lower pH is problematic because diverse ions can differentially influence structure, stability, and/or activity of the propeptide and protease system. Nonetheless, comparing the folding transition profiles of PRO^{FUR} and PRO^{PC1} suggests that PRO^{PC1} is more stable with regard to pH-dependent unfolding when compared with PRO^{FUR}. Under similar conditions, PRO^{SUB} and PRO^{AQU} are stable with minor changes in conformation. Due to its intrinsically unstructured state, PRO^{SUB} would not be expected to undergo conformational changes as a function of pH. However, studies have suggested that an increase in proton concentrations can induce molten-globule like states into unfolded proteins [220]–[223]. Our studies suggest that acid-induced folding is not observed in case of PRO^{SUB}. It is noteworthy that the pH-associated structural transitions PRO^{PC1} and PRO^{FUR} correlate with organelle-specific pH values necessary for activating the mature catalytic domains, MAT^{PC1} and MAT^{FUR} [33]. We next investigated whether propeptides alone are sufficient for pH-dependent activation of cognate proteases *in vitro*.

3.4.3 Swapping propeptides between PC1 and furin reassigns pH-dependent activation

To monitor *in vitro* activation of propeptide:protease inhibition complexes, we measured enzyme activity as a function of pH (see Materials and Methods).

Figure 3-2d demonstrates that $\text{PRO}^{\text{FUR}}:\text{MAT}^{\text{FUR}}$ and $\text{PRO}^{\text{PC1}}:\text{MAT}^{\text{PC1}}$ show maximum activation at pH ~6.5 and pH ~5.5, respectively, consistent with the optimal activation pH of their zymogens [33]. However, the $\text{PRO}^{\text{PC1}}:\text{MAT}^{\text{FUR}}$ complex (wherein PRO^{PC1} substitutes PRO^{FUR}) forces the catalytic domain of furin (MAT^{FUR}) to now display PC1-like activation. Similarly, replacing PRO^{PC1} with PRO^{FUR} causes the catalytic domain, MAT^{PC1} , to alter its activation to mimic furin (pH ~6.5; Figure 3-2d). Together, the CD spectroscopy, sequence/structural congruence with PRO^{SUB} , and the reassignment of activation pH by swapping PRO^{FUR} and PRO^{PC1} support the hypothesis that eukaryotic propeptides recognize and regulate pH-dependent activation of their cognate proteases *in vitro*. This finding could also be confirmed *in vivo* using constructs that trap complexes in the endoplasmic reticulum [190].

3.4.4 Histidine protonation alters conformational dynamics of eukaryotic propeptides

Based on experimental studies, we had hypothesized that the protonation of His₆₉ and potentially other histidine residues may induce conformational changes within PRO^{FUR} to mediate pH-dependent activation [189]. Moreover, although His₆₉ is conserved, PRO^{PC1} undergoes its pH-dependent activation at a much lower pH (5.0). To better understand how histidine protonation may influence propeptide conformations, we conducted MD simulations on PRO^{FUR} and PRO^{PC1} with unprotonated (pH 7) or protonated (pH 6) histidine residues, using PRO^{SUB} and PRO^{AQU} from prokaryotes as controls. MD simulations can provide information that complements biophysical and biochemical studies on

mechanisms of propeptide-mediated protease activation in eukaryotes [161]. Early MD simulations of the unfolding of reduced bovine pancreatic trypsin inhibitor on a 500 ps time scale suggest the formation of a molten-globule-like state that was compact but expanded relative to the native bovine pancreatic trypsin inhibitor (11-25%), which is consistent with experimental data [224], [225]. MD simulations have also analyzed the structure and fluctuations of “native” apomyoglobin in aqueous solution for a period of greater than 0.5 ns and have yielded a detailed model for structure and fluctuations in apomyoglobin, which complements the experimental studies [226]. Unfolding simulations using MD methods have yielded insights into the mechanism of extreme unfolding cooperativity in the kinetically stable alpha-lytic protease, a protein that exploits the mechanism of propeptide-dependent folding [227]. In these studies, the simulated alpha-lytic protease unfolding pathway produces a robust transition - state ensemble that is observed within the 10 ns simulation and is consistent with prior biochemical experiments demonstrating that unfolding proceeds through a preferential disruption of the domain interface. Furthermore, the authors demonstrate that α LP unfolds extremely cooperatively, while trypsin, a protein that folds independent of its propeptide, undergoes gradual unfolding under identical conditions of simulations. MD simulation studies have also been used to investigate the role of hydrogen bonding involving the backbone in hen egg white lysozyme, using native as well as partly and fully thionated lysozyme [228]. The results of the simulations show that the structural properties of fully thionated lysozyme clearly differ from

those of the native protein, while partly thionated lysozyme changes only slightly when compared to native lysozyme. In these studies, the extent of observed unfolding remains constant after 10 ns. Hence, in our studies, MD simulations are performed on a 10 ns time scale. We compared the similarity of structures to the starting conformation by measuring the root-mean-square deviation (RMSD) values at C_α in every residue of the propeptide domain, along equally spaced snapshots of the simulation trajectory. Simulations suggest that while PRO^{SUB} and PRO^{AQU} are stable, PRO^{PC1} and PRO^{FUR} display enhanced conformational dynamics (Figure 3-3 A and B). Our time-evolved, pH-dependent, residue-specific conformational dynamics suggest that although eukaryotic propeptides display local fluctuations at neutral pH, histidine protonation enhances overall movement and potentially exposes the compartment-specific second cleavage site loop for proteolysis in PRO^{FUR} (residues 70 to 80) when compared with PRO^{PC1}, which is more stable at pH ~6.0-7.0 (Figure 3-3A). Under identical conditions, PRO^{SUB} and PRO^{AQU} from prokaryotes display remarkable stability towards histidine protonation (Figure 3-3B). To further dissect the structural changes, we plotted the global unfolding of PRO^{FUR} and PRO^{PC1} as a function of time and at the two different pH values (Figure 3-3C). Global unfolding (Q score), which was computed using the fraction of native contacts that are retained as a function of time during the simulation at different pH values, demonstrates that PRO^{FUR} appears to undergo significant changes in the native-like contacts upon protonation of the histidine

residues. Under similar conditions, PRO^{PC1} appears to be more stable at both pH values.

Since our model for PRO^{FUR} is based on a homology model derived from the NMR structure of PRO^{PC1}, it can be argued that the model may not correspond to an energetically favorable conformation and the simulations may be biased by the homology model. To address this issue, we have performed two additional independent simulations on PRO^{FUR} and PRO^{PC1} and for a longer time scale (Figure 3-3D). To analyze the structural changes, we plotted the RMSD of the core and the secondary cleavage site loop between the initial structure and equally spaced snapshots of the trajectory of simulation, both as a function of time and at two different pH values (Figure 3-3D). While PRO^{PC1} remained stable at both pH values, PRO^{FUR} showed increasing RMSD values throughout the simulation at pH 6, while remaining stable at pH 7. The results confirm our earlier simulations on a shorter time scale and suggest that protonation/deprotonation of histidine residues play a role in the conformational destabilization of PRO^{FUR} compared to PRO^{PC1}. While our simulations do not provide information on why PRO^{PC1} is more stable than PRO^{FUR} towards pH-dependent unfolding, they corroborate our experimental observations on the pH-dependent stabilities of the propeptides. His₆₉ in furin and the corresponding His residue in PC1 reside closely to other histidine residues and charged residues in the cleavage loop (Figure 3-3E). The interaction of this protonated His with these other residues may provide key insights into why the activation pH values of furin and PC1 differ dramatically.

Together with our biophysical, biochemical, and cell-based studies, the MD simulations suggest that upon protonation of His residues, PRO^{FUR} undergoes conformational changes that may potentially destabilize the propeptide domain to expose the internal cleavage site for proteolysis. Given that PRO^{PC1} undergoes activation at pH ~5.5 in the DCSGs and remains stable upon His protonation, we can conclude that either additional residues must play a role in the activation of PRO^{PC} or the time scale of the simulations is too short to capture the unfolding event.

3.4.5 The constitutively deprotonated mimic of the pH Sensor, His₆₉Leu-PRO^{FUR}, is more stable than WT-PRO^{FUR}

To understand the mechanism by which His₆₉ functions as a pH sensor, we undertook detailed structural analyses of WT-PRO^{FUR} and the His₆₉Leu-PRO^{FUR} variant reported previously, which mimics the nonprotonated state of the pH sensor, using CD and intrinsic fluorescence spectroscopy. Prior studies indicate that the His₆₉Leu-PRO^{FUR} chaperones efficient folding of the catalytic domain of furin (MAT^{FUR}), as measured by autoprocessing of His₆₉Leu-PRO^{FUR} to form a stable His₆₉Leu-PRO^{FUR}-MAT^{FUR} complex [189]. However, unlike the WT-PRO^{FUR}-MAT^{FUR} complex, His₆₉Leu-PRO^{FUR}-MAT^{FUR} remains trapped in a stable state, unable to become active at acidic pH. On the other hand, the His₆₉Lys-PRO^{FUR} variant, which represents a constitutively protonated state of the pH sensor, fails to fold correctly and is rapidly degraded inside the cell [189]. Hence, it was not used in our analyses.

The far UV CD spectrum of the isolated WT-PRO^{FUR} revealed the existence of a significant secondary structure (Figure 3-4A). Substituting the pH sensor, His₆₉, with Leu caused a slight increase in the secondary structure, as seen by the shift in the peak from 206 to 208 nm with a concomitant increase in negative ellipticity at 222 nm in His₆₉Leu-PRO^{FUR}. Because α -helices absorb strongly at 222 nm and 208 nm [229], our results suggest that substituting residues that mimic the deprotonated state of the pH sensor marginally increase the α -helicity within the isolated His₆₉Leu-PRO^{FUR}.

Likewise, we examined the tertiary structure of the protein by exciting the protein using a wavelength of 295 nm where the tryptophan emission spectrum is dominant over the weaker tyrosine and phenylalanine fluorescence [230]–[232]. As seen in Figure 3-4B, WT-PRO^{FUR} displays a maximum peak at 342 nm. Under identical conditions, the intrinsic tryptophan fluorescence is enhanced slightly with a blue shift in its emission spectrum (maximum at 339 nm) when His₆₉ is substituted by leucine. This indicates that the tryptophan residues are less exposed to solvent when His₆₉ is replaced by a leucine, suggesting that the structure may be more packed.

To better understand the extent of stabilization, we next measured the thermodynamic stability of WT-PRO^{FUR} and its variant relative to their unfolded states. Thermodynamic stability occurs when a system is in its lowest energy state when compared with all other accessible states within the same reaction environment. It can be measured by monitoring changes in the secondary structure with progressive addition of chaotropes such as urea or guanidine

hydrochloride [233]. Figure 3-4C compares chaotrope-induced conformational changes in WT-PRO^{FUR} or His₆₉Leu-PRO^{FUR} using circular dichroism spectroscopy. The transitions were fitted using a standard Marquardt algorithm with constraints for the base line set from using the circular dichroism ellipticity of the folded and unfolded proteins [17], [212]. The data demonstrate that His₆₉Leu-PRO^{FUR} ($\Delta G_{\text{NU}} = 1.424 \pm 0.12$ kcal) is more stable than WT-PRO^{FUR} ($\Delta G_{\text{NU}} = 0.921 \pm 0.09$ kcal/mol). This indicates that the constitutively deprotonated variant His₆₉Leu -PRO^{FUR} is stabilized by ~0.5 kcal/mol when compared with WT-PRO^{FUR}.

Because propeptides are bona fide temporary inhibitors of proteases [217], we next asked how an increase in thermodynamic stability affects the inhibitory function of isolated WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR} by comparing IC₅₀ values as described under “Experimental Procedures” (Figure 3-4D). Analysis of the data gives an estimated IC₅₀ concentration for WT-PRO^{FUR}, at pH 6.5, of ~33 nM, 3-fold higher than that estimated for His₆₉Leu-PRO^{FUR} (IC₅₀ ~11 nM). This establishes a link between the increased thermodynamic stability of the His₆₉Leu substitution and its ability to act as an inhibitor of MAT^{FUR}, as indicated by the decrease in IC₅₀. Taken together, the circular dichroism and fluorescence spectra, along with the analyses of thermodynamic stabilities, suggest that the nonprotonated mimic of the pH sensor subtly increases both secondary and tertiary structure, and enhances the overall thermodynamic stability and inhibitory function of His₆₉Leu-PRO^{FUR}.

3.4.6 His₆₉Leu-PRO^{FUR} is more stable toward pH-dependent unfolding

Because pro-furin undergoes its primary cleavage in the neutral environment of the ER to form a cleaved, non-covalently associated PRO^{FUR}-MAT^{FUR} complex that transits in to the mildly acidic TGN to become active, we next examined how changes in pH affect the structure, stability, and binding affinity of WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR}. Propeptides were purified, refolded, and analyzed for their secondary structure content using CD spectroscopy as described (“Materials and Methods”). The results show that WT-PRO^{FUR} undergoes pH-dependent unfolding with an isosbestic point at ~208 nm (Figure 3-5A). A plot of the changes in CD signal at 222 nm as a function of pH suggests that WT-PRO^{FUR} undergoes a cooperative sigmoidal transition to a more unstructured state. Interestingly, the midpoint of this transition occurs at pH ~6.0, close to the optimal pH for activation of furin. Under identical conditions, His₆₉Leu-PRO^{FUR} is more stable. Although it does undergo some pH-dependent unfolding, with a midpoint of transition, likewise, at pH ~6.0, it is critical to note that not only it is more stable at neutral pH than the WT, but also that it is not unfolded to the same extent, suggesting that the unfolding response to pH is blunted. In comparison, the change in structure of isolated MAT^{FUR} across this pH range is not significant (Figure 3-5C, colored lines), nor does the addition of glycerol markedly change the structure of MAT^{FUR} (gray line). Taken together, this suggests that the protonation status of His₆₉ may drive pH-dependent conformational changes in the isolated furin propeptide. Our results indicate that lowering the pH triggers a transition between a folded state at pH 7.4 and a less

folded, but not completely unstructured, state at pH 5.0. Mutations of titratable group His₆₉ to leucine marginally increase secondary structure at pH 7.4 and, to a larger extent, at pH 5.0 (Figure 3-5B), suggesting that protonation of His₆₉ is essential for the pH-dependent transition between the two states.

Next we measured the changes in thermodynamic stability of the WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR}, as described earlier (Figure 3-5C), under conditions of varied pH. Our data suggest that the overall thermodynamic stability of the proteins decreases when the pH becomes more acidic (Figure 3-5D). The greater change in thermodynamic stability of His₆₉Leu-PRO^{FUR} as a function of pH suggested that the His₆₉Leu substitution enhanced the thermodynamic stability in the isolated propeptide when compared with WT-PRO^{FUR}.

Because the concentration of protons affects the conformation of WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR}, we next measured how this conformational change affects the IC₅₀ values as a function of pH (Figure 3-5D). The data demonstrate that the IC₅₀ values for WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR} change as a function of pH, with the maximum inhibitory concentration required for both proteins at pH 6.0 (Figure 3-5D) and the midpoint of the conformational transition being ascertained using CD spectroscopy (Figure 3-5B). Moreover, three important features in Figure 3-5D are noteworthy. 1) The IC₅₀ value for WT-PRO^{FUR} at pH 7.4 (~12 nM) is about 4-fold higher than at pH 6.0 (IC₅₀ ~50 nM); 2) the IC₅₀ value for His₆₉Leu-PRO^{FUR} at pH 6.0 (~17 nM) is about 3-fold lower than that for WT-PRO^{FUR} (IC₅₀ ~50 nM); and 3) when the pH is lower than the optimum for activation (pH ~6.0), the IC₅₀ value drops to lower concentrations of propeptides

for both WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR}, suggesting an apparent increase in binding affinity. To examine whether pH denatures or inactivates MAT^{FUR}, we also monitored changes in the secondary structure (Figure 3-5C) and the activity of furin across this pH range with no propeptide present (Figure 3-5F). It is worth noting that although the activity of furin does decrease as pH drops, it remains active, with an activity at pH 5.0 roughly 50% of that observed at pH 6.0, where IC₅₀ is highest. This suggests that furin remains structurally stable, indicating that increased affinity at pH 5.0 is likely a chemical phenomenon. Hence, the change in IC₅₀, which is roughly 40-fold lower at pH 5.0 than at pH 6.0, cannot be explained by changes in activity alone. We are currently unable to examine directly how pH affects the propeptide-furin complex because of the high concentrations of mature furin required to create stoichiometric complexes.

3.4.7 Molecular Dynamics of WT and His₆₉Leu PRO^{FUR}:

We hypothesized that introduction of the His₆₉Leu into our model should stabilize the structure. To test this hypothesis, we first compared the root mean square fluctuations (RMSF) values of WT-PRO^{FUR} at pH 7.0 and pH 6.0 (Figure 3-6A). Although histidine protonation increased fluctuations at almost all residues, the highest increase was observed in the loop flanked by β 2 and β 3 in proximity of residue 61 and within the C-terminal half of the cleavage loop. The largest increase was observed for His₈₀, whereas His₆₉ did not show any change upon protonation and remained stable. The His₆₉Leu substitution reduced conformational fluctuations during the simulation at pH 7, with an even greater

stabilization observed at pH 6. Interestingly, in His₆₉Leu-PRO^{FUR}, His₈₀ appears to be the most stabilized compared with residues at pH 6.

Analysis of the root mean square deviation values compared with the starting structure (Figure 3-6B) as well as ribbon representation of the starting and end structures (Figure 3-6C) revealed that during the simulation, the core region remains largely stable at pH 7.0 in the WT and His₆₉Leu-PRO^{FUR}. It is important to note that the loop regions, which had a very high root mean square deviation during the simulation of WT-PRO^{FUR}, were stabilized substantially by the His₆₉Leu variant. At pH 6, the core domain of the WT lost its native structure, indicated by rising root mean square deviation values. Compared with WT-PRO^{FUR}, the His₆₉Leu variant stabilized the core and loop region significantly, although a slight increase in root mean square deviation was still observed.

Hence, our MD simulations suggest that although the loop region shows a high degree of flexibility during simulations using both protonated and unprotonated histidines in WT-PRO^{FUR}, the core remains stable in simulations using unprotonated histidines but loses structure in simulations using protonated histidines. Introduction of the His₆₉Leu substitution into our model greatly increased stability of the core region during our simulation using protonated histidines, confirming that protonation of His₆₉ alone plays an important role in the pH-mediated structural changes. Interestingly, His₆₉Leu mutations also lead to an increase in stability in the loop regions that stayed buried during the simulation. Taken together, the MD simulations are consistent with our hypothesis that changes in physical properties of the side-chain at the

pH sensor position has a strong influence on the structure of the activation loop and also suggest that leucine, because of its greater hydrophobicity, may not be a faithful representation of unprotonated histidine.

3.4.8 Protonation of histidine residues in addition to His₆₉ are involved in pH-sensing

Since the His₆₉Leu variant still displays pH-dependent conformational changes (Figure 3-5) we asked whether the conformational change in WT PRO^{FUR} is caused by two protonation events. To do this we used Near-UV circular dichroism, which characterizes the structural environment of tyrosine and tryptophan residues. PRO^{FUR} displays strikingly different Near-UV CD spectra at low and high pH (Figure 3-7A). Titration from pH 8.0 to 4.0 using 200 mM HCl showed a sigmoidal transition of the Near-UV signal at 285nm, that was perfectly reversible if the same solution was titrated back to pH 8.0 using 200 mM NaOH (Figure 3-7B). This indicates that the pH-dependent conformational transition is indeed thermodynamically controlled and that pH and not ionic strength is modulating PRO^{FUR} structure. A fit of equation 10 derived in Chapter 1.2.2.2 resulted in an apparent pK_a of 6.32±0.04 and an exponent n of 1.6±0.2. This indicates that more than one titratable site is involved in the structural transition, as one would expect n of 1 for titration of a single site. The same experiment performed for the His₆₉Leu variant resulted in an apparent pK_a of 6.05±0.07 and an exponent n of 0.9±0.2 (Figure 3-7C). This again is consistent with an additional titrating residue that is coupled to the structure of PRO^{FUR}.

To examine the possibility that another histidine residue is the additional titrating site, individual histidine residues in PRO^{FUR} were mutated to leucine or arginine to mimic the protonated and unprotonated form of histidine, respectively. Secondary structures of the various variants were then analyzed using far-UV circular dichroism with the wild-type PRO^{FUR} as a control (Figure 3-8). If protonation of a specific histidine residue contributes to the loss of secondary structure in PRO^{FUR} at low pH, we would expect that mutation of this histidine to an arginine may lead to loss of secondary structure at neutral pH, while mutation to leucine may reduce the loss of secondary structure at acidic pH. Consistent with earlier result, the His₆₉Leu variant stabilizes PRO^{FUR} against unfolding under conditions of acidic pH, although the H₆₉R variant does not substantially destabilize the structure of PRO^{FUR} at neutral pH. It is noteworthy that His₅₂Arg leads to a loss of secondary structure at neutral pH, while a His₅₂Leu variant stabilizes PRO^{FUR} at acidic pH, indicating a role of His₅₂ in pH-mediated structural changes. Surprisingly, both His₆₆ variants strongly stabilize the secondary structure of PRO^{FUR} at acidic pH, while having no substantial effect at neutral pH. Substitution of leucine and arginine at His₈₀ and His₈₄ display no substantial effect on secondary structure loss at both acidic and basic conditions, suggesting that these histidine residues do not augment the sensitivity of the pH sensor. In summary these data suggest that both His₅₂ and His₆₆ might contribute to the pH-mediated structural changes of PRO^{FUR}.

3.5 Discussion

3.5.1 Propeptides are sufficient to impart pH-sensing

Our work provides insight as to why nature may have imposed differential selective constraints that alter both sequence and the asymmetrical distribution of histidine residues in two functional domains, namely, the propeptides and their cognate catalytic domains within furin and PC1. In this article, we demonstrate that PRO^{FUR} and PRO^{PC1} are enriched in histidine content when compared with cognate proteases and prokaryotic orthologs (Figure 3-1C), consistent with the general enrichment of histidine in eukaryotic propeptides described in chapter 2. Spectroscopic studies demonstrate that changes in pH can induce conformational changes only within PRO^{FUR} and PRO^{PC1}, while their prokaryotic orthologs, PRO^{SUB} and PRO^{AQU}, are largely unaffected (Figure 3-2A and B). Since swapping propeptides between eukaryotic paralogs transfers pH-dependent protease activation in a propeptide-dictated manner (Figure 3-2D), while allowing folding and cellular localization [191], our results argue that PRO^{FUR} and PRO^{PC1} may have evolved from prokaryotic orthologs to encode histidine-driven pH sensors that enable furin and PC1 to recognize and adapt to cellular organelles. Our MD simulations suggest that histidine protonation may be sufficient to induce conformational changes that enable the second activating cleavage of the propeptide and are consistent with our spectroscopic analysis. While it would be interesting to compare the structures of the chimeras with those of the wild-type complexes and examine how their structures are affected

by changes in pH, such experimentation is currently unfeasible due to the high concentrations of protein required for CD spectroscopic analysis.

It is important to note that despite histidine enrichment, the specific location of these residues within the amino acid sequences of propeptides can vary significantly (Figure 3-1A and Figure 3-3E). Moreover, the His₆₉ that was identified as a primary pH sensor in PRO^{FUR} [59] is also conserved in PRO^{PC1}, although the pH-dependent activation of furin and PC1 differs significantly [33]. This suggests that additional undetermined residues and/or cellular factors must play a significant role in pH-dependent activation of their cognate protease domains. Propeptides also contain several charged residues [217] that may interact with protonated and non-protonated histidine residues, thereby enabling subtleties in their sensitivity to compartment-specific pH. Hence, our studies emphasize the necessity of more detailed analyses of the differences between pH sensors of PRO^{FUR} and PRO^{PC1} using detailed site-directed mutagenesis studies, to tease out the interplay with residues in the proximity of their cognate pH sensors.

3.5.2 Mechanism of pH-dependent activation of furin

There are at least three possible mechanisms (Figure 3-9) through which the stoichiometric inhibition complex (PRO^{FUR}-MAT^{FUR}) can become active upon reaching the TGN. The first mechanism posits that protonation of the pH sensor, along with other histidine residues, can induce dissociation of PRO^{FUR} from MAT^{FUR}, which then triggers the second proteolytic cleavage [215], [234]. The second mechanism postulates that pH causes partial unfolding of the propeptide,

which enables the second cleavage site to access the active site, promoting proteolysis in a cis reaction. The third mechanism hypothesizes that protonation of the pH sensor induces conformational changes that allow a mature furin molecule to access the loop that harbors the second internal cleavage site in trans [2], [59]. In these cases, subsequent cleavage facilitates propeptide dissociation.

To test these possibilities, we undertook various *in vitro* experiments. Our results are most consistent with the hypothesis that the cleavage loop is critical to the overall structure and stability of the propeptide for the following reasons.

The data demonstrate that a 25-fold increase in proton concentration observed between pH 7.5–6.0 causes a ~3.5-fold change in apparent binding affinity for WT-PRO^{FUR}. Moreover, the His⁶⁹Leu pH sensor variant affects the apparent binding affinity ~3.5-fold. To further test this, we built a mathematical model that assumed that protonation of the pH sensor promoted dissociation (data not shown here, but in [235]), allowing it to be processed. By varying K_a , we determined the effect of varying the affinity between the propeptide and protease on the rate of activation, which is a stochastic process in bacterial subtilisin [17]. Our simulation results demonstrate that ~10-fold changes in the affinity between PRO^{FUR} and MAT^{FUR} have only a minimal effect on the rate of activation within physiologically relevant range of values. It is noteworthy that our experimental data establish that the change in affinity because of the substitution is ~3-fold, which argues that dissociation alone cannot account for lack of activation of the His⁶⁹Leu-PRO^{FUR}-MAT^{FUR} complex observed in cell-based studies.

MD simulations suggest that His₆₉ protonation affects furin activation by increasing the conformational dynamics of the cleavage loop. At acidic pH, histidine residues within WT-PRO^{FUR} are protonated, including the pH sensor, His₆₉. However, when His₆₉ is replaced by Leu, the conformational dynamics of the loop are reduced dramatically, despite all of the remaining histidine residues being protonated (Figure 3-6). Hence, the protonation status of His₆₉ alone affects the dynamics of the activation loop of the wild-type and mutant propeptides at two different pHs. Our simulations suggest that at pH 6.0, where the imidazole side-chain of histidine is protonated, a dramatic movement in the loop region of PRO^{FUR} precedes the overall unfolding of the propeptide domain. This movement is reduced substantially in the case of the His₆₉Leu variant, resulting in diminished unfolding of the propeptide as seen in Figure 3-6.

It is important to note that His₆₉ is solvent-accessible yet abuts a pocket formed partly by the hydrophobic core residues. Above the pH optimum of activation (pH > 7.0), the packing of the deprotonated His₆₉ into the core maintains a well packed structure that favors the bound state. However, upon protonation, the imidazole ring becomes charged, disrupting the packing and resulting in destabilization and local unfolding that exposes the cleavage site.

Hence, the results of our experimental and simulated data indicate that structural changes alter the accessibility of the cleavage site, thus raising the question of how the cleavage site becomes available to the active site. Although we cannot definitely distinguish between the possibility that the loop movement simply moves the cleavage site into a position more accessible to the active site in

cis, or that there is a larger destabilization of the packing of the hydrophobic core that allows processing by a second molecule of furin in trans (Figure 3-9), findings reported previously may lend some insight. In earlier work we observed that when the pH sensor, His₆₉, was mutated to leucine, no activation of furin takes place under basal conditions [59]. Experiments where excess active furin was added to the inhibition complex indicated that at a non-permissive pH, exogenous furin was unable to affect activation of the inhibited furin. This suggests that the cleavage loop is inaccessible to free furin molecules at a pH outside of its optimum. Therefore, we argue that activation is mediated by proteolysis permitted by movement in the cleavage loop that only occurs upon protonation of the pH sensor and that dissociation occurs subsequently to processing.

Given this model, it is interesting to consider the possibility that activation is not concomitant with processing but, rather, that the C-terminal part of the propeptide sits in the substrate binding pocket and likely remains bound there for a period of time before it too dissociates to release inhibition. This is consistent with studies that demonstrate that the C-terminal propeptides fragments are potent inhibitors of furin [236], [237]. We do not know whether, upon cleavage, there is a change in affinity or structure or if another protease plays a role in the dissociation [238]. We can speculate that after the propeptide is cleaved at Arg₇₅, the shorter peptide fragment that lies in the substrate binding pocket is simply too short to make efficient contacts with residues of the protease and dissociates or, alternatively, that the cleavage allows a structural change to

take place that promotes dissociation. A final alternative possibility is that the peptide fragment then becomes a substrate for cleavage in trans by another protease, such as carboxypeptidase [239]. Although we cannot yet distinguish between these possibilities, future work will undoubtedly shed further light on this step of activation.

3.5.3 Implications of the pH Sensor in the activation of proprotein convertases

The data presented here suggest an overarching model for activation of the PCs using furin as an example. Upon entering its window of activation, His₆₉, the pH-sensor in furin, is protonated to destabilize the hydrophobic pocket in which the pH sensor sits. Destabilization pushes the cleavage loop outward, thus allowing the catalytic site access to the secondary cleavage site. At the same time, a certain amount of flexibility in the association of the propeptide with the protease domain is preserved, which allows this cleavage to act as a finely tuned trigger.

In this work, we have employed biophysical, biochemical, and computational approaches to investigate pH-dependent activation of furin, the canonical proprotein convertase. Our results provide insight into the way in which mature furin recognizes and responds to the changing pH of the secretory pathway and allow us to propose a mechanism for regulated activation. We have demonstrated that the propeptides of furin and PC1 alone contain information necessary for their compartment-specific activation. However, the residue that corresponds to His₆₉ in furin is conserved within all PCs, suggesting that additional factors must augment the subtle differences between the pH optima of

individual PCs. This may be in part mediated by the distribution of additional histidines, as we demonstrate that in the case of PRO^{FUR} additional histidine contribute to pH-dependent conformational changes (Figure 3-7 and Figure 3-8), and other charged residues within the propeptides of PCs.

Figure 3-1: Comparison of sequences, structures, evolution, and composition biases of propeptides in prokaryotic and eukaryotic subtilases.

The pink and gray background in (a) through (d) indicates prokaryotes and eukaryotes, respectively. (a) Multiple sequence alignment (MSA) displaying conservation between eukaryotic subtilases and prokaryotic orthologs. Numbering is based on furin. Residues shaded black are 100% conserved, those shaded dark gray are > 80% conserved, and those shaded light gray are > 50% conserved. The conserved pH sensor in furin is shaded green and the secondary cleavage loop is indicated by the red box. Red X's represent an insertion of five residues in aqualysin. Pink shading represents prokaryotes while the light gray represents eukaryotes. Secondary structures displayed below MSA are based on PRO^{PC1} (1KN6). Motifs N1 and N2 depict folding nucleation sites for MAT^{SUB}. (b) Structures of propeptides displayed as ribbon diagrams. PRO^{SUB} structure was extracted from the propeptide:subtilisin (1SCJ), while PRO^{AQU} structure is a homology model based on 1SCJ and 2W2M. The structure of PRO^{PC1} is derived from the NMR (1KN6) while PRO^{FUR} represents a homology model of PRO^{PC1}. (c) Heat map displaying amino acid content within the propeptides and catalytic domains of prokaryotic subtilisin and aqualysin and eukaryotic PCs, furin and PC1. Protein sequences for furin (n = 26), PC1 (n = 14), subtilisin (n = 69), and aqualysin (n = 7) families were obtained from the 50% sequence identity clusters UniRef50_P09958, UniRef50_P29120, UniRef50_P00782, and UniRef50_P08594 in the UniRef database, respectively. Amino acid content for each family of propeptides and protease domains was calculated and averaged. Contents of amino acids belonging to individual groups were added and divided by the sum of their content in the whole UniProt database (release 2011_12) to obtain the fold change. Within an individual group, fold values greater than 1 indicate residue enrichment, values less than 1 indicate residue depletion, while a value of 1 indicates no change.

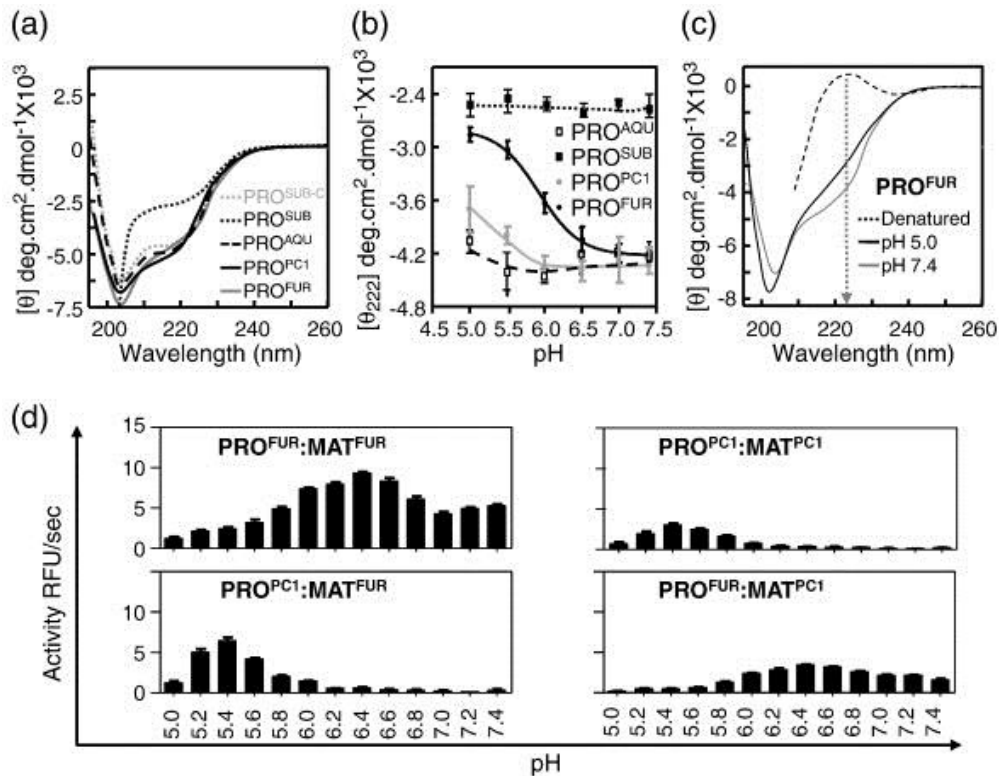


Figure 3-2: pH-dependent structure and function of propeptides.

(a) Secondary structures determined using CD spectroscopy performed at pH 7.0 and plotted as molar ellipticity $[\theta]$ deg cm² dmol⁻¹. (b) Structural stability of propeptides monitored by changes in ellipticity at 222 nm as a function of pH. (c) The secondary structure of PRO^{FUR} at pH 7.4 and 5.0, compared with completely denatured furin. The arrow marks 222 nm on the scale. (d) Type of eukaryotic propeptide dictates pH optimum for activation of the propeptide:protease complex. The activation optimum for MAT^{FUR} shifts from pH ~ 6.5 in the presence of PRO^{FUR} to pH ~ 5.5 when PRO^{PC1} forms the complex. Conversely, MAT^{PC1} activation shifts from pH ~ 5.5 in the presence of PRO^{PC1} to pH ~ 6.5 when PRO^{FUR} forms its complex. Values are measurements of three independent experiments.

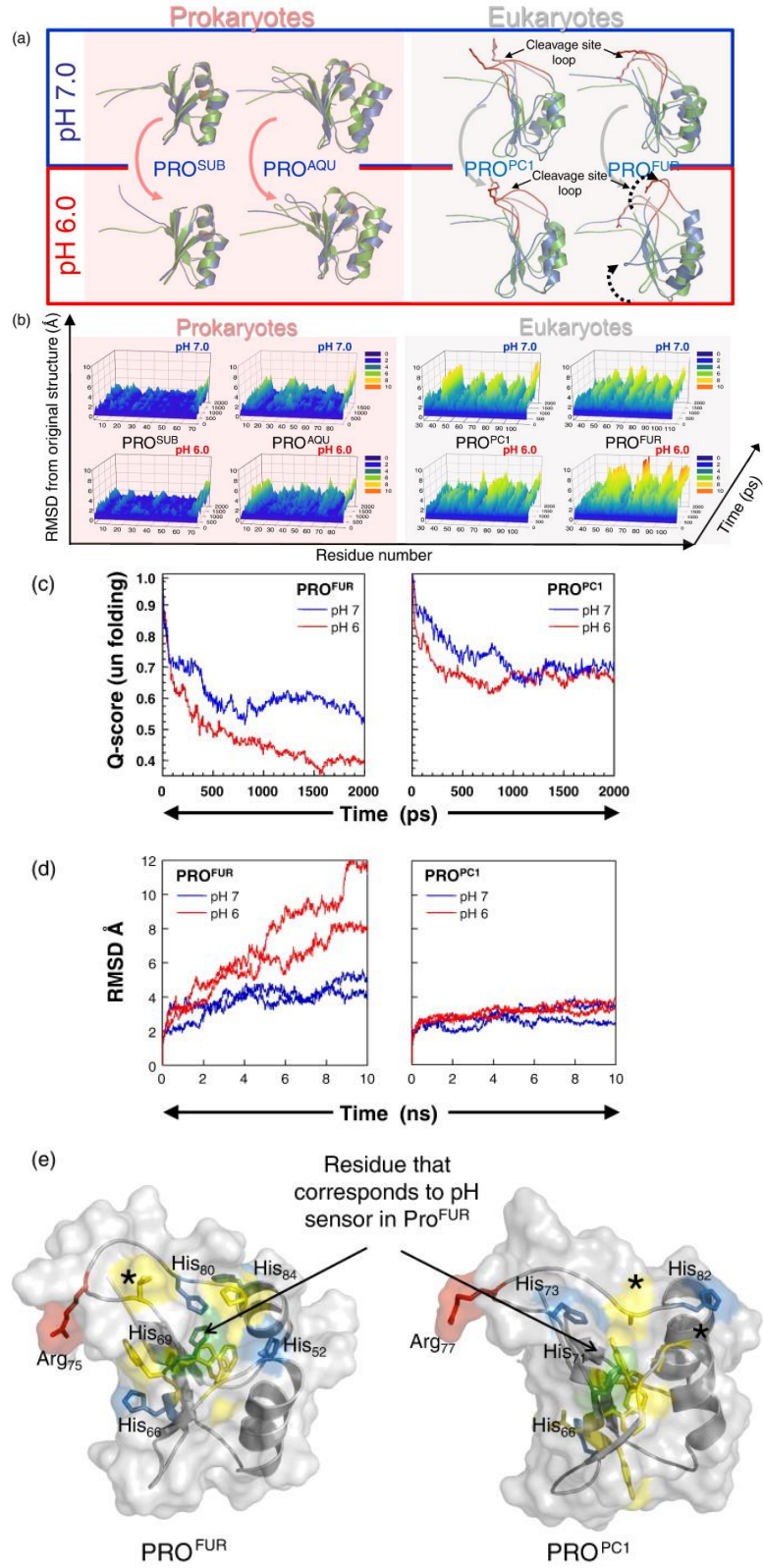


Figure 3-3: pH-dependent structural dynamics of prokaryotic and eukaryotic propeptides.

(a) Green and blue cartoons represent initial and final structures of the simulations, respectively. The second cleavage site loop (red/salmon) in PRO^{FUR} (structure on the right side) is stable when histidines are deprotonated (pH 7.0; bordered by black box) but changes conformation upon histidine protonation (pH 6.0; bordered by red box). The dynamics of the loop are unaffected by the histidine protonation status of PRO^{PC1} (cartoons on the left side). Under identical conditions, PRO^{SUB} and PRO^{AQU} show insignificant changes in dynamics as a function of pH. (b) Protonation - status - dependent, time-resolved, residue-specific dynamics of PRO^{SUB}, PRO^{AQU}, PRO^{PC1}, and PRO^{FUR}. Arrowhead indicates secondary cleavage site and color scale represents RMSD from initial structures. (c) Global unfolding (Q score) of PRO^{FUR} and PRO^{PC1} at different pH values. Unfolding was computed using the fraction of native contacts that are retained as a function of time during the simulation at different pH values and suggests that PRO^{FUR} undergoes global unfolding at a pH of 6.0 when compared with pH 7.0 and with PRO^{PC1} at both pH 7.0 and 6.0, respectively. (d) Evaluating the robustness of independent MD simulations using different models and longer time scales. We compared the similarity of structures to the starting conformation by measuring the RMSD within the propeptide domain, along equally spaced snapshots of the simulation trajectory. Our results suggest that while PRO^{PC1} appears stable at different pH values, PRO^{FUR} displays significantly larger conformational changes, which may contribute to its increased proteolytic susceptibility at pH 6.0, and is consistent with our spectroscopic studies. (e) A comparison of the structural locations of various histidine residues in PRO^{FUR} and PRO^{PC1}. The pH sensor His69 in PRO^{FUR} (green) along with other histidine residues (blue) and their corresponding residues with PRO^{PC1} are depicted. Hydrophobic residues surrounding His69 in PRO^{FUR} are depicted in yellow, while the asterisks denote residue substitutions at cognate histidine residues.

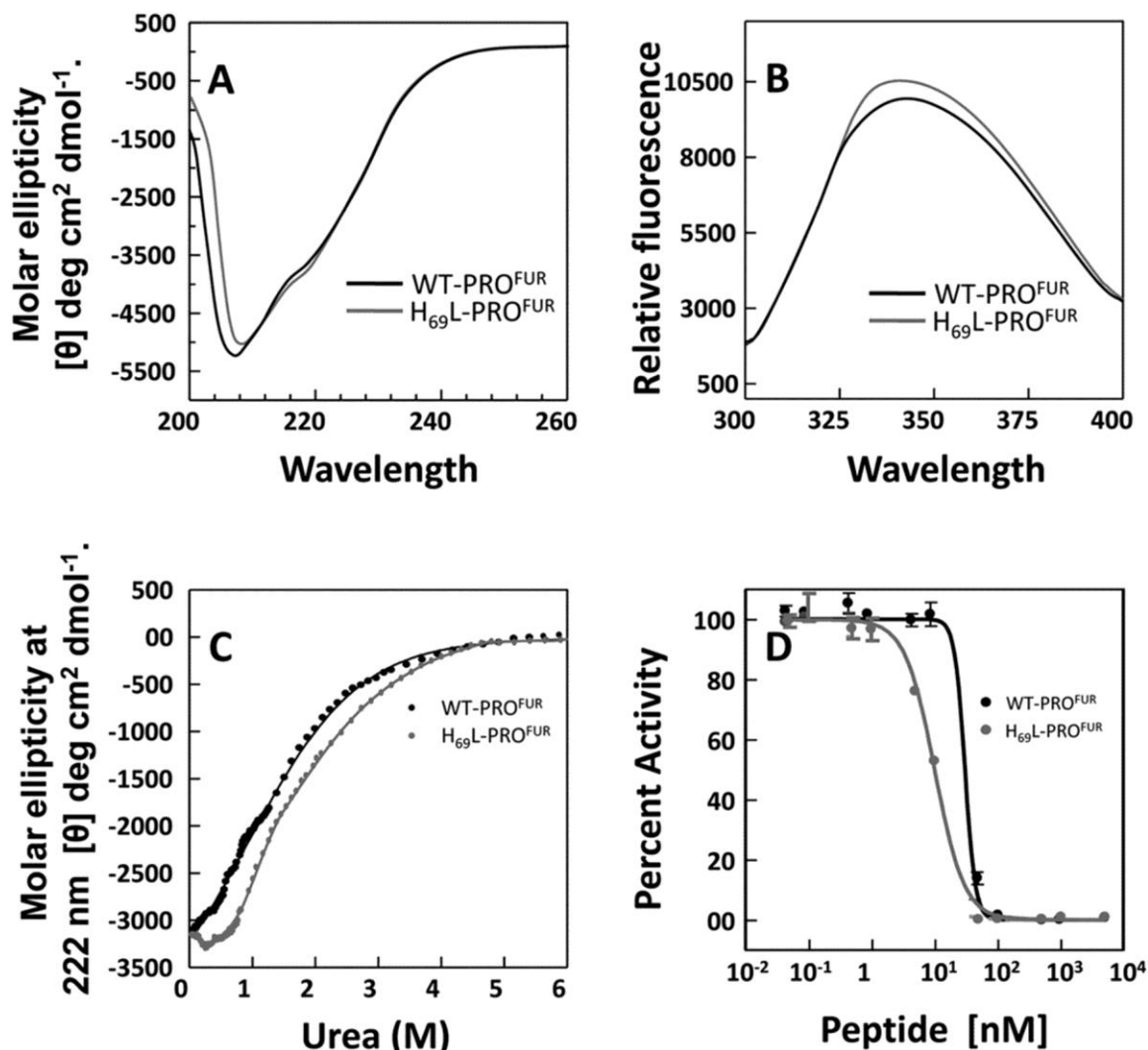


Figure 3-4: His₆₉Leu-PRO^{FUR} is more structured than WT-PRO^{FUR}.

A, the secondary structure of WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR} determined via CD spectroscopy at far UV, performed at pH 7.0, and plotted as molar ellipticity (θ) deg/cm²/dmol. B, tertiary structure of wild-type or mutant propeptide determined by measuring intrinsic tryptophan fluorescence after excitation with $\lambda = 295$ nm. C, thermodynamic stability of the propeptides monitored by changes in ellipticity (θ) at $\lambda = 222$ nm as a function of urea concentration. Data were fit to a standard three-state equation using a Marquardt algorithm. D, normalized activity, used to estimate IC₅₀ values, determined by monitoring cleavage of the fluorogenic peptide substrate Abz-RVKRGLA-Tyr[2-NO₂] with increasing amounts of WT-PRO^{FUR} or His₆₉Leu-PRO^{FUR} present. All data are averaged over three independent experiments.

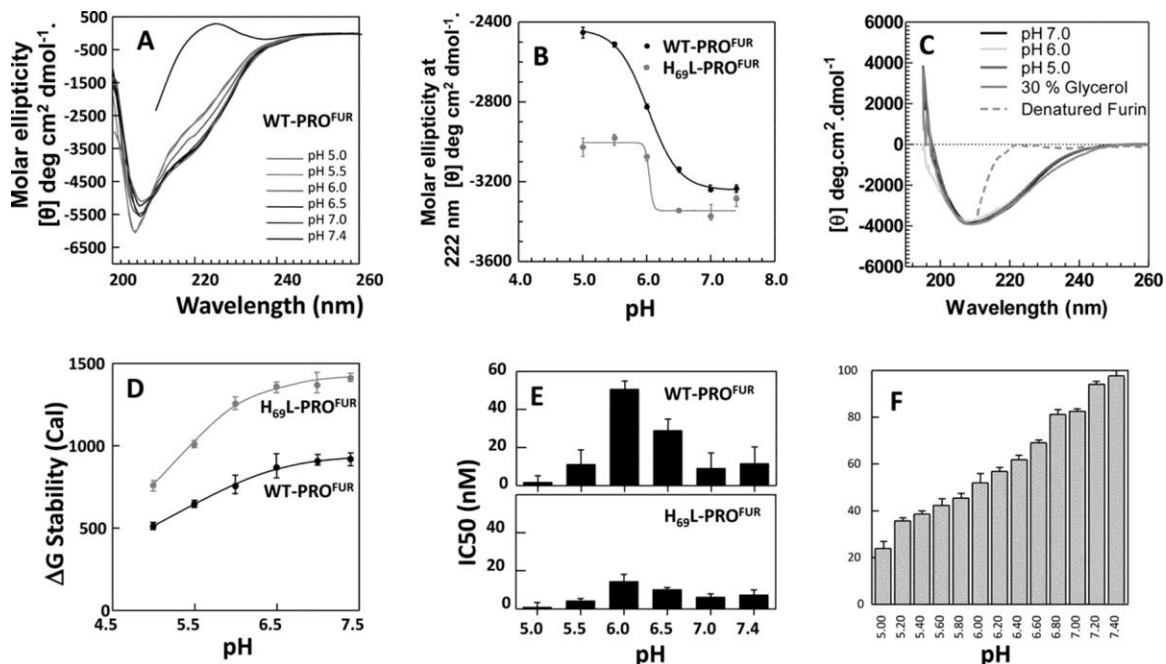


Figure 3-5: His₆₉Leu-PRO^{FUR} is more stable than WT-PRO^{FUR} to pH-induced unfolding.

A, pH-dependent secondary structure of WT-PRO^{FUR} performed at pH 7.0–5.0 and plotted as molar ellipticity. B, changes in secondary structure of the isolated propeptides monitored by changes in ellipticity at $\lambda = 222$ nm and plotted as a function of increasing pH. The midpoint of the unfolding transition for both peptides occurs at pH ~ 6.0 . C, CD structure of MAT^{FUR} at varying pH (colored lines) and with the addition of 30% glycerol (gray line). The dotted line represents the spectra of denatured MAT^{FUR}. D, thermodynamic stability of WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR} as a function of pH, given in calories (Cal). E, IC₅₀ values for WT-PRO^{FUR} (top panel) and His₆₉Leu-PRO^{FUR} (bottom panel) as a function of pH. F, activity of furin in the absence of the propeptide at varying pH. All values are given as a percentage of maximum activity and are the average of three independent experiments.

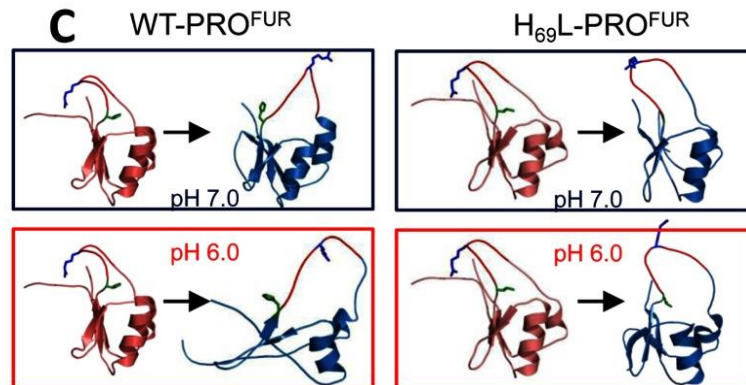
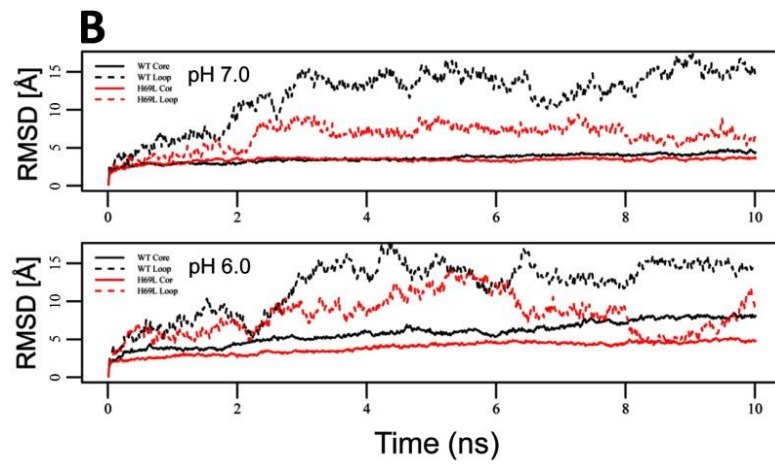
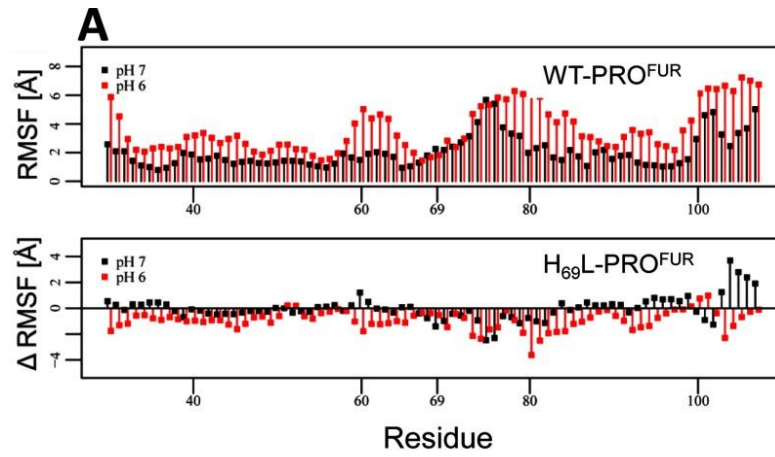


Figure 3-6 Molecular dynamics simulations of His₆₉Leu PRO^{FUR}

MD simulations were performed on WT-PRO^{FUR} and His₆₉Leu-PRO^{FUR} using NAMD as described under "Materials and Methods." A, top panel, RMSF as a function of residue number for WT-PRO^{FUR}. Bottom panel, differences between the RMSF of simulation with His₆₉Leu-PRO^{FUR} and the WT (Δ RMSF). Negative values indicate reduced fluctuations, and positive values indicate increased fluctuations because of the His₆₉Leu point mutation. Values obtained under a simulated pH of 7 are shown in black, whereas values obtained under simulated pH of 6 are shown in red. B, RMSF values for the core region (all except for the loop) are shown by dashed lines, and loop regions (residues 70–79) are depicted by solid lines and plotted as a function of simulation time. Black lines represent WT-PRO^{FUR}, and red lines represent His₆₉Leu-PRO^{FUR}. C, ribbon representation of the starting (red) and final (blue) structures of the simulations. The secondary cleavage site, Arg₇₅, is indicated in the cleavage loop (blue), and the pH sensor, His₆₉, is indicated in green. All simulations were done over 10 ns.

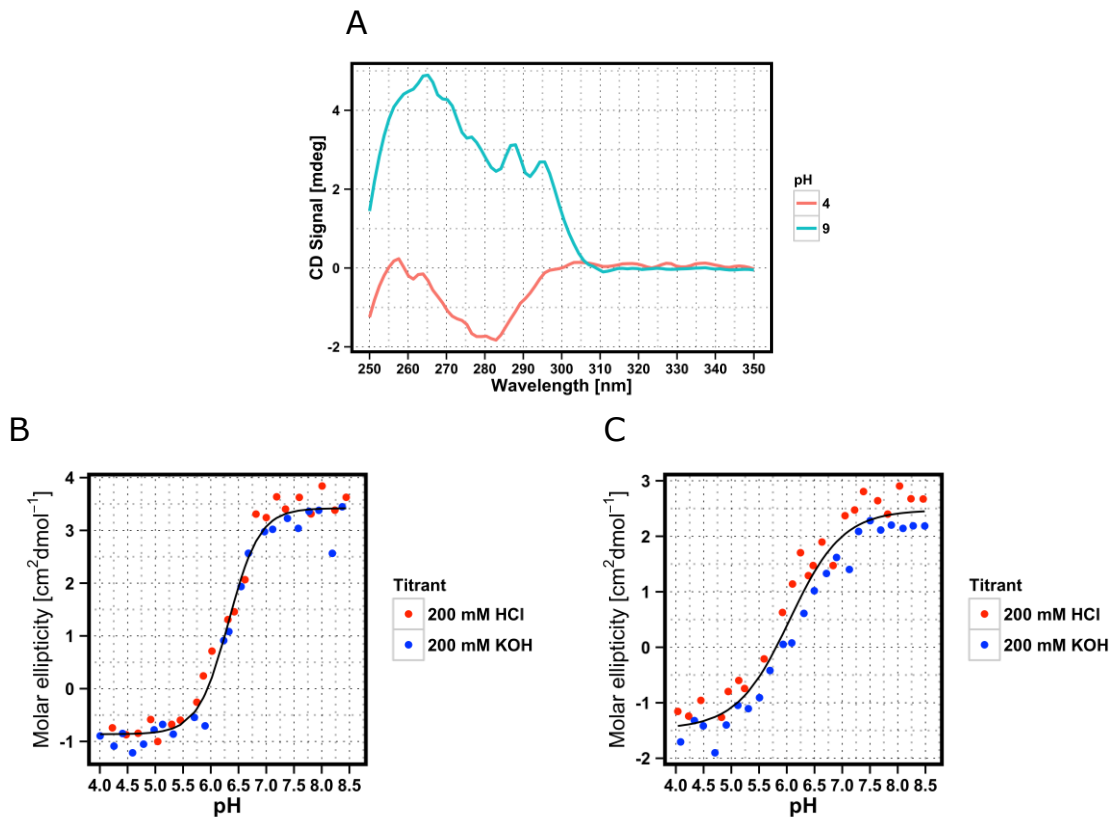


Figure 3-7: Near-UV titration of WT and His₆₉Leu PRO^{FUR}

(A) Near-UV circular dichroism spectra of WT PRO^{FUR} at pH 4 or pH 9.
 (B-C) Titration of WT PRO^{FUR} (B) and His₆₉Leu PRO^{FUR} (C). Structural transition is monitored by circular dichroism at 285 nm. Titration started at pH 8.5 and was performed by injecting small amounts of acid to elicit an about 0.2 pH unit shift. After the solution reached the titration solution was changed to base and titration was performed back to pH 8.5. Data is fitted using equation 10 in chapter 1.2.2.2

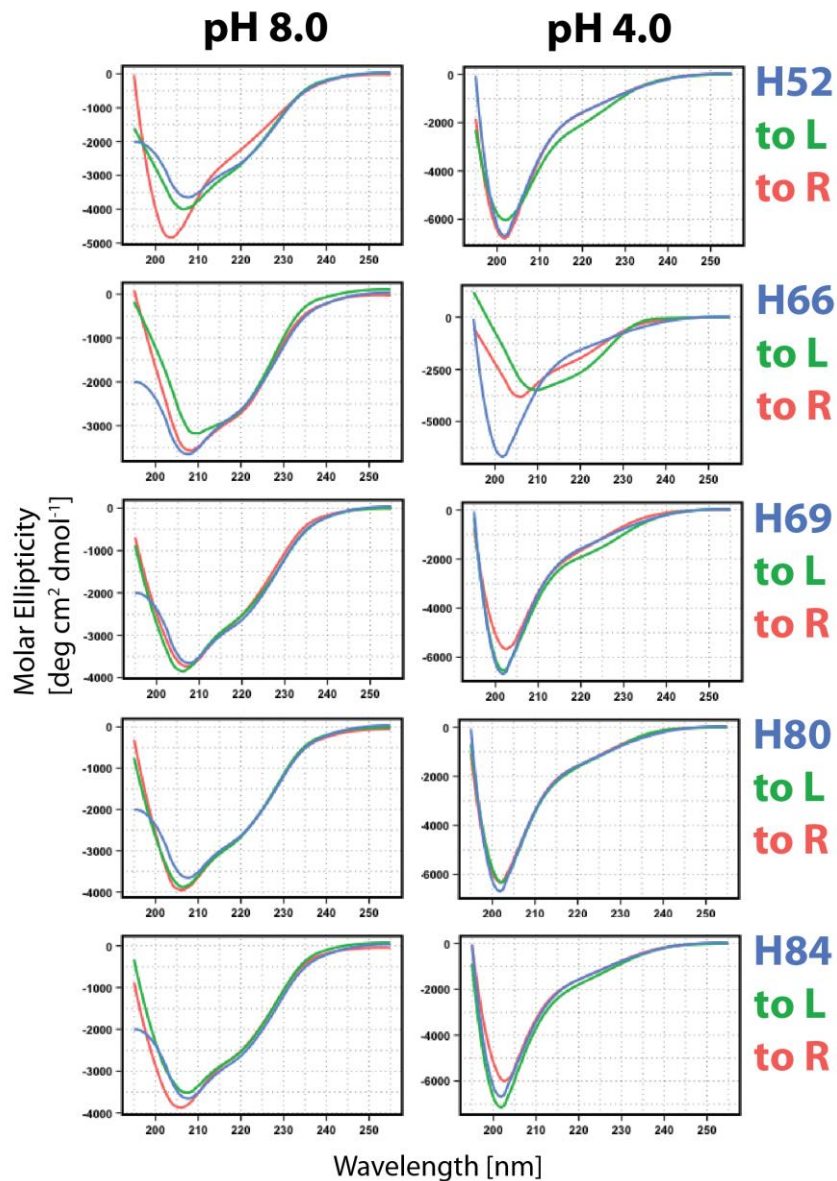


Figure 3-8: Far-UV circular dichroism spectra of single-point variants of PRO^{FUR} .

Every histidine was mutated to either leucine or arginine. Spectra were obtained at pH 4.0 and pH 8.0. Spectra of the leucine (green) and arginine (red) mutant at every histidine position are shown superimposed with the spectrum of the wild-type PRO^{FUR} (blue) for comparison.

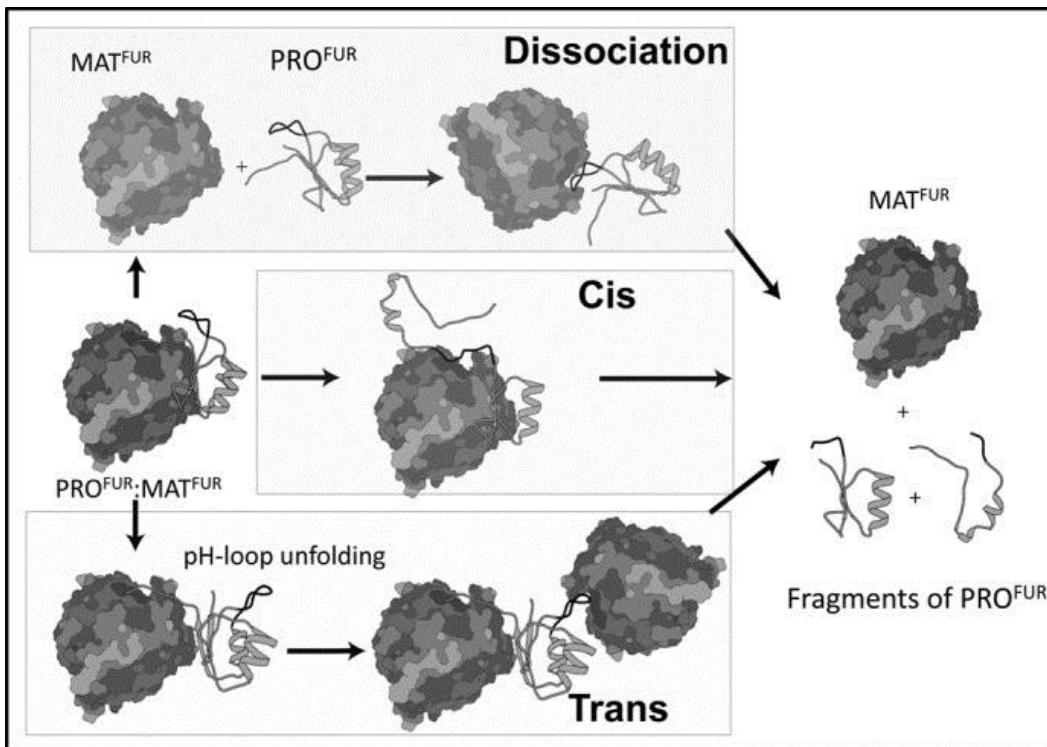


Figure 3-9: Potential mechanisms of furin activation

4

**pKa determination of histidine
residues in the propeptides of furin
and PC1/3 using histidine hydrogen-
deuterium exchange mass
spectrometry**

Johannes Elferich, Danielle M. Williamson, Larry L. David, and Ujwal Shinde

This manuscript was published in *Analytical Chemistry* on August 4th 2015 in
volume 87(15) pages 7909-17

4.1 Abstract

Propeptides of proprotein convertases regulate activation of their protease domains by sensing the organellar pH within the secretory pathway. Earlier experimental work highlighted the importance of a conserved histidine residue within the propeptide of a widely studied member, furin. A subsequent evolutionary analysis found an increase in histidine content within propeptides of secreted eukaryotic proteases compared to their prokaryotic orthologs. However, furin activates in the trans-golgi network at a pH of 6.5 while a paralog, proprotein convertase 1/3, activates in secretory vesicles at a pH of 5.5. It is unclear how a conserved histidine can mediate activation at two different pH values. In this manuscript we measured the pKa values of histidine residues within the propeptides of furin and proprotein convertase 1/3 using a histidine hydrogen-deuterium exchange mass spectrometry approach. The high density of histidine residues combined with an abundance of basic residues provided challenges for generation of peptide ions with unique histidine residues, which were overcome by employing ETD fragmentation. During this analysis we found slow hydrogen-deuterium exchange in residues other than histidine at basic pH. Finally, we demonstrate that the pKa of the conserved histidine in proprotein convertase 1/3 is acid-shifted compared to furin, and is consistent with its lower pH of activation.

4.2 Introduction:

The addition or removal of a proton represents the smallest possible chemical alteration of a protein, but can change the charge by one unit and alter the status of hydrogen bond donors and acceptors[240]. Changes in cellular proton concentrations (pH) can therefore induce rapid reversible chemical modifications by shifting the equilibrium of titratable amino acid side-chains between their unprotonated and protonated forms, thus perturbing the electrostatic potential to drive changes in protein structure, dynamics, and interactions. Hence, there is significant interest in experimentally measuring equilibrium constants (pKas) for protonation/deprotonation of specific amino acid residues within a protein[84].

The ligand binding affinities and catalytic activities of numerous proteins are sensitive to physiologically relevant changes in cellular pH[181]. Histidine is a unique amino acid because the pKa of its imidazole side-chain is close to physiological pH, thus positioning the side-chain as a charge relay system within the catalytic sites of proteins and as a sensor that recognizes subtle perturbations in local pH. An evolutionary analysis demonstrates that eukaryotic propeptide-dependent, secreted proteases are specifically enriched in histidine residues within their propeptides, but not cognate catalytic domains or prokaryotic orthologues[241]. Selective enrichment of histidine residues in propeptides is hypothesized to regulate the organelle-specific, pH-dependent activation of eukaryotic proteases, a finding that was experimentally confirmed in furin, a constitutively expressed proprotein convertase (PC)[58].

PCs take inactive proteins and peptides and, via endoproteolytic cleavage, produce active hormones, enzymes and other critical components of the cellular machinery along the secretory pathway, in the extracellular matrix, and at the cell surface[207]. The PC family includes nine members, furin, PC1/3, PC2, PC4, PACE4, PC5/PC6, PC7/LPC/PC8, SKI/S1P, and NARC-1/PCSK9, all of which are initially synthesized as proproteins at the endoplasmic reticulum (ER) and undergo folding in the ER lumen. Upon completion of folding, the N-terminal propeptide is cleaved in an autocatalytic reaction, but remains associated with the protease and blocks proteolytic activity by occupying the catalytic active site. Final activation requires a second cleavage within the propeptide, which occurs only after the propeptide:PC complex trafficks into the correct organellar compartment[2]. PCs have been implicated in a wide array of pathologies where regulation of the timing and/or location of their activation are altered through genetic aberration or environmental stress[242].

Organellar pH seems to be the major biochemical cue that triggers the final activation step of PCs as individual paralogs display differences in their pH-dependent activation. For example, furin activates at pH 6.5, the pH of the early trans-golgi network (TGN)[18], while PC1/3, a neuroendocrine specific PC activates at pH 5.5, the pH of mature dense core secretory granules[243]. Swapping propeptides between furin and PC1/3 switches their sensitivity to pH-dependent activation, demonstrating that the pH-sensor, which recognizes organellar pH and regulates protease activation, is localized within the propeptide[191]. Studies identify histidine 69 (His₆₉) as the pH-sensor in the furin

propeptide (PRO^{FUR}), as its substitution with a leucine blocks pH-mediated activation[189]. Protonation of His₆₉ does not substantially decrease the affinity of the propeptide at acidic pH, but induces a local conformational change that increases accessibility to a second cleavage site for proteolysis[235].

Nuclear magnetic resonance (NMR) is the most powerful and preferred method to investigate pH-mediated mechanisms, as titratable protons as well as conformational changes can be directly observed within a protein[123]. However, NMR spectral interpretations become difficult if a protein is intrinsically unstable at neutral and acidic pH, as observed in PRO^{FUR} [219]. Mass spectrometry on the other hand identifies peptides by their characteristic mass and is independent of conformational changes that may occur in the protein[244]. The slow hydrogen-deuterium (HD) exchange within the imidazole ring of histidine residues can be exploited to assess the local pK_a of histidine residues[245]. In proteins, the C2 hydrogen in imidazole rings can be exchanged with deuterium with a half-life of days, which is substantially slower than hydrogen bound to oxygen, nitrogen, or sulfur atoms. Therefore, by incubating a protein in deuterated buffer for several days, followed by a shorter incubation (~30min) in hydrogen-containing buffer, one can selectively label the imidazole ring with deuterium. This uptake can then be quantified with mass spectrometry. Since the exchange rate depends on the protonation state of the histidine residue, one can determine histidine pK_as by performing this experiment at different pH values. This method has been successfully used to determine the pK_a of individual histidine residues in RNase A[245] and dihydrofolate reductase[246],

and helped to probe the mechanism by which long range interactions can stabilize the formation of a complex between anthrax protective antigen and its receptor capillary morphogenesis protein-2[247].

Although the pH sensor in furin has been identified[189], several important questions about the mechanism of pH-mediated activation of PCs remain unanswered. For example, the residue corresponding to His₆₉ in PRO^{FUR} is conserved in the propeptide of PC1/3 (PRO^{PC1}). Nevertheless PC1/3 requires a ten-fold higher proton concentration for its activation when compared with furin, suggesting that additional factors are critical for fine-tuning pH sensitivities of individual PC paralogs. While it is clear that His₆₉ in PRO^{FUR} is critical, it is likely that other histidine residues also play roles given that H₆₉L-PRO^{FUR} variant displays moderate structural changes as a function of pH [235]. In this study we address these important mechanistic questions about the pH-mediated activation of PCs by measuring the histidine pKa values in PRO^{FUR} and PRO^{PC1} using hydrogen-deuterium (HD) exchange mass spectrometry. The high density of histidine residues and positively charged residues in these propeptides provides challenges for proteolytic separation, which we overcome by combining the use of Electron-Transfer Dissociation (ETD), Collision Induced Dissociation (CID) and proteolysis using pepsin. Analysis of this data demonstrated a thus far unappreciated slow-dynamic HD exchange in residues other than histidine, which may provide an approach to probe the local environment of additional amino acid residues. We find that His₇₂ in PRO^{PC1}, which corresponds to the established His₆₉ pH-sensor of PRO^{FUR}, has an acid-

shifted pKa of about 5.6, a value that is consistent with its pH for activation.

These results now provide a chemical basis for how PRO^{FUR} controls activation of its cognate catalytic domain at the more neutral pH in the TGN (pH~ 6.5) when compared with PRO^{PC1}, which requires a more acidic pH for activation of the catalytic domain of PC1/3 in the dense core secretory granules (pH~ 5.5).

4.3 Experimental Section:

4.3.1 Purification of propeptides:

PRO^{FUR} and PRO^{PC1} were expressed in E. coli BL21/DE3 and purified by ion-exchange chromatography in 5M urea. After purification proteins were concentrated and stored in 6M guanidinium hydrochloride. Before experiments were performed, proteins were refolded by dialyzing twice against a 100x volume of 50mM Tris pH7.4/50mM NaCl. After refolding the protein was centrifuged for 30 minutes at 100,000g to remove aggregates and concentration was determined by the absorption at 280nm. Point mutations in PRO^{FUR} were generated using the Quikchange protocol and purified identically as wild-type PRO^{FUR}.

4.3.2 HD exchange:

D₂O buffers contained 50mM sodium acetate (pH3.5-4.5), 50mM MES (pH5.0-7.5), and 50mM HEPES (pH 8.0-9.0). The pH was adjusted using DCl or NaOD, with a pH electrode calibrated with standard buffer solution without correcting for the isotope effect, as in previous studies[245], [246]. Conductivity was adjusted with sodium chloride to match a 50mM MES/50mM NaCl buffer at pH

5.0. About 20 μ g of propeptides were diluted 1:50 into 100 μ l of deuterium buffer. A control sample was diluted into the same amount of H₂O buffer. The samples were kept at 37°C for 72h. After this incubation the exchange was quenched by addition of 10 μ l formic acid and the samples were dried using a SpeedVac. For pepsin digestion the sample was resuspended in 20 μ l of potassium phosphate buffer at pH 2.3 containing 0.4 μ g of pepsin. After a 30 minute digest at 37°C the samples were dried again and resuspended in 50 μ l of 100mM ammonium bicarbonate. The samples were incubated for 30 minutes at room temperature to allow for back exchange of polar hydrogens. Samples for whole-protein ETD analysis were directly resuspended in ammonium bicarbonate and allowed to back-exchange for 30 minutes. After drying in a SpeedVac the samples were stored at -80°C until measurement.

4.3.3 Mass analysis:

Mass analysis was performed using an Orbitrap Fusion instrument (Thermo Scientific). Samples were resuspended in 30 μ l 0.1% formic acid. For whole-protein ETD analysis, the sample was automatically desalted using a 1x10 mm protein Opti-Trap™ cartridge (Optimize Technologies, Oregon City, OR) and eluted using 50% acetonitrile/0.1% formic acid (v/v) directly into an electrospray ionization (HESI-II) probe (Thermo Scientific). The instrument was set up to cycle between different ETD reaction times (5ms/10ms/20ms) and MS3 analysis using collision-induced dissociation (CID) as indicated in the Results section. Data was collected using the OrbiTrap mass analyzer in a mass range from m/z 500-1600 at 60,000 resolution and ETD performed on the most intense +14 charge

state of the propeptide. For peptic digest analysis the sample was desalted using a peptide 1x10 mm Opti-Trap™ cartridge and then separated on a 0.5x150 mm SB C-18 reverse-phase column (Agilent Technologies). Peptides were eluted at a 10µl/min flow rate using a linear gradient from 2% to 50% acetonitrile in water containing 0.1% formic acid (v/v). The instrument was set up to collect survey scans at a mass range from m/z 400-2000 and cycle with ETD MS2 scans targeting histidine containing peptides at specific elution times. Specific elution times were determined by an initial test digest of unexchanged propeptide with the instrument set up for data-dependent MS2 fragmentation using higher-energy collisional dissociation (HCD) for peptides with charge states 1-2 and ETD for peptides with higher charge states.

4.3.4 Data processing:

For the whole protein ETD analysis, scans during the elution of the protein were averaged. The spectra generated using the three different reagent reaction times were searched for the expected isotopic distribution of all possible c and z fragments at charge states from +1 to +10. Fragments were considered identified if the root mean square deviation (RMSD) between the expected and measured isotopic distribution was smaller than 10% at one of the different reaction times. If a fragment was found at multiple reaction times, the reaction time with the best RMSD was chosen. MS3 spectra were processed similarly, by searching for features with the expected mass of all theoretical fragments. To identify peptic peptides, survey scans were searched for the expected isotopic distribution of all possible peptides, assuming no preferential cleavage of pepsin. Matches were

ordered by intensity and the identity of peptides with histidines was verified by manual inspection of triggered MS2 scans.

To determine the rate of hydrogen deuterium exchange one must calculate the deuterium uptake from mass measurements of the peptide or fragment ions. This information is encoded in the shifting pattern of the isotopic distribution. Previously the uptake has been derived either from the ratio of the monoisotopic (I) and I+1 peak[245] or by calculating the average mass from the isotopic distribution, where the difference from the unexchanged average mass is used to quantify the uptake of deuterium[248]. The first method has the disadvantage that it cannot be used in the cases of peptides containing two or more histidines. We also found that using the average mass was susceptible to artifacts due to the lowest intensity peaks being unreliably detected. To overcome these problems we chose to calculate the uptake by a fit of the observed isotopic distribution with a linear combination of the theoretical isotopic distributions of the non-exchanged and fully exchanged peptides, which, in the case of peptides containing single histidines, can be derived by simply shifting the peaks by one mass unit. This concept can easily be extended to peptides containing multiple histidines by deriving theoretical models of complete exchange at two or more sites by shifting the peaks by two or more mass units. It also allows calculation of the uptake with only a subset of peaks, as long as the number of peaks is larger than the number of exchanging sites. For analysis of exchange at pH 10.0, where the number of exchanging site was unknown we used the average mass for quantification of deuterium uptake. After calculation of uptake the exchange rate

was calculated using the following formula[245], where u is the calculated uptake and t is the exchange time:

$$k = -\frac{\ln(1-u)}{t} \quad (1)$$

pKa and k_{max} values were derived by fitting the exchange rates at different pH values to the following equation[245]:

$$\log\left(\frac{k_{max}-k}{k}\right) = pKa - pH \quad (2)$$

Data processing was performed using python scripts based on the mass spectrometry library of the mMass program[249].

4.3.5 Homology modeling and pKa prediction:

Homology models of PRO^{FUR} were build using the MODELLER program[178] using the automodel module. Five models were built for each of the 20 models deposited for PRO^{PCl} NMR structure[215]. pKa values were predicted using the PROPKA program version 3.1 [128].

4.4 Results and Discussion:

4.4.1 Identifying ions containing a single histidine residue using whole-protein electron-transfer dissociation and proteolytic digestion:

To measure deuterium uptake at individual histidine residues within a protein, it is essential to obtain peptides that contain a single histidine residue[245]. This is often achieved using trypsin, a protease that cleaves proteins mainly at the C-terminus of lysine and arginine amino acids. However, a tryptic digestion of PRO^{FUR} fails to produce peptides that contain a single histidine residue, most likely because of the abundance of lysine (6%), arginine (13%) and histidine (6%)

residues, as well as the presence of a 15 amino acid region that contains nine positively charged residues; four histidines and five arginine and lysines, as illustrated in Figure 4-1A. On the other hand, the high positive charge makes PRO^{FUR} a promising candidate for electron transfer dissociation (ETD), a process that induces fragmentation along the peptide backbone in a sequence-independent manner. Hence, we initially employed the easy-ETD module of the OrbiTrap Fusion instrument to fragment PRO^{FUR} inside the mass spectrometer in a “top-down” manner.

Injections of PRO^{FUR} into the OrbiTrap produce multiple charged states that range from +8 to +14 (Figure 4-1B). ETD fragmentation of PRO^{FUR} with a +14 charge produces a fragmentation pattern that is highly dependent on the reaction time (Figure 4-1C), where many fragments can only be identified at specific reaction times, either because they are not present or because they overlap with other fragments. Using various reaction times, 15 and 21 of the 83 possible c and z fragments, respectively, can be identified from the resulting MS2 spectra. A peptide fragment is considered suitable to quantify deuterium uptake only when the root mean square deviation between the observed and the expected isotopic distribution is smaller than 10%. The above criteria identify fragment ions c29 and z24 containing His₅₂ and His₈₄, respectively, along with ion z28 that includes His₈₄ and His₈₀, which can be used to calculate deuterium uptake of His₈₀ by subtracting the deuterium uptake of His₈₄. Although the fragment c46, which includes His₅₂, His₆₆, and His₆₉, is observable, differentiating mass increases

between His₆₆ and His₆₉ is not feasible due to the lack of appropriate smaller fragment ions.

To overcome this problem we subjected PRO^{FUR} with a +14 charge to collision-induced dissociation (CID), which is a process that displays greater sequence specific cleavage than ETD. Top-down methods for hydrogen/deuterium exchange of backbone amide hydrogen have been shown to be successful only when using ETD as a fragmentation mode[130], [250], since CID causes scrambling of exchanged deuterium along the protein backbone[251]. This is of no concern in this study, as the deuterium is bound to C2 in the imidazole side-chain. The CID-MS2 spectrum identifies fragment y47 (+8 charged state) as the most intense ion (Figure 4-1D). The OrbiTrap Fusion allows the selection of the y47 ion for subsequent MS3 fragmentation using CID, which yields the fragment y47_b6 that only contains His₆₆. The ETD fragmentation of y47 generates the y47_c17 ion that contains both His₆₆ and His₆₉. The difference between the masses of y47_c17 and y47_b6 ions yields the mass increase of His₆₉. Figure 4-1E shows the observed isotopic distribution of the fragment ions used for deuterium uptake determination.

To test whether this whole protein CID/ETD approach allows for accurate measurements of deuterium uptake into the imidazole side-chain, we repeated the experiment by cleaving PRO^{FUR} using pepsin. Cleavage by pepsin produces peptides that allow the measurement of His₅₂ and His₆₆ (Figure 4-2A-B).

Although pepsin digest produces a peptide containing His₆₉ alone, it is of very low abundance due to low probability for cleavage at Phe₆₇ and Thr₇₃ and poor

retention of the resulting peptide on the reverse phase chromatography column. Moreover, cleavage between His₈₀ and His₈₄ residues is not observed when using pepsin, and hence ETD was employed to obtain accurate masses of fragments that contain His₆₉, His₈₀ and His₈₄ (Figure 4-2C). While this results in lower intensity of the fragment ions when compared with the whole peptide masses, it does not require determination of deuterium uptake by subtracting two mass measurements. Thus, using a combination of both ETD and CID in a “top-down” approach or pepsin cleavage and ETD, we successfully measured the mass of fragments that contain individual histidine residues.

4.4.2 Individual histidine residues in PRO^{FUR} display comparable pKa values, but vary in their solvent accessibility:

To calculate the pKa and k_{\max} values for protonation of individual histidine residues, we performed HD exchange for 72h at 37°C at different pH-values as described in the methods and fitted the exchange rates of individual histidine residues to equation 2 (Figure 4-3). Since the fit indicates more than one exchanging site at high pH, we excluded the data points at pH 8.5 and 9.0 for His₅₂ generated using the whole protein ETD/CID approach. Moreover, since PRO^{FUR} displays changes in conformation as a function of pH[191], the precise fitting of the exchange rates to equation 2, which assumes no changes in protein conformation, can only report the apparent pKa values (pK_{aapp}) for the protonation of histidine residues.

Both whole protein ETD/CID and pepsin digestion approaches demonstrate that the pK_{aapp} values for all histidine residues are close to 6.0, while

measurement of the maximal exchange rate k_{\max} at the imidazole ring shows that individual histidine residues in PRO^{FUR} differ substantially in their solvent accessibility (Table 4-1). For example, His₅₂ displays the highest rate of HD exchange when compared with all other histidine residues in PRO^{FUR}. It is worth noting that the k_{\max} values obtained by the two methods are different for His₅₂, His₆₆ and His₆₉, but are similar for His₈₀ and His₈₄ (Table 4-1). The anomaly in k_{\max} values for His₅₂, His₆₆ and His₆₉ obtained using the different approaches is due to at least two factors; first, while the pepsin digest approach measures mass increase in ten residues around His₅₂, the whole-protein ETD approaches measures mass increase in 28 residues N-terminal of His₅₂. The analyses of these peptides demonstrates additional deuterium incorporation occurs in the R-group of amino acids located within the extra 19 residues present in the ETD, but not the pepsin digestion approach. The additional exchange is greatest at alkaline pH, explaining the multi-site exchange observed at pH 8.5 and 9.0. This finding is described in more detail below (Figure 4-4). Second, the lower maximal exchange rate for His₆₆ identified in the ETD/CID approach, is likely a result of the low intensity of the y47_b7 fragment, which may lead to an underestimation of HD exchange. The exchange at His₆₉ in the ETD/CID approach is computed by subtracting the exchange of His₆₆ from the exchange into the y47_c16 ion, which likely leads to overestimation of deuterium uptake in His₆₉. In retrospect, the pepsin digest approach appears to be less prone to artifacts, although the ETD approach provides insights into additional amino acids that could undergo HD exchange under alkaline conditions. Since in-instrument fragmentation

techniques, such as CID and ETD, generate a single break in the peptide backbone, only exchange in the first N-terminal and C-terminal histidines can be directly measured using MS2 spectra. To quantify deuterium uptake at internal histidine residues, one must either use the difference between two fragments, or employ MSn spectra to further fragment peptide ions. While the first approach correctly assigns deuterium uptake at His₈₀ and His₈₄ in PRO^{FUR}, it poses a disadvantage because errors in the measurement of the mass of individual fragments tend to be magnified. However, the second approach, which employs the MS3 spectra to measure the mass increase of His₆₆, poses a disadvantage by severely reducing the intensity of the fragment ions, which may lead to the underestimation of histidine uptake. Certainly further advances in top-down mass spectrometry, such as charge-state pooling, will help to overcome these difficulties.

Overall these findings suggest that the local protein environment only marginally influences the protonation equilibrium of the imidazole side-chain in PRO^{FUR}, but influences the HD exchange rate due to its solvent accessibility and local hydrogen bonding as seen by the higher k_{\max} values for His₅₂ and His₈₀ compared to His₆₆, His₆₉, and His₈₄. This indicates that the local protein environment is different for these histidine residues.

4.4.3 Alkaline pH induces HD exchange of carbon-bound protons in addition to the C2-proton in the imidazole ring:

Since the c29 fragment in PRO^{FUR} appears to exchange at more than one site under conditions of alkaline pH (pH \geq 8.5) as described earlier, we next

investigated whether (i) carbon-bound protons residues other than the C2-proton of the imidazole ring in histidine can undergo HD exchange and (ii) deuterium uptake can be analyzed in pepsin derived peptides that lack histidine residues. Quantification of deuterium uptake by fitting the experimental data to a model with two exchanging sites shows that an increase in pH above 8.0 enhances deuterium uptake into the c29 fragment (Figure 4-4A). Analysis of pepsin-generated peptides establishes that peptides ²⁵MQGQKVFTNTW³⁶ and ³⁷AVRIPGGPANSVA⁴⁹, which are N-terminal of His₅₂ and are part of the C29 fragment ion, demonstrate deuterium uptake at pH 8.5 and 9.0 (Figure 4-4B). The magnitude of these mass increases is consistent with the additional HD exchange observed in the c29 fragment obtained using ETD. However, peptides ⁵⁷GQIFGGDY⁶⁴ and ⁸⁸QREPQVWL⁹⁵, which also lack histidine residues, demonstrate no deuterium uptake at pH 8.5 and 9.0. Since the additional HD exchange at alkaline pH may result from the fact that an increasing OH⁻ concentration likely removes carbon-bound protons at other residues, experiments were also conducted at a pH of 10.0. The results demonstrate the magnitude of deuterium uptake increases in the peptides ²⁵MQGQKVFTNTW³⁶ and ³⁷AVRIPGGPANSVA⁴⁹, but also causes deuterium uptake within peptides ⁵⁷GQIFGGDY⁶⁴ and ⁸⁸QREPQVWL⁹⁵, which display no HD exchange at pH of 9.0 (Figure 4-4C). Identical experiments conducted in non-deuterated buffer at pH 10.0 show no mass increase, thus confirming the mass increase observed in deuterated buffers results from HD exchange and not due to chemical

modifications, such as a deamidation event that would also cause a 1Da mass increase[252].

To investigate which residues undergo HD exchange at alkaline pH, we monitored deuterium uptake within CID fragments of ²⁵MQGQKVFTNTW³⁶ and ³⁷AVRIPGGPANSVA⁴⁹ at pH 10.0 (Figure 4-4D-E). Our analyses show that in both cases, asparagine residues are most prone to uptake of deuterium, with additional (albeit smaller) uptake in glycine, glutamine and threonine residues. While in the ²⁵MQGQKVFTNTW³⁶ peptide deuterium uptake seems to be mediated by multiple residues, in the ³⁷AVRIPGGPANSVA⁴⁹ peptide deuterium uptake is clearly limited to Asn₄₆ and Gly₄₃, while Gly₄₂ shows no uptake. This indicates that slow HD exchange in residues other than histidine is not only dependent on the chemical nature of the side-chain, but also on the context of the local environment within the protein.

The mass increases may be due to HD exchange of backbone amides with very slow kinetics, as they are occasionally noted in proteins[253]. However, they are still observed after back-exchange of already digested protein. It is hard to imagine that small peptides could provide such strong protection of backbone amides. This leads to the possibility that protons bound to carbon atoms other than the C2 carbon of the imidazole ring can exchange at alkaline pH. One such example is the uptake of deuterium by racemization at C α -atoms of serine residues during HD exchange of immunoglobulins at pH 8 and 40°C for four weeks[254]. HD exchange at C α atoms by base-catalyzed racemization also occurs at N-substituted glycine-containing peptides at a pD of 12.3[255] and may

explain the exchange we observe in Gly₂₇ and Gly₄₃ at pH 10.0. However, the majority of the exchange at pH 8.5 and 9.0 can probably be attributed to asparagine (Figure 4-4D-E). Asparagine is prone to racemization due to formation of a succinimide intermediate[256], whose product due to the low concentration of ammonia compared to water is almost always an aspartate. The tetrahedral intermediate during succinimide formation can however racemize back to an asparagine containing peptide[257]. It is likely that the low stability of PRO^{FUR} allows sufficient protein backbone flexibility to permit racemization.

This observation shows that caution must be used when measuring histidine pKa by HD exchange to remove this influence from curve fittings. This is easily achieved in practice since this exchange shows a characteristic pH profile, where no saturation of the exchange rate is observed at basic pH. While in this study the unexpected exchange observed at residues other than histidine was an artifact it may be useful in the study of protein aging and misfolding. Since racemization depends on local flexibility of the protein backbone it may also be helpful in studying disordered protein regions and warrants further study.

4.4.4 His₇₂ in PRO^{PC1} displays a pK_{a,app} shifted to a more acidic value:

PRO^{FUR} and PRO^{PC1} adopt similar three-dimensional structures but differ in the density and distribution of histidine residues within their otherwise conserved sequences. Circular dichroism studies demonstrate that PRO^{PC1} requires about ten-fold higher proton concentration to undergo a 50% loss in its secondary structure when compared with PRO^{FUR} [191]. Consistent with these studies,

biochemical approaches show that PRO^{FUR} mediates activation of MAT^{FUR} at a pH of 6.5, while its paralog, PRO^{PC1} modulates activation of MAT^{PC1} at a pH of 5.5. The lower pH of activation of PRO^{PC1} may result from lower pKa values of a subset or all histidine residues in the propeptide. To examine this possibility we measured the pKa values of individual histidine residues in PRO^{PC1} using the pepsin/ETD hybrid approach that gives the best results for PRO^{FUR} (Figure 5). The results demonstrate that three of the four histidine residues have pK_{a,app} values of ~6.0, similar to those in PRO^{FUR}. However, His₇₂, displays a substantially lower pK_{a,app} value of 5.6 (Table 4-2). While the maximal exchange rates are slightly less than those observed for PRO^{FUR}, this finding is consistent with the higher structural stability PRO^{PC1}.

4.4.5 A structural interpretation of histidine pKa values:

Our data suggests that histidines within PRO^{FUR} display similar pK_{a,app} values, but differ in their observed exchange rates, while individual histidines in PRO^{PC1} display substantial differences in their pK_{a,app} values. How can one rationalize these differences using available sequence and structural information? A sequence alignment between PRO^{FUR} and PRO^{PC1} establish that only one histidine, His₆₉ in PRO^{FUR}, has a direct equivalent in His₇₂ within PRO^{PC1} (Figure 4-6A). His₅₂ in PRO^{FUR} is unique with no direct counterpart in PRO^{PC1}. The other histidine residues are not directly identical in their positions but are all situated around the second cleavage site Arg₇₂ and Arg₇₈ in PRO^{FUR} and PRO^{PC1}, respectively. To analyze this in a structural context we created homology models of PRO^{FUR} based on the solution NMR structure of PRO^{PC1} (Figure 4-6B-C).

In order to rationalize the measured $pK_{a,app}$ values we used the PROPKA program to predict histidine pK_a values based on the NMR solution structures of PRO^{PC1} and the PRO^{FUR} homology models (Figure 4-6D-E). In the case of PRO^{PC1}, the predicted pK_a values are consistent with the measured $pK_{a,app}$ values for His₆₇, His₇₅, and His₈₅. His₇₂ is predicted to have the most basic pK_a values of all histidines, while it has the most acidic $pK_{a,app}$. Interestingly for two of the 20 NMR structures the pK_a of His₇₂ is predicted to be strongly acidic. In these structures the imidazole side-chain is not exposed to the solvent but is packed under residues of the cleavage loop. Theoretical consideration of pH sensing suggests that titratable residues involved in pH-driven structural changes should have divergent theoretical pK_a values in the two different structural states with an observed pK_a value that is between these two values[258]. Since the observed $pK_{a,app}$ is between the pK_a predicted for His₇₂ buried within the loop and His₇₂ pointing towards the solvent, this suggests that movement of His₇₂ from the buried conformation towards the solvent is part of the pH sensing mechanism. The fact that His₇₂ is the residue corresponding to the previously identified primary pH-sensor His₆₉ in PRO^{FUR} confirms a central role in pH sensing, although specific histidine residues are likely to augment the sensitivity of His₇₂ in PRO^{PC1} (Williamson, et al manuscript in preparation).

Predicted pK_a values for PRO^{FUR} vary substantially between histidines compared to the consistent $pK_{a,app}$ values that were measured. This might be due to low quality of the homology model due to the low sequence identity (~40%) and the fact that PRO^{FUR} in solution shows strong structural dynamics and probably is in

a molten-globule like state[219]. The predicted pKa for His₅₂ is especially acidic, probably due to burial of His₅₂ between the two alpha helices. Since His₅₂ displays the fastest observed HD exchange rate (Table 4-1), this suggests that His₅₂ is more solvent exposed than this structural model suggests.

Since association with the protease likely stabilizes the propeptide structure, this might change the effect of histidine protonation. However, since the propeptide:protease complex is not stable enough *in vitro* (data not shown) we are currently limited to studying isolated propeptides. Since the PRO^{PC1} solution structure is similar to structures of propeptides in complex with proteases[215], it is likely that principles observed in isolated propeptides also apply to the complex.

4.5 Conclusion:

Previous knowledge about the pH sensing mechanism of proprotein convertases was based on mutagenesis of histidine residues, which might introduce artifacts beside the removal of a titratable group. Therefore the use of the histidine HD exchange method on wild-type propeptides provides orthogonal knowledge. We show that while all histidines in PRO^{FUR} show a similar pKa_{app} they do experience different local environments. We furthermore show that the activation of PC1/3 at lower pH can be explained by a shift of the pKa of the primary pH-sensor His₇₂ to a more acidic value. Comparison of this value with predictions based on a NMR structure suggests the movement of the conserved primary pH-sensor His₇₂ from a protected pocket towards the solvent may be one of the key events in pH-mediated activation.

Table 4.1: Parameters of histidine HD exchange rate fits for PRO^{FUR} at different pH values using equation 2

Histidine	Whole protein ETD/CID		Pepsin/ETD	
	pK _a _{app}	k _{max}	pK _a _{app}	k _{max}
H52	6.04±0.04	0.0175±0.0003	6.07±0.02	0.0124±0.0001
H66	5.84±0.08	0.0045±0.0002	5.98±0.03	0.0069±0.0001
H69	6.13±0.05	0.0155±0.0003	6.04±0.05	0.0063±0.0002
H80	6.04±0.08	0.0112±0.0003	6.02±0.03	0.0094±0.0002
H84	5.96±0.08	0.0077±0.0003	6.04±0.07	0.0073±0.0002

Table 4.2: Parameters of histidine HD exchange rate fits for PRO^{PC1} at different pH values using equation 2

Histidine	pK _a _{app}	k _{max}
H67	6.31±0.03	0.0068±0.0001
H72	5.61±0.06	0.0034±0.0001
H75	5.97±0.03	0.0057±0.0001
H85	5.85±0.04	0.0052±0.0001

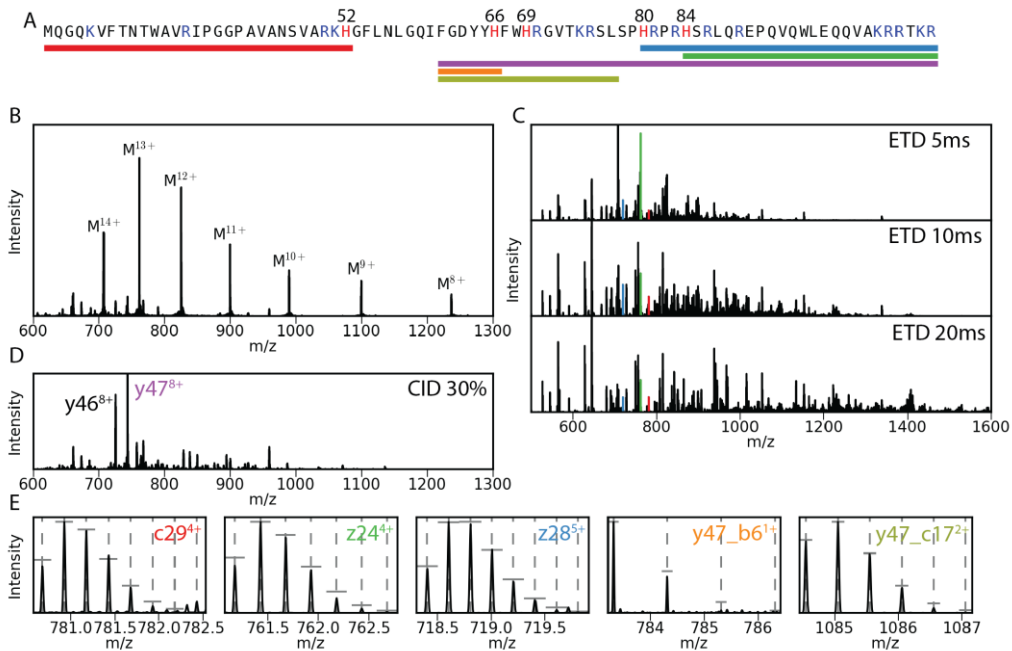


Figure 4-1: Identification of ions for HD exchange measurements for individual histidines using the whole-protein ETD/CID approach.

(A) Sequence of PRO^{FUR} with histidines colored red and other basic residues colored in blue. Numbers indicate sequence position of histidines. Colored bars indicate the fragments that were used to measure individual histidine masses. (B) Mass spectrum of whole PRO^{FUR} . (C) Mass spectra after ETD reaction of the M^{14+} ion. Reaction time of ETD is indicated. Colored peaks indicate fragments used for later analysis. (D) Mass spectra after CID fragmentation at 30% collision energy. Most abundant fragments $y47$ and $y46$ are indicated. (E) Close-up of isotopic peaks of fragments used to measure deuterium uptake. Dashed vertical lines indicate expected m/z, and horizontal vertical lines indicate expected relative intensity given the natural isotopic abundances in the given peptides.

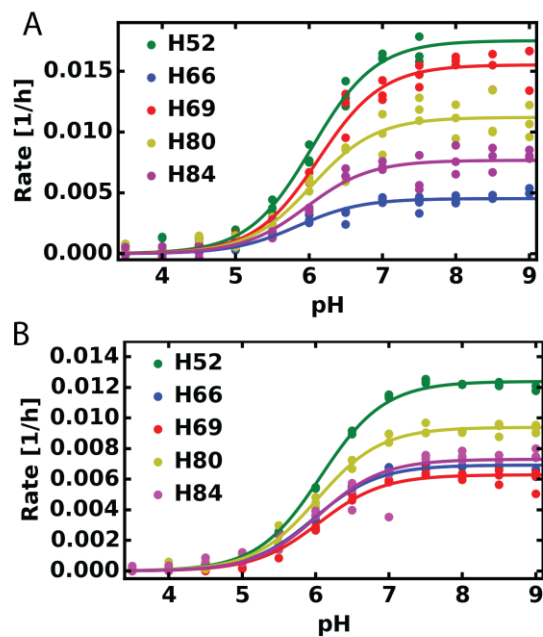


Figure 4-3: pH-dependent HD exchange rate of PRO^{FUR}

(A) Rate of HD exchange as a function of pH for individual histidine residues as determined by the whole-protein ETD/CID approach for PRO^{FUR}. Three data points for each pH are shown and are derived from independent experiments. The result of a non-linear fit against equation 2 is shown as a solid line. Results of the fit are listed in Table 1. (B) Same as in panel A, but with data derived from the pepsin/ETD approach.

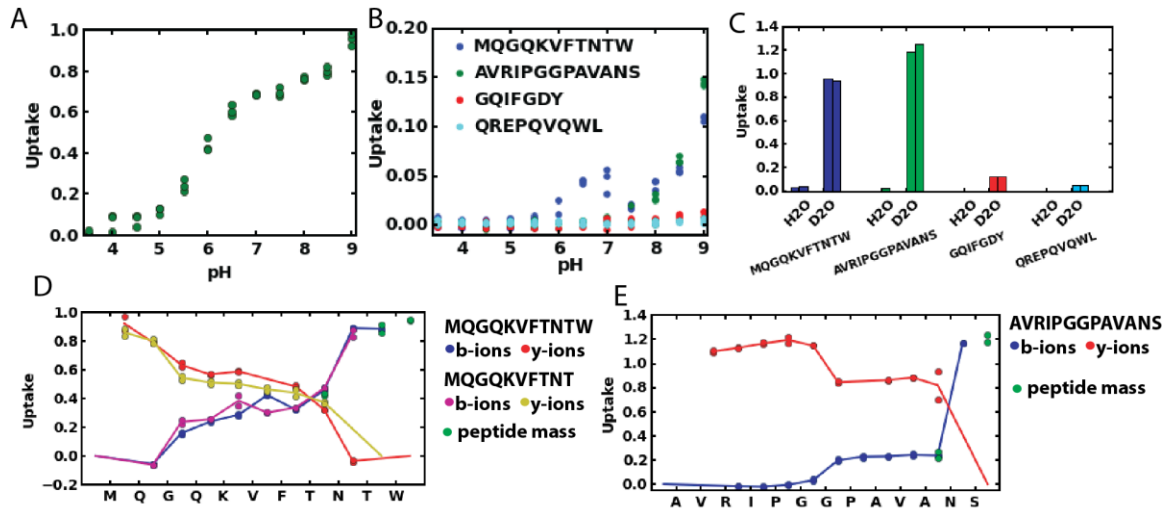


Figure 4-4: HD exchange in residues other than histidine

(A) Deuterium uptake as a function of pH into fragment c29 of PRO^{FUR} determined by a fit of the isotopic distributions against a model of 2 exchanging sites. (B) Deuterium uptake into histidine-free peptides as a function of pH. (C) Deuterium uptake at pH 10.0 into histidine free peptides either in H₂O or D₂O buffer. (D) Deuterium uptake into CID-generated fragments of the ²⁵MQGQKVFTNTW³⁶ and ²⁵MQGQKVFTNT³⁵ peptides after exchange for 72h at pH 10. The different b- and y-series are color-coded and connected by lines. Uptake as determined by whole peptide masses is indicated in green, including the uptake into the ²⁵MQGQKVFT³³ peptide. (E) Deuterium uptake into CID-generated fragments of the ³⁷AVRIPGGPAVANS⁴⁹ peptide after exchange for 72h at pH 10. The b- and y-series are color-coded and connected by lines. Uptake as determined by whole peptide masses is indicated in green, including the uptake into the ³⁷AVRIPGGPAVA⁴⁷ peptide.

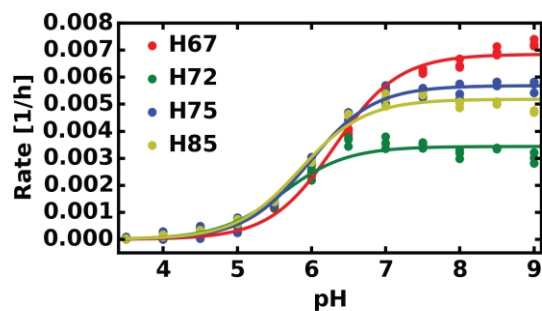


Figure 4-5: pH dependent HD exchange in PRO^{PC1}

The rate of HD exchange as a function of pH for individual histidine residues in PRO^{PC1}. Three data points for every pH are shown and are derived from independent experiments. The result of a non-linear fit against equation 2 is shown as a solid line. Results of the fit are listed in Table 2.

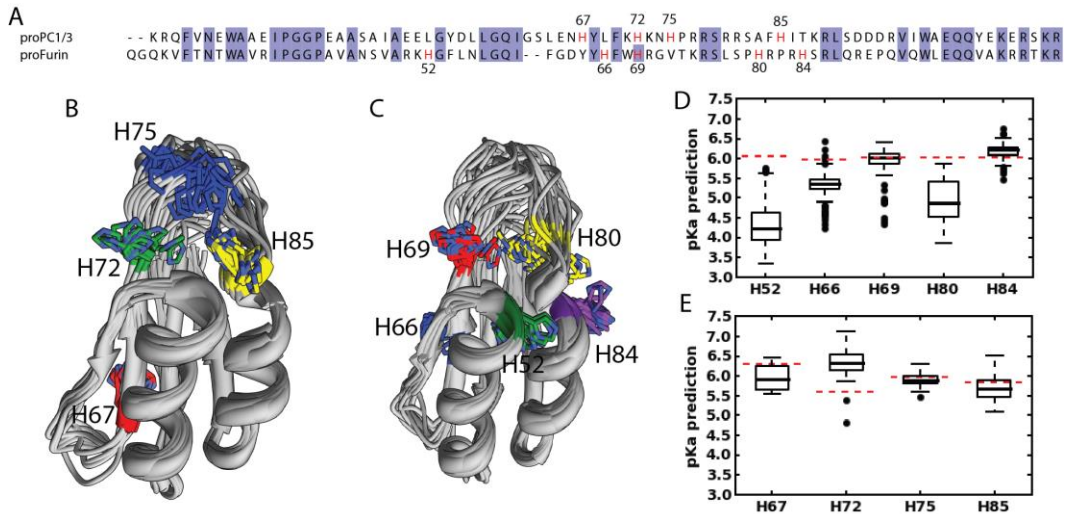


Figure 4-6: Structural interpretation of measured pK_{aapp} and maximal exchange rates

(A) Sequence alignment of human PRO^{FUR} and PRO^{PC1}. Conserved residues are shaded in grey and histidines are shown in red. (B) Solution NMR-structure of PRO^{PC1} [215]. All 20 deposited structures are superimposed. Histidine residues are shown in stick representation and color coded according to figure 5. (C) Homology model of PRO^{FUR} based on PRO^{PC1}. Best scoring model based on each of the 20 deposited structures of PRO^{PC1} are superimposed. Histidine residues are shown in stick representation and color coded according to figure 3. (D) Histidine pKa predictions of PRO^{FUR} using the PROPKA program. Predictions based on 5 models per template structure are shown as a boxplot with the median shown as a bold line, the interquartile range shown as a box, and the range defined as 1.5 times the interquartile range shown as whiskers. Outliers are shown as black spheres. The experimentally determined pK_{aapp} values are shown as a dashed red line. (E) Histidine pKa predictions of PRO^{PC1} using the PROPKA program. Predictions are based on each of the 20 deposited structures. Data is depicted as described for panel D.

5

Using computational techniques to understand protein function in disease

This chapter is based on four manuscripts and focuses on my contributions of computational techniques in collaborative projects. The collaborating laboratories generated the experimental data shown in figures 5-3,5-4, and 5-5. The experiments were designed based on my models.

Identification of the intracellular gate for a member of the equilibrative nucleoside transporter (ENT) family.

Valdés R, **Elferich J**, Shinde U, Landfear SM

This manuscript was published by The Journal of Biological Chemistry on February 4th 2014 in volume 289 pages 8799-8809

Monoubiquitination Is Critical for Ovarian Tumor Domain-containing Ubiquitin Aldehyde Binding Protein 1 (Otub1) to Suppress UbcH5 Enzyme and Stabilize p53 Protein

Yuhuang Li, Xiao-Xin Sun, **Johannes Elferich**, Ujwal Shinde, Larry L. David and Mu-Shui Dai

This manuscript was published by The Journal of Biological Chemistry on January 8th 2014 in volume 289 pages 5097-5108

Cotranslational folding inhibits translocation from within the ribosome-Sec61 translocon complex.

Conti BJ, **Elferich J**, Yang Z, Shinde U, Skach WR.

This manuscript was published by Nature Structural and Molecular Biology on February 23rd 2014 in volume 21 pages 228-235

Foretinib is a potent inhibitor of oncogenic ROS1 fusion proteins

Monika A. Davarea, Anna Saborowskib, Christopher A. Eidea, Cristina Tognona, Rebecca L. Smitha, **Johannes Elferich**, Anupriya Agarwala, Jeffrey W. Tynera, Ujwal P. Shinded, Scott W. Loweb, and Brian J. Druker.

This manuscript was published by Proceedings of the National Academy of Sciences on November 26th 2013 in volume 110 number 48 pages 19519-19524

5.1 Abstract

The exponential increase of computational power coupled with the development of increasingly sophisticated algorithms and the standardization of protocols for deposition of sequence and structure data into publicly available databases have created an environment where computational biology is a powerful technique to generate and test hypotheses. This chapter demonstrates several examples how computational methods can be combined with experiments. The first example shows how structure prediction can be used to gain insight into a transporter that was not amenable to structural studies *in vitro*. Hypotheses about the structure and mechanism of a purine transporter of *Leishmania donovani* were generate *in silico* and then tested *in vitro*. In the second example, a specialized structure prediction technique was first used to test a hypothesis about the mechanism by which mono-ubiquitination of an enzyme called Otub1 stabilizes p53. After arguing against this hypothesis the structural models were used to propose an alternative hypothesis that was then tested using site-directed mutagenesis. In the third example sequence and structure databases were combined to test a hypothesis about folding during cotranslational transport into the ER. The results did not support the hypothesis that evolution selected against small domains in N-terminal regions of secreted proteins and therefore points to more complex mechanisms. The last example shows how structure prediction can be used to quickly rationalize inhibitor binding to pharmacological targets and to explain potential resistance mechanisms. This can speed up iterations of initial studies, clinical studies, and inhibitor improvements.

5.2 LdNT1: Understanding purine uptake of parasites on a structural level

5.2.1 Introduction

Leishmania donovani is a protozoan parasite and the pathogenic agent of visceral leishmaniasis, a disease that currently affects 12 million people worldwide. Since current treatment options have problems with toxicity, high cost, and emerging resistance, new therapeutic strategies that target biochemical processes unique to the parasite are needed [259].

One potential strategy relies on the fact that *Leishmania* are unable to synthesize purines *de novo* and rely on uptake of purines from their hosts. In order to transport nucleosides through the membrane *Leishmania donovani* encodes two nucleoside transporters, LdNT1 and LdNT2 [260]. Inhibition of these transporters would be a promising approach for treatment of leishmaniasis, but structural information about these transporters is sparse. Direct structural studies of these transporters are hindered by the inability to overexpress these transported in orthogonal expression system or in the parasite. Given the lack of close homologs in prokaryotes that might be more amenable to structural studies, this means that experimental insights into the structure-function relationships are limited to genetic and biochemical approaches, such as site-directed mutagenesis and crosslinking. In order to guide these studies and interpret their results, computational techniques have been successfully implemented to predict the structure using *ab initio* techniques and relying on experimentally determined structures of proteins with remote homologies [261].

Models built using these approaches have been successfully used to predict point mutations that affect the transporter function [261] and guide crosslinks at the extracellular gate of the transporter [262].

Despite the success of these computational approaches, questions remain regarding the mechanism by which purines are transported across the membrane remain. LdNT1 belongs to the major facilitator superfamily (MFS) of transporters, which are believed to function using an “alternating access” mechanism, where the transporter can exist in two different conformational states, where one is open to the outside of the cell and the other is open to the cytosol [263]. Using HMM sequence profiles to identify a remote homology of LdNT1 to the fucose transporter whose structure had recently been solved in an outward facing conformation [264], a homology model was built. The model was then used to predict point mutations in the structure at a position that likely corresponds with the intracellular gate. The mutations were then analyzed for loss of transporter function and were used as targets for creating crosslinks in the transporter. Results from these studies suggest that LdNT1 also functions by use of an “alternating access” mechanism.

5.2.2 Materials and Methods

The sequence of LdNT1.1 was first aligned to FucP using a Hidden Markov profile comparison, generated by HHsearch software, employing the HHPred web server and default parameters [265]. This alignment produced a probability score of 98.4% and an E value 2.0×10^{-5} . The profile for LdNT1.1 was built automatically by searching a profile database based on UniProt clustered at 20%

sequence identity using HHblits [266]. The profile for FucP was one that was catalogued in the HHPred database. A homology model of LdNT1.1 was then generated from this alignment and the FucP crystal structure employing the MODELLER software [178]. The model was then visualized using PyMOL software (PyMOL Molecular Graphics System, Schrödinger LLC).

5.2.3 Results

Previously, our laboratory had developed an *ab initio* model of LdNT1.1 that showed an inside-open conformation (Figure 5-1). Since LdNT1.1 is a member of the major facilitation superfamily (MFS), we speculated that like other members of this family it would function by an alternating-access mechanism, where the two pseudo-symmetrical halves of the transporter undergo a ~38 degree “rocker-switch” motion [262]. When we approximated the resulting outside-open conformation by applying this motion to the *ab initio* model, we saw as expected an opening of the outside gate, where residues Ala₆₁, Phe₃₇₄, and Gly₃₅₀, that were shown to crosslink, move apart from another. However, upon inspecting the inside gate we found strong clashes of helices 11 and 2 as well as helices 10 and 5, indicating that the *ab initio* structure did not represent a biophysical meaningful model of the inside gate.

To acquire information about the inside gate we used the recently published structure of the *e. coli* fucose transporter (FucP) [264] as a template for homology modeling. FucP was crystallized in an outward-open conformation. A HMM-profile search identified FucP as a homolog of LdNT1.1 as described in “Materials and Methods”. Even though the sequence identity is extremely low

(13%), the alignment had strong statistical support (Figure 5-2). While the alignment is ambiguous in some of the transmembrane helices, the C-termini of helices 5 and 11, as well as the N-termini of helices 4 and 10 are well aligned. Since these ends make up the inside gate in the FucP structure, we reasoned that a homology model may provide meaningful information about the inside gate.

The homology model derived from the FucP template and the alignment in Figure 5-2, as expected, showed that the inside site of the transporter was shielded from solvent by interactions made by helices 4, 5, 10, and 11. To test the accuracy of the inside gate structure extensive side-directed mutagenesis of residues at the ends of helices 4,5,10, and 11 were performed. Eighteen residues (TM4: Thr₁₆₀, Tyr₁₆₁, Gly₁₆₂, Met₁₆₃, Phe₁₆₄, Phe₁₆₇; TM5: Thr₁₇₄, Met₁₇₅, Met₁₇₆; TM10: Leu₄₄₄, Val₄₄₅, Leu₄₄₆; TM11: Met₄₆₆, Gly₄₆₇, Ile₄₆₈, Ser₄₆₉, Ile₄₇₀, Leu₄₇₁) were mutated to alanine and introduced into a Δ ldnt1 Δ ldnt2 *L. donovani* double null mutant that is genetically deficient in the LdNT1.1, LdNT1.2, and LdNT2 genes and consequently provides a null background for transport of nucleosides. This allowed for testing of the nucleoside uptake capability of these mutants in a native environment (Figure 5-3D). Six of these residues strongly impaired nucleoside uptake (>90%) compared to wild-type, mutation of one residue slightly impaired uptake (88%), while the other residues had only modest or no impact on nucleoside uptake (<70%). The reduction of uptake could not be explained by the amount of surface expression (Figure 5-3E). When we mapped the position of the mutations that impacted transporter function onto the homology model, we found that they clustered strongly at the interface formed

between the helices 4,5,10, and 11 (Figure 5-3A), while residues that did not impact transporter function were not part of the interaction between the N- and C-terminal half of LdNT1.1 (Figure 5-3B). Mapping of the critical mutations onto the previous inside-open *ab initio* model shows that these residues do not appear to interact in the inside-open state. However, only after the transition of the transporter to the model that represents the in the inside-closed state do these residues appear to interact. Since five of the critical residues within the putative inside gate are hydrophobic, while the two others are glycine residues, I proposed that the inside gate is most likely formed by residues Tyr₁₆₁, Met₁₇₅, Val₄₄₅, Ile₄₆₈, and Ile₄₇₀.

To further test the prediction that these residues likely form the inside gate, I proposed sites for crosslinking using site-specific introduction of cysteine. For this purpose I chose sites, that would crosslink the interface between (1) helices 4 and 11 (Ser₁₅₈Cys/Leu₄₆₅Cys), (2) between helices 5 and 10 (Met₁₇₆Cys/Met₄₄₂Cys), as well as crosslinks between (3) helices 4 and 5 (Gly₁₆₂Cys/Ser₁₇₃Cys) and (4) helices 10 and 11 (Gly₄₄₇Cys/Gly₄₆₇Cys) (Figure 5-4 A-B). The distances between the C α atoms of the introduced cysteine residues was 5.8 and 7.8 Å, for crosslink 1 and 2, respectively and 12.4 and 13.5 Å for crosslinks 3 and 4. Introduction of these mutations into a cysteine-less background resulted in a nonfunctional transporter and therefore their effect was assessed in the wild-type sequence, which contains five endogenous cysteine residues. For crosslink 3, no activity could be detected even in the wild type background and is therefore no longer discussed.

Notably, all three double cysteine mutant transporters completely lost the ability to catalyze active adenosine transport upon cross-linking with the relatively short reagents MTS-3-MTS (~5 Å in length and flexible; range of S-S distances, ~3–6 Å) (38) or o-PDM (~6 Å in length and rigid; range of S-S distances 7.67–10.47 Å) [267], [268]. However, these crosslinkers also caused significant reduction in transporter activity when used on single cysteine mutants. In contrast, preincubation with the longer length cross-linking reagents p-PDM (~10 Å in length and rigid; range of S-S distances, 9.20–12.29 Å) [267], [268] or MTS-17-O5-MTS (~22 Å in length and flexible; range of S-S distances, ~3–22 Å) [269] strongly inhibited adenosine influx in Δ ldnt1 Δ ldnt2 *L. donovani* parasites transfected with the three double cysteine mutants, S158CTM4loop/L465CTM11 (p-PDM, ~6-fold inhibition; MTS-17-O5-MTS, ~5-fold inhibition), M176CTM5/M442CTM10 (p-PDM, ~28-fold inhibition; MTS-17-O5-MTS, ~20-fold inhibition), and G447CTM10/G467CTM11 (p-PDM, ~8-fold inhibition; MTS-17-O5-MTS, ~8-fold inhibition) double mutant constructs, but had almost no effect on single cysteine mutants (data not shown here, but in [270]). Moreover, analysis of cysteine mutants with p-PDM by western blots, shows shifts of protein mobility after p-PDM treatment, consistent with the distance between the modified residues, but no such shifts in single-cysteine mutants (Figure 5-4). This indicates that these residues are indeed close in the structure of LdNT1.1 in some sampled conformational states.

5.2.4 Discussion

Although the sequence identity between LdNT1.1 and FucP is low (13%), homology models of other proteins with similar or lower degrees of identity, such as the HIV protease modeled on distantly related aspartyl proteases [271], [272], have successfully provided valuable structural models to interrogate the function of those proteins. In the case of the HIV protease model the computational structure was subsequently validated by the crystal structure of a retroviral protease [273]. Indeed, modeling from remote sequence similarities has been justified by “the general observation that protein structures are better conserved through evolution than are the sequences which overlay them” [271]. Furthermore, we have performed site-directed mutagenesis and chemical cross-linking that support this model and define specific residues likely to contribute to the intracellular gate.

The cysteine cross-linking studies also support the identification of the intracellular gate, as residues within predicted gate helices are close enough in the inward-closed conformation (Figure 5-4) to be cross-linked to each other. Because p-PDM cross-links all three cysteine-cysteine pairs, they must be within 9.20–12.29 Å of each other, not far from the predictions of the model. The shorter predicted distances between introduced Cys-158–Cys-465 (5.8 Å) and Cys-176–Cys-442 (7.8 Å) could either indicate that these residues are somewhat farther apart than predicted by the model or that they are able to react with p-PDM as they breathe apart during opening of the gate.

It is noteworthy that the inward-closed and outward-closed conformations are probably not the only two structural states important for function of LdNT1.1. Crystallographic studies have identified “occluded states” of multiple MFS members [274]–[277], intermediate structures in which the substrate is not directly accessible to solute on either the extracellular or intracellular face of the membrane. Such states are likely to exist for LdNT1.1 as well.

5.3 Otub1: Novel mechanism of p53 stabilization

5.3.1 Introduction

The p53 tumor suppressor is one of the major guardians of the cell against cancer. Upon sensing DNA damage or other unfavorable conditions, p53 protects the cell by functioning as a transcription factor that induces apoptosis or cell cycle arrest to stop propagation of the damage. In order to overcome this protection against uncontrolled cell division, tumors have either mutations that inactivate p53 or severely reduce p53 levels. Stabilization of p53 as way to increase its effective cellular concentrations might be an approach to prevent tumor growth [278].

The levels of p53 inside a cell are controlled by many factors. Mouse double minute 2 homolog (Mdm2) is the most important negative regulator. Mdm2 affects p53 by multiple mechanisms: It inactivates p53 by binding, downregulates p53 expression, and has E3-ligase activity causing ubiquitination of p53 and subsequent degradation of p53 by the proteasome [279]. Overexpression of Mdm2 is one of the mechanisms by which cancers reduce p53 levels. Positive

regulation of p53 levels can be achieved by deubiquitination, catalyzed by deubiquitination enzymes. Surprisingly, one of these enzymes, ovarian tumor domain-containing Ub aldehyde-binding protein 1 (Otub1), stabilizes p53 not by deubiquitination but by inhibiting the cognate E2-ligase of Mdm2, UbcH5 [280]. Several other studies have established that Otub1 can bind to other E2-ligases. Most importantly a crystal structure of a complex of a different E2-ligase Ubc13 with Otub1 showed that Otub1 can bind to the ubiquitin-charged E2-ligase (Ubc13~Ubq), thereby blocking access of substrate to the E2-ligase [281]. The Dai laboratory discovered that Otub1 needs to be monoubiquitinated on either residue Lys₅₉ or Lys₁₀₉ in order to stabilize p53 and that monoubiquitination increased binding of Otub1 to UbcH5 [282]. While structure determination using X-ray crystallography or cryo-electron microscopy would be the gold standard to understand the structural basis of the role of monoubiquitination, these techniques are labor and time intensive and are often precluded by limited stability or intrinsic disorder of the sample. Using a specialized program based on the ROSETTA package I was able to demonstrate that monoubiquitination does not stabilize an Otub1-UbcH15 complex similar to the Otub1-Ubc13 complex, but rather involves binding of the Otub1 bound ubiquitin to the “backside” of UbcH15. My prediction was subsequently confirmed by site-directed mutagenesis.

5.3.2 Materials and Methods

All modeling was performed using the UBQ_Gp_LYX-Cterm application of the Rosetta 3.5 suite [283] using standard parameters according to the

documentation. For modeling of the quaternary complex, the crystal structure 4DDI was used. UbcH5b and UbcH5b~Ub were included using the “extra-bodies” option. A total of 500 independent models were created for either Lys₅₉ or Lys₁₀₉ ubiquitination. For modeling of monoubiquitinated Otub1, the crystal structure 2ZFY was used. A total of 200 models were created for either Lys₅₉ or Lys₁₀₉ ubiquitination. The resulting models were clustered, and the highest scoring structure of the most populated cluster was used for further analysis.

5.3.3 Results

How does monoubiquitination promote Otub1-E2 interaction? Biochemical [284] and structural [281], [285] studies have shown that Otub1 preferentially binds to Ub-charged E2 (E2~Ub). Because UbcH5 preferentially binds to monoubiquitinated Otub1 (Otub1-Ub) [282], we reasoned that monoubiquitination at Lys₅₉ or Lys₁₀₉ may stabilize the Otub1-UbcH5 interaction by forming a quadruple molecular complex containing Otub1-Ub-UbcH5~Ub via direct interaction of Otub1-linked Ub with UbcH5~Ub within the Otub1·UbcH5~Ub complex characterized previously [285]. Thus, we modeled the quaternary complex between UbcH5b~Ub and Otub1-Ub using Rosetta software [283]. After extensive sampling of the conformational space available to Ub, we found that Ub linked to Lys₁₀₉ was unable to form a direct interaction with UbcH5b~Ub (Figure 5-5A). Ub linked to Lys₅₉ formed interactions with UbcH5b~Ub in a small subsets of sampled conformations (Figure 5-5B). Because ubiquitination of either Lys₅₉ or Lys₁₀₉ has a similar effect to regulate p53, we suggest that Ub linked to Otub1 does not stabilize Otub1-UbcH5 interaction

through direct interactions within the characterized Otub1·UbcH5~Ub complex [285].

To further understand how monoubiquitination promotes Otub1-UbcH5 interaction, we modeled the structures of monoubiquitinated Otub1 alone. Interestingly, for monoubiquitination at either Lys₁₀₉ or Lys₅₉, the most frequently sampled poses exposed the canonical Ile₄₄ interface of Ub, potentially allowing for the interaction of Ub with the “backside” of UbcH5 against the catalytic Cys known previously for donor Ub binding [286], [287]. When we superimposed the Ub linked to Otub1 with Ub (charged to another UbcH5) bound to the backside of UbcH5b (PDB code 3A33), we found no steric clashes between UbcH5b and Otub1. This suggested that the Ub covalently linked to Otub1 could facilitate the Otub1-UbcH5 interaction through binding to the backside Ub-interacting surface of UbcH5 (Figure 5-5C). This backside interaction of UbcH5 with Ub involves key residues, including Ser₂₂ on UbcH5 and Ile₄₄ on Ub (Figure 5-5C, right panel) [287]. To test this possibility, we first examined whether mutating Ser₂₂ on UbcH5 to Arg (UbcH5Ser₂₂Arg) could affect this backside UbcH5-Ub interaction because this mutation has been shown to disrupt the Ub-UbcH5 interaction [287]. As seen in Figure 5-5C, although wild-type GST-UbcH5 can bind to the monoubiquitinated Otub1, the GST-UbcH5Ser₂₂Arg mutant fails to bind. Similarly, mutating Ser₂₂ to Leu (UbcH5Ser₂₂Leu) also disrupted the binding between UbcH5 and the monoubiquitinated Otub1 (Figure 5-5D). These mutants generate a steric clash with Ub Ile₄₄ because of the larger side-chain of Arg or Leu compared with Ser [287], [288]. Next, we asked whether mutating Ile₄₄ of Ub to

Ala with a smaller side-chain can also affect the binding. The mutant Ub-I44A (UbIle₄₄Ala) can still be used for Otub1 monoubiquitination *in vitro* (Figure 5-5E). Using the reaction mixture containing Otub1 monoubiquitinated with UbIle₄₄Ala (Otub1-UbIle₄₄Ala) for GST pulldown assays, we found that wild-type GST-UbcH5 indeed failed to interact with Otub1-UbIle₄₄Ala (Figure 5-5E, last lane). Interestingly, the Ile₄₄Ala mutation likely relieves the steric clash generated by the Ser₂₂Ile mutation through complementation, resulting in the rescued binding between UbcH5Ser₂₂Leu and Otub1-UbIle₄₄Ala (Figure 5-5E). Together, these results demonstrate that monoubiquitinated Otub1 interacts with the backside Ub-interacting surface of UbcH5 through the covalently linked Ub. This backside binding could potentially cripple the self-assembly of UbcH5~Ub conjugates thought to be critical for ubiquitin transfer to substrates and poly-Ub chain formation [286], [287], thereby suppressing the Ub-conjugating activity of UbcH5.

5.3.4 Discussion

The novel mechanism described here could exist in parallel with a mechanism described previously wherein Otub1 binds to E2~Ub [281], [285] and they act in concert to potentiate the function of Otub1 to suppress E2. Previous structural studies have shown that binding of a free Ub to the distal Ub-binding site on Otub1 causes a conformational change of Otub1 to form an N-terminal Ub-binding helix, allowing for the binding of the donor Ub [281], [285]. This Otub1 binding blocks the interaction of the donor Ub with another E2 and the attack on the thioester bond by an acceptor Ub and also suppresses Ub transfer [287].

Future studies are warranted to address how the two distinct mechanisms could interplay, timely and spatially, and whether the mechanism described here is also regulated by the cellular concentration of free Ub or the molecular ratio of charged versus uncharged E2 molecules. Alternatively, monoubiquitinated Otub1 could bind to Ub-charged E2 to form a quadruple inhibitory complex (Otub1-Ub·E2~Ub) through the backside UbcH5-Ub interaction. This quadruple complex would then form an infinite spiral, as in the case of the self-assembly of E2~Ub conjugates [287] through the donor Ub interaction with Otub1. This complex could, therefore, disrupt the assembly of E2~Ub conjugates and suppress the Ub transfer and efficient polyubiquitination of substrates [286]. In this scenario, it is important to test whether monoubiquitination of Otub1 could result in a conformational change to promote the donor Ub binding to Otub1. Thus, future characterization of the structure of the inhibitory Otub1-Ub-UbcH5~Ub complex would provide further insights into how Otub1 suppresses E2.

5.4 Ribosome-Sec61 translocon: The role of small structured domains in cotranslational secretion

5.4.1 Introduction

Biosynthesis of secreted proteins involves recognition of an amino-terminal signal sequence by the signal recognition particle and subsequent targeting of the ribosome and the nascent chain to the endoplasmic reticulum. There the ribosome binds to the Sec61 channel and the protein is cotranslationally inserted into the ER [289]. While this simple mechanism suggests that only the signal

sequence determines whether a protein is inserted into the ER, experimental evidence has shown that the cargo protein can also influence translocation. For example, interactions of basic residues of the native chain with the translocon complex determines the topology of transmembrane proteins [290] and domains that fold before the ribosome is targeted to the ER can block translocation [291]. During formation of the ribosome-Sec61 complex about 50 to 60 residues of the nascent chain can accumulate in a vestibule formed by the ribosome and Sec61 before being inserted through the Sec61 pore [292]. In order to answer the question whether folding of small cargo domain within this vestibule could influence translocation, the Skach laboratory demonstrated that insertion of a small zinc-finger domain can block translocation in a zinc-dependent manner, even after the ribosome-polypeptide complex has been recruited to the ER [293]. Even though these studies demonstrate that folding of a small domain can influence translocation, they are based on an artificial system. In order to answer the question whether this phenomenon caused evolution to select against small independent folding units in secretory protein right after the signal peptide, I performed a sequence analysis of the almost completely determined human proteome. By annotating domains using HMM sequence profiles of known folded domains stored in available databases I found that no such selection is found and that the cell probably uses other mechanisms to prevent premature folding of small secreted domains.

5.4.2 Material and Methods

Reviewed entries of human proteins (20,266 total) from the UniProt database (<http://www.uniprot.org/>) were parsed into soluble secretory (1,905) and soluble cytosolic and nuclear proteins (8,485) on the basis of their cellular location and signal-peptide annotations. Proteins with annotated transmembrane segments were excluded. The search strings were 'organism:9606 AND reviewed: yes AND annotation:(type: location Cytoplasm OR Nucleus) NOT annotation:(type:transmem) NOT annotation: (type:signal)' for cytosolic and nuclear proteins and 'organism:9606 AND reviewed: yes AND annotation:(type: signal) NOT annotation: (type:transmem)' for secretory proteins.

To determine whether proteins in the two cohorts contained predicted structured domains within their N-terminal regions, UniProt identifiers were mapped to the Superfamily (<http://supfam.org/>) and Gene3D (<http://gene3d.biochem.ucl.ac.uk/>) databases. These databases annotate domains within protein sequences with hidden Markov models on the basis of the structural classification databases SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop/>) and CATH (<http://www.cathdb.info/>). From these annotations, the percentage of proteins with domains within the first 100 residues downstream of the predicted signal-sequence cleavage (secretory proteins) or within 100 residues after the N-terminal methionine (cytosolic and nuclear proteins) was calculated.

5.4.3 Results and Discussion

In this study the zinc-finger domain was selected as the model for analyzing the effect of domain folding during translocation. Although zinc-finger domains are

common in the N-terminal region of cytosolic and nuclear proteins, analysis of the SCOP and CATH structural databases revealed that they are rare in the secretory proteome (Figure 5-6 b). We therefore tested whether the hypothesis that small folding domains are selected against in the first 100 residues in secretory proteins can be verified by a lower frequency of secretory proteins with a small domain encoded within the first 100 residues. Surprisingly and contrary to the hypothesis, small structurally defined (sub)domains are nearly twice as common within the first 100 residues within secreted proteins as are domains of similar size within the first 100 residues of cytosolic and nuclear proteins (Figure 5-6 a). One of the plausible reasons for this higher abundance of small domains within N-terminal region of secretory proteins may arise from the fact that secretory proteins are on average shorter than cytoplasmic ones. Indeed, some of the most common domains found at the N-terminus of secreted proteins are small tightly folded signaling proteins, such as interleukins or small adhesion mediating SCR-like domains (Figure 5-6 b).

This means that protection against premature folding of small proteins during translocation must occur through some other mechanism. As expected, approximately 70% of these secretory domains contain disulfide bonds (Figure 5-6 c), which would stabilize folding only after ER entry or perhaps under conditions of oxidative stress. These findings raise the possibility that folding kinetics or stability may confer unique translocation properties on secretory cargo that are coupled to their cognate signal sequences [294]. However, further work is needed to determine the extent to which protein domains that are less

stable than the zinc finger domain examined in this study may affect the general translocation processes.

5.5 ROS1: New drugs for treatment of small cell lung cancers

5.5.1 Introduction

Receptor tyrosine kinases (RTK) are activated by extracellular ligands, such as growth factors, cytokines, and peptide hormones. Upon binding their respective ligands by their extracellular domains, the signal is passed along their transmembrane helix to the cytosolic kinase domain, which then catalyzes autophosphorylation or phosphorylation of substrates. Since the downstream signaling of RTKs often leads to cell growth and survival pathways, aberrant activation of RTKs by chromosomal rearrangements or point mutations drives many cancers. By using small molecules that specifically inhibit a particular RTK that drives a specific tumor, one can prevent further growth of the cancer or even reduce the cancer burden. Such “targeted therapy” has the advantage of having minimal side effects when compared to classical chemotherapy treatments and often leads to a higher quality of life for the patient while allowing higher therapeutic doses [295]. However, an intrinsic problem of this approach is the emergence of resistance through point mutations, which accumulate in the targeted RTK to prevent the inhibitor from binding or to display activity in spite of inhibitor binding. Therefore, rapid development of new inhibitors that can overcome the resistance acquired by the RTK against the first-line treatment is required [296].

ROS1 is an RTK that is activated by chromosomal rearrangements in a subset of glioblastomas, non-small-cell lung cancers and cholangiocarcinomas [297]–[299]. The close similarity of ROS1 with another RTK, ALK, have lead to clinical trials testing the effect of the ALK inhibitor crizotinib on ROS1-driven cancers [300]. Although the results are promising, a subset of patients acquires point mutations in ROS1 making the kinase domain develop resistance against crizotinib due to point mutations in ROS1. To overcome this problem, the Druker laboratory developed an unbiased screen of potential RTK inhibitors and identified foretinib as a novel ROS1 inhibitor. Furthermore, foretinib maintains its potency against point mutations in ROS1 that confer resistance to crizotinib [301]. Since no structural data for ROS1 was available, I built homology models of ROS1 bound to either crizotinib or foretinib to rationalize the potency of foretinib against crizotinib resistant ROS1.

5.5.2 Materials and Methods

Homology models were created using the MODELLER software package. To create the model of ROS1 in complex with crizotinib the crystal structures of ALK bound to crizotinib (PDB: 2XP2) was used as template. The model of ROS1 bound to foretinib was built using the crystal structures of c-Met bound to foretinib (PDB: 3LQ8) and of ALK R1275Q bound to type-II inhibitor N-(4-chlorophenyl)-5-[(6,7-dimethoxyquinolin-4-yl)oxy]-1,3-benzoxazol-2-amine (PDB: 4FNY) as templates. For each ligand 800 models were created and ranked based on their DOPE scores. The top ten models were inspected and showed consistent results. The top structure was chosen for the structural depiction.

5.5.3 Results and Discussion

A BLAST search of the sequence of the ROS1 kinase domain against the PDB identified ALK as the closest homolog with 50% sequence identity. Since a crystal structure of ALK bound to crizotinib was available, binding of crizotinib to ROS1 was modeled using this structure as a template (PDB: 2XP2). Crizotinib is a type-I inhibitor, meaning it binds only to the ATP binding pocket, but not to the so-called DFG-pocket, which in this structure is occupied by the phenylalanine residue of the DFG-motif (DFG-in). Foretinib is a type-II inhibitor, meaning it also binds to the DFG-pocket. Therefore binding of foretinib is not compatible with a DFG-in conformation of the kinase. In order to overcome this problem, I used a multi-template modeling approach using the crystal structure of ALK bound to another type-II inhibitor (PDB: 4FNY) and the crystal structure of MET bound to foretinib (PDB: 3LQ8) as templates. This is necessary due to the lower sequence identity (40%) of MET to ROS1. By using the approach the drug interaction are more closely modeled by the kinase bound to the right inhibitor (MET-foretinib), while the rest of the structure is modeled using the better template (ALK). Comparison of the resulting models (Figure 5-7) shows that overall the two structures are similar, with the exception of the activation loop that has to acquire different structures due to the variation in binding of a type-I and type-II inhibitor.

We then mapped mutations (Gly₁₉₇₁Glu, Leu₁₉₄₇Arg, Leu₁₉₈₂Phe, Val₂₀₉₈Ile, Cys₂₀₆₀Gly), which were found to confer resistance to crizotinib [301], onto the structures obtained through homology modeling. While the Gly₁₉₇₁Glu mutation

in the N-terminal lobe is located in a loop that is distant from the ligand binding site, both the Leu₁₉₄₇Arg and Leu₁₉₈₂Phe mutations are in proximity of structural features involved in ligand binding (nucleotide binding loop and helix α C, respectively). In the C-terminal lobe, the Val₂₀₉₈Ile mutation is proximal to the DFG motif of the activation loop, while Cys₂₀₆₀Gly is located within the alpha-helical part of the C-terminal lobe. Importantly, mutations of Leu₁₁₅₂ in ALK, a residue homologous to Leu₁₉₈₂ in ROS1, confer crizotinib-resistance [302] and a mutation of Leu₁₁₅₂ was recently identified in a crizotinib-treated NSCLC patient with clinical resistance [303]. Also, mutations in ALK that are proximal to the DFG motif exhibit increased resistance to crizotinib [302].

Since none of the mutations directly affect interactions between the inhibitor and the kinase, they most likely affect drug binding through allosteric interactions. While Leu₁₉₄₇, Leu₁₉₈₂, and Val₂₀₉₈ map closely to the active site and may disturb the active site, residues Gly₁₉₇₁ and Cys₂₀₆₀ are distant and may affect the kinetics of an induced fit structural rearrangement that was recently identified as critical for drug affinity in the kinase proteases Src and Abl [304].

5.6 Conclusion

These examples demonstrate a variety of scenarios where sequence and structure based computational methods were critical to understand the underlying biology. In the first example, structure prediction using homology modeling of LdNT1.1 was the starting point of the study. All experiments were designed to test the proposed model. In the absence of such a structural model many more mutations would need to be screened and would be much harder to interpret. In

the second example predictions of Otub1-Ub structures were used to the first test a hypothesis of the mechanism by which ubiquitination of Otub1 causes p53 stabilization. After the computational model was not consistent with the initial hypothesis, the modeling was used to propose an alternate hypothesis, which was subsequently tested through experimentation. In the third example bioinformatics was used to examine a hypothesis that is not testable by experimentation. The hypothesis, which poses that during evolution, secreted proteins selected against small folded domains at their N-termini, cannot be readily tested by experimentation. However, the combination of publicly available genomic sequences and experimentally determined structures allowed us to address this question. While the results were not consistent with the hypothesis, they provided an alternate explanation for how premature folding of passenger domains might be prohibited. In the last example, computational methods provided a way to interpret data obtained using cell-based approaches. While advancements in high-throughput cell culture allow for rapid testing of compounds and mutations, it is often not always feasible to directly follow them up by structure determination by X-ray crystallography or NMR spectroscopy. In the absence of experimental structural data, modeling is the best way to interpret the results.

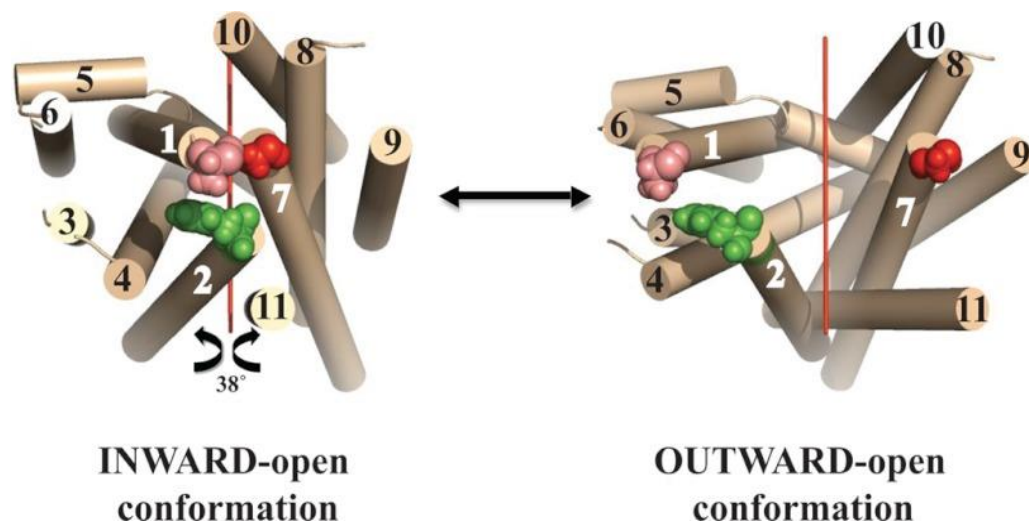


Figure 5-1: Ab initio computational model of LdNT1.1

Left, indicated are residues that were mutated to cysteines. *Tan cylinders* represent predicted TM helices and are numbered 1–11. Residues at the extracellular termini of helices 1, 2, and 7 that were mutated to cysteines are indicated by space filling models: *pink* is Ala-61_{TM1}, *green* is Phe-74_{TM2}, and *red* is Gly-350_{TM7}. The view is from the extracellular surface toward the interior, indicating that the *ab initio* model predicted an inward-open conformation. The figure was generated using PyMol. A suggestive model for the outward-open conformation (*right*) is given by rotating the N-terminal domain (helices 1–6) and the C-terminal domain (helices 7–11) 38° around an axis (*red line*) parallel to the lipid bilayer

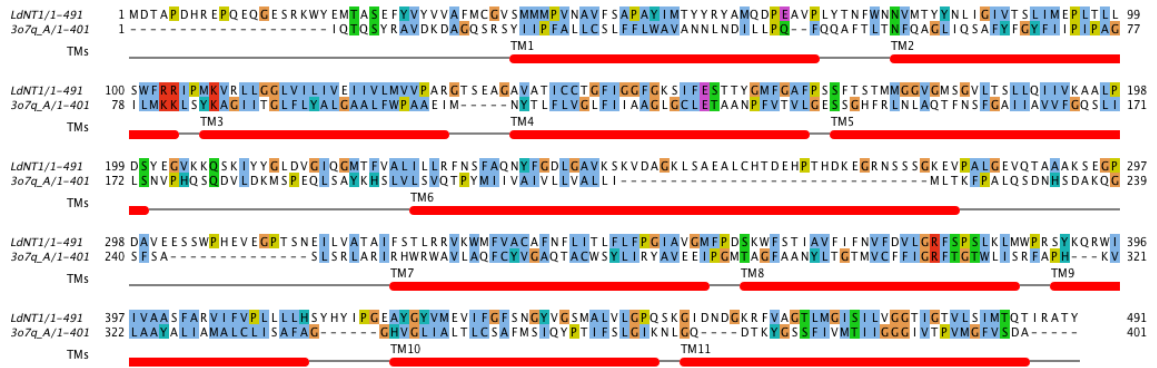


Figure 5-2: Alignment of LdNT1.1 and the *e. coli* sucrose transporter
 Red bars indicate position of transmembrane helix segments in FucP.

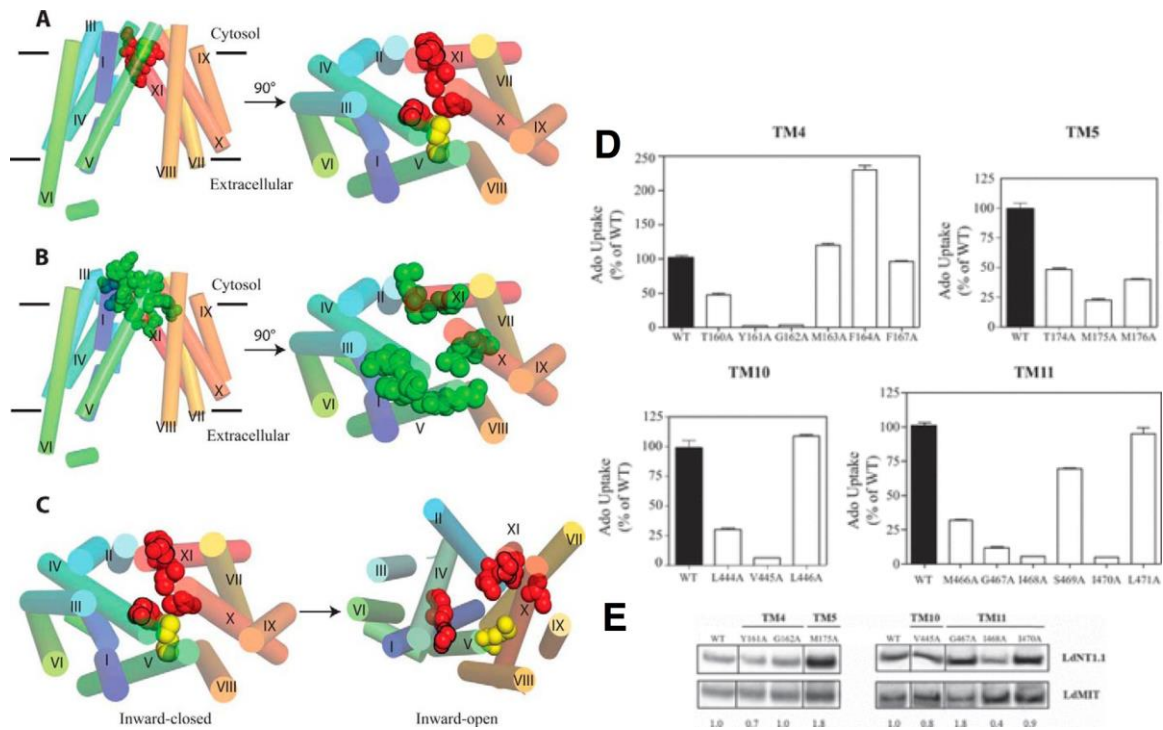


Figure 5-3: Homology model of outward-open state of LdNT1.1 and mutations that affect activity

(A-C) TM helices are indicated as cylinders and are numbered I-XI, and the connecting hydrophilic loops are not shown. Images on the left represent views along an axis parallel to the membrane (inside of the cell on top, outside of the cell on bottom), whereas images on the right are views perpendicular to the membrane from the inside toward the outside. A, six critical residues whose mutation to alanine inhibits adenosine uptake by >90% are indicated as space-filling structures in red. The yellow residue is Met-175, discussed in the text, whose mutation to alanine inhibits uptake by 78%. B, residues at the inner ends of TM helices, whose mutation to alanine either inhibits adenosine uptake more modestly or activates uptake, are shown as space-filling structures in green. C, image of LdNT1.1 in the inward-closed and inward-open conformations showing the inner gate transitioning from closed to open state. The inward-closed image and the space-filling structures are as in A, whereas the inward-open image represents the ab initio model for LdNT1.1 that was reported previously (15).

(D) Uptake of 1 μ M [3 H]adenosine (y axes) was quantified for the Δ nt1/ Δ nt2 double null mutant of *L. donovani* expressing either wild type (WT, filled bars) LdNT1.1 or this transporter with the designated alanine point mutations (open bars). The data are separated according to each TM helix. The level of uptake for WT was set at 100%, and numbers on the y axes represent percent uptake of each mutant relative to wild type LdNT1.1. Each uptake value in this and subsequent figures represents the mean and S.D. (error bars) for at least three independent uptake measurements. (E) Surface expression of WT and each mutant was quantified by surface biotinylation followed by purification of biotinylated proteins on streptavidin beads. The streptavidin bound fractions were separated by SDS-PAGE, blotted, and probed with antibody directed against the NT1-loop VII (top) or the myo-inositol transporter LdMIT (bottom). The numbers under each lane represent the relative surface expression of each alanine point mutant relative to that of WT LdNT1.1 after normalization to the LdMIT signal for each lane.

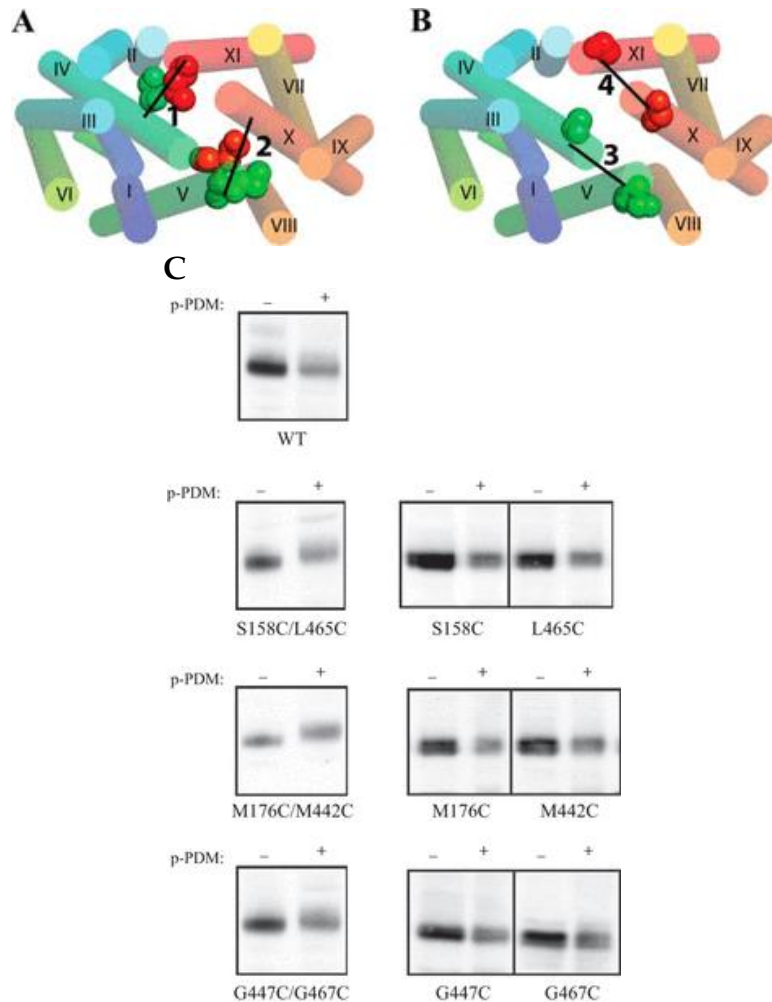
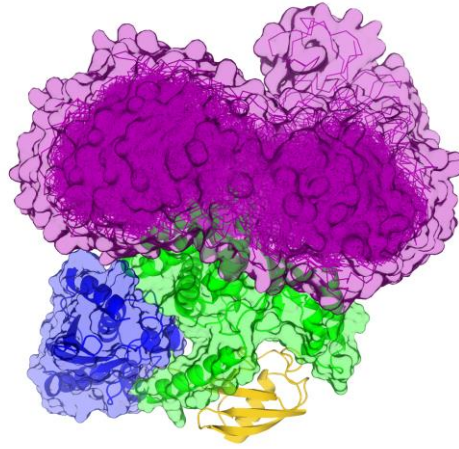


Figure 5-4: Crosslinks based on homology model

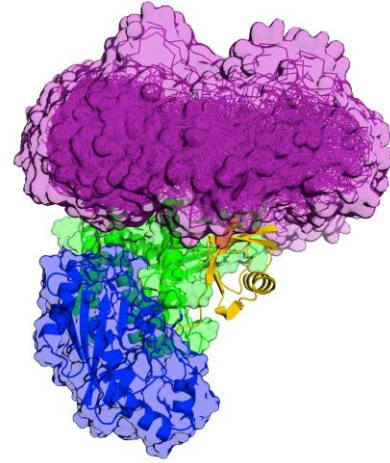
(A-B) For each substituted cysteine, amino acids located in the N-terminal helix bundle (TM1-TM6) are shown as space-filling structures in green, and those in the C-terminal helix bundle (TM7-TM11) are shown in red. The images show the LdNT1.1 inward-closed model seen from the inside toward the outside of the cell. Image A depicts the dual cysteine mutants between TM4 (IV) and TM11 (XI) (S158C/L465C) (labeled 1) and the cysteine pair between TM5 (V) and TM10 (X) (M176C/M442C) (labeled 2). Image B shows the dual cysteine pair between TM4 (IV) and TM5 (V) (G162C/S173C), labeled 3, and that between TM10 (X) and TM11 (XI) (G447C-G467C), labeled 4. The black lines show the interconnections between each cysteine pair.

(C) Membranes were isolated from parasites expressing WT or double cysteine mutant LdNT1.1 cross-linked (+) or not (-) with p-PDM, separated on SDS-PAGE, blotted onto a membrane, and probed with the NT1-loop VII antibody (left panels). Single cysteine mutants were treated identically (right panels).

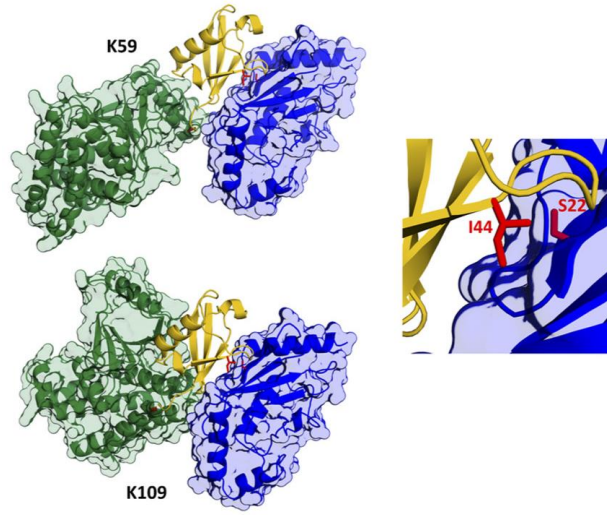
A



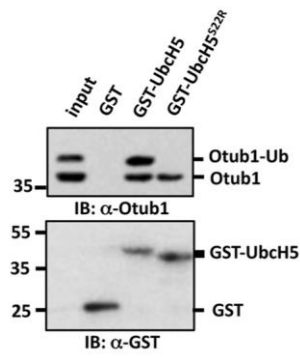
B



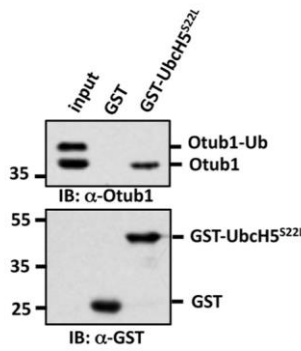
C



D



E



F

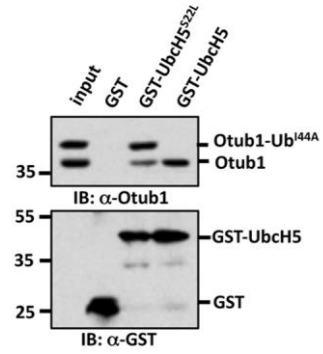


Figure 5-5: Ubch5 binds preferentially to monoubiquitinated Otub1 through backside Ubch5-Ub interaction.

(A and B) Conformational freedom of ubiquitin tether to Lys-59 (A) or Lys-109 (B) in the ternary Oub1-Ub/Ubch5b-Ub complex. Otub1 is shown as a green cartoon with transparent surface, Ubch5b is shown in blue, with the ubiquitin tethered to the active site cysteine shown in yellow. Models for ubiquitin tethered to Otub1 are shown superimposed in purple, with the surface representing the maximal extend of the available conformational space.

(C) Modeling of the Ub-Otub1/Ubch5b interaction through docking of Ub linked to either Lys-59 (top left panel) or Lys-109 (bottom left panel) with the backside of Ubch5b. Otub1, Ubch5b, and Ub are colored green, blue, and yellow, respectively. Lys-59 and Lys-109 of Otub1, Ile-44 of Ub, and Ser-22 of Ubch5 are indicated in red. The enlarged view in the right panel shows the canonical Ub Ile-44 interacting with Ser-22 on Ubch5.

(D and E) Ubch5S22R and Ubch5S22L do not bind to monoubiquitinated Otub1. The *in vitro* ubiquitination reaction mixture containing both monoubiquitinated and unmodified Otub1 was incubated with GST alone, GST-Ubch5c, GST-Ubch5S22R, or GST-Ubch5S22L immobilized onto GSH beads. After washing, bead-bound proteins were assayed by IB analysis.

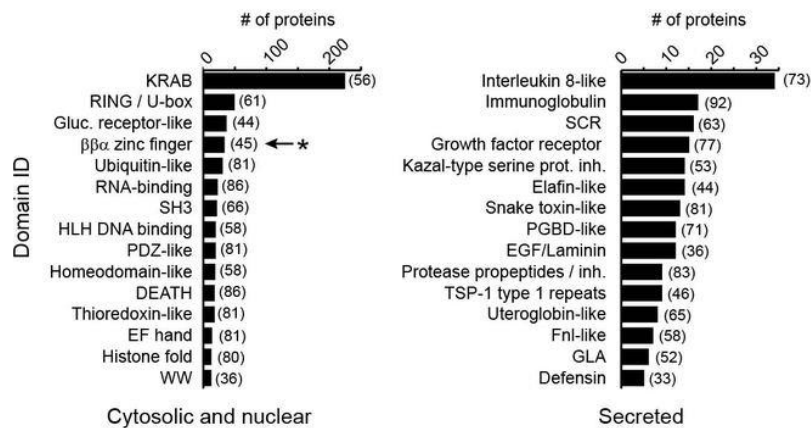
(F) Mutating Ser-22 to Leu rescues the binding defect of Ubch5 with Otub1-UbI44A. His-Otub1 was subjected to an *in vitro* ubiquitination reaction using recombinant UbI44A. The reaction mixture containing both monoubiquitinated (Otub1-UbI44A) and unmodified Otub1 was incubated with GST alone, GST-Ubch5c, or GST-Ubch5S22L immobilized onto GSH beads. After washing, bead-bound proteins were assayed by IB analysis.

a

		CATH database		SCOP database	
		Cytoplasmic and nuclear	Secreted	Cytoplasmic and nuclear	Secreted
Domain within first 100 residues	Domain length				
	< 100 aa	11.2 % (953 / 8485)	19.7 % (375 / 1905)	9.5 % (803 / 8485)	12.5 % (239 / 1905)
	< 50 aa	3.0 % (252 / 8485)	8.3 % (159 / 1905)	1.8 % (154 / 8485)	3.5 % (66 / 1905)

b

15 Most frequent domains within first 100 residues



c

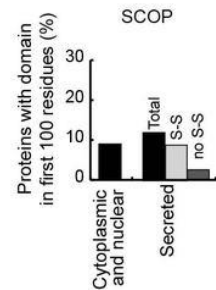


Figure 5-6: Frequency and identity of N-terminal domains in secretory versus cytosolic and nuclear proteins.

a) Table showing percentage of cytosolic and nuclear or secretory proteins that contain structurally defined domains in the first 100 residues from the N-terminus or the first 100 residues beyond the predicted signal sequence based on structural classification via SCOP and CATH databases. Top row includes all domains, whereas bottom row includes only those domains shorter than 50 residues. Actual number of proteins with domains and total proteins are shown in parentheses. Although the number of proteins varies between the databases, both analyses show that discrete domains are more commonly found in N-terminal regions of secretory proteins. b) Identity of the 15 most frequent domains found based on the SCOP database is indicated and plotted as the number of cytosolic and nuclear or secretory proteins that contained the domain. Average domain length is shown in parenthesis. Similar results were obtained using the CATH database, but are not shown since CATH uses numerical identifiers. c) Of the domains located within the first 100 residues of the secretory cohort based on the SCOP database, 70.4% had annotated disulfide bridges in the Uniprot database.

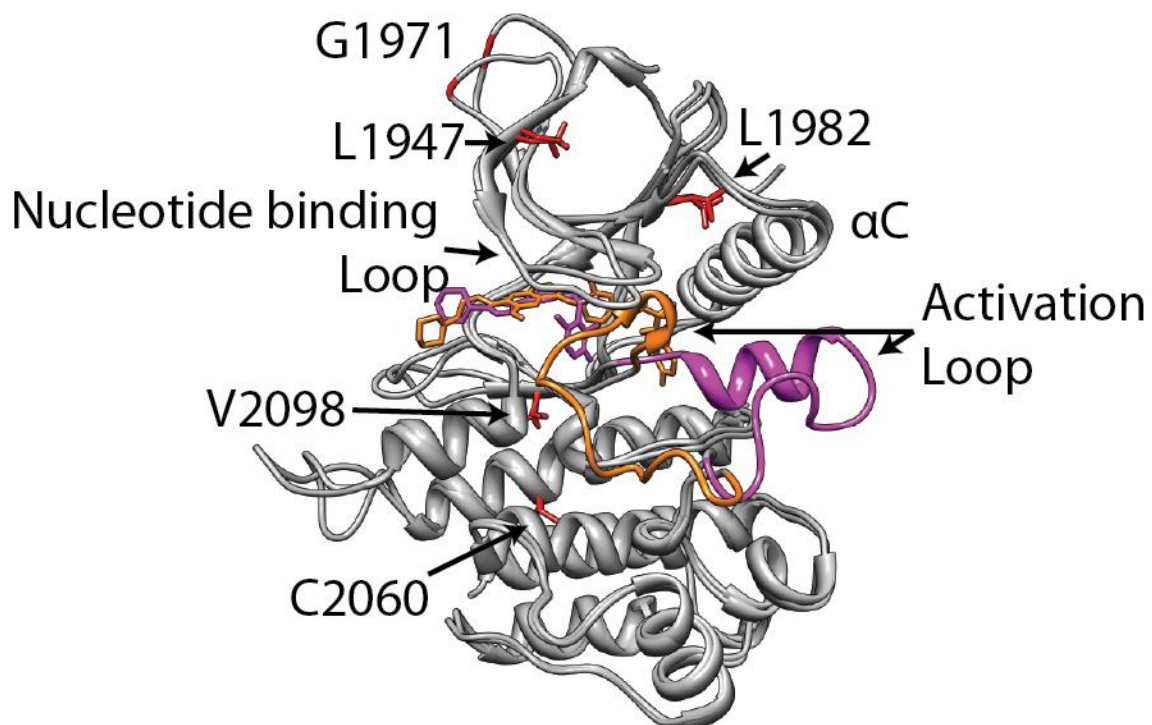


Figure 5-7: Homology models of ROS1 bound to crizotinib and foretinib
 Homology models of ROS1 bound to crizotinib (magenta) and foretinib (orange). The ligands are shown in colored stick representation, while the protein is shown in grey ribbon. Differences in the activation loop conformations are highlighted by coloring them according to the respective ligands. Residues, which were found to confer resistance to crizotinib when mutated, are highlighted in red and shown as stick representation.

Summary and conclusions

I would like to close my thesis by reflecting on the biological problems addressed in this thesis as well as on the approach of using iterations of computational predictions with experimental verification. These iterations are highly synergistic. First computationally generated hypotheses or models, such as the LdNT1 homology model (Chapter 5.2), can be the starting point for experiments testing this hypothesis or vice versa new experimental insights can be the starting point for computational testing of a hypothesis, such as the question whether folding of small domains at the N-terminus of secreted protein blocks translocation (Chapter 5.4). The use of computational methods can drastically increase the insight gained by experiments, for example by testing a hypothesis based on experiments on a single model protein using thousands of protein sequences, as in the histidine bias analysis (Chapter 2), or by simulating an experimentally observed phenomenon. For example the molecular dynamics simulation of the furin propeptide with protonated histidines (Chapter 3.4.4 and 3.4.7) allowed a deeper understanding of the consequences of histidine protonation than what would be experimentally possible. However, almost all computational experiments use simplified representation of the biological systems involved. While these assumptions are often well justified, nevertheless

the value of computational predictions is drastically increased if the predictions are experimentally verifiable. For example the prediction of an alternative binding mode of Otub1 and UbcH5b (Chapter 5.3) does only contribute to our knowledge because we were able to design experiments based on the predicted model to verify its accuracy. In the following sections I will summarize the biological insights that were gained in this thesis and highlight how combination of computation and experiments worked synergistically.

6.1 Histidine protonation in the propeptides of furin and PC1/3 is responsible for pH-mediated activation of cognate catalytic domains

Multiple lines of evidence in this thesis support the idea that histidine protonation in the propeptides of furin and PC1/3 is responsible for the pH-mediated activation of furin and PC1/3. First, swapping the propeptides of furin and PC1/3 exchanges their pH of activation (Figure 3-2) and the secondary structure content of the propeptides is reduced at the pH of activation (Figure 3-2). Both propeptides share a high content of histidine residues (Figure 3-1) and molecular dynamics simulations show that protonation of histidine residues has a strong impact on the structural stability of the furin propeptide (Figure 3-3). We did not observe similar differences for the PC1/3 propeptide, which might be due to the limited timescale of the simulation. In the case of furin one particular histidine, His₆₉, had been previously implicated as a pH-sensor [59]. Consistent with the hypothesis we found that mutation of His₆₉ to leucine increased the overall stability of the propeptide and reduced the amount of secondary

structure loss of the propeptide (Figure 3-4 and Figure 3-5). Furthermore, molecular dynamics simulations showed that mutation of His⁶⁹ to leucine leads to a reduced loss in structure upon histidine protonation, especially in the loop that contains the secondary cleavage site (Figure 3-6).

6.2 Histidine enrichment is an evolutionary mechanism for propeptides of proteases to become pH-sensitive

Even though the high degree of sequence diversity in the propeptides of proprotein convertase made it hard to identify absolutely conserved features in eukaryotic homologs, we found that an increase in histidine content in the propeptide is a unique feature of eukaryotic subtilases (Figure 2-2). The fact that we observed a similar enrichment in the propeptides of a completely unrelated protease family that is also activated by low pH (Figure 2-5), suggests that use of histidine residues in propeptides is an example of convergent evolution. This is supported by the fact that we do not observe such enrichment in a cytosolic protease family (Figure 2-6).

Presumably, introduction of histidine residues is an easy way to evolve pH regulated activation. Propeptides of subtilases are only under a weak evolutionary pressure to maintain their precise structure, compared to the protease domain, which has to maintain the precise arrangement of the catalytic center (Figure 2-1). Furthermore, the structural stability of propeptides appears to be optimized to balance structural integrity so as to maintain inhibition, but be flexible enough to allow cleavage for activation [15]. By inserting a histidine at a position that can accommodate an uncharged imidazole chain, but not a positive

charge, the activation is immediately coupled to pH. So mutation of a single residue is sufficient for a gain of function. Subsequently, introduction of additional histidine residue and fine-tuning of the stability of the propeptide allows for a precise yet facile modulation of the pH of activation.

6.3 Multiple histidine residues are involved in shaping the pH-dependent energy landscape of PRO^{FUR}

Coupling multiple titration sites to the structure of a protein is functionally advantageous, because it allows for modulation of the effective pH of structural loss by altering the protein environment of either titratable site, potentially allowing for quicker adaptation to optimal pH of activation. Most importantly, it allows for a stronger dependence of the structure on proton concentration (See Chapter 1.2.2.2, equation 10). This allows greater pH specificity, because the protein can achieve a stronger perturbation of the equilibrium between structured and unstructured state due to a smaller change in proton concentration.

We found that the propeptide of furin indeed shows a stronger dependence on pH than one would expect from one titrating site (Figure 3-7). Further mutagenesis studies suggest that beside His₆₉, the primary pH-sensor, His₅₂ and His₆₆ could be involved in pH-sensing (Figure 3-8), even though mutation of His₆₆ alone does not impact pH-mediated activation [59]. Data from our laboratory (Williamson, et. al. manuscript in preparation) shows that additional His residues fine-tune the sensitivity of the propeptide of PC1/3 towards pH-dependent activation.

6.4 Differences in pH-activation are due to modulation of the pKa-value of a conserved histidine residue

Protein convertases use pH to direct their activation. It is therefore critical that their propeptides are optimally tuned to confer activation in the right organelle. Measurement of histidine pKa values using histidine hydrogen-deuterium exchange mass spectrometry shows that activation of furin and PC1/3 at different pH values is at least partially due to a difference in the pKa of the primary pH sensor (furin: His₆₉, PC1/3: His₇₂). This histidine is probably critical for pH-mediated activation since it contributes to the structural flexibility of the cleavage loop and also provides a positive charge at the P6 position upon protonation, which is favored by furin [59]. Therefore a shift from a pKa of 6.0 in furin to 5.6 in PC1/3 can explain why PC1/3 requires a higher proton concentration for activation. Comparison of the measured pKa values of His₇₂ in PC1/3 with computational predictions and thermodynamic theory of pH-mediated conformational changes (Figure 4-6) suggests that outward movement of His₇₂ from a pocket formed by the cleavage loop is an important determinant of activation.

6.5 Combination of proteolysis, CID, and ETD fragmentation allows for measurement of residue-specific deuterium uptake in side-chains

Even though the high resolution of an Orbitrap mass spectrometer makes measurement of the effectively sub Dalton mass increase in histidine hydrogen-deuterium uptake possible, an important technical hurdle is the measurement in

a residue specific manner. While in the case of histidine hydrogen-deuterium exchange a residue specific analysis does not require cleavage at every peptide bond, but only between every histidine residue, the high density of histidine residues in the furin and PC1/3 propeptides made this a challenging endeavor. My work demonstrated that this hurdle can be overcome either by a “top-down” combination of using CID and ETD on the undigested propeptide or by combining pepsin digestion with ETD (Figure 4-1 and Figure 4-2). However, the use of pepsin digestion and ETD was much more robust, suggesting that further technique refinement is needed for top-down approaches.

An interesting discovery made during analysis of the exchange data was that residues other than histidine were able to undergo hydrogen-deuterium exchange at long time scales at basic pH (Figure 4-4). By using CID, I was able to measure this exchange in a residue specific manner, my results demonstrate that asparagine and glycine are likely the primary exchanging sites. While backbone scrambling prevents the use of CID in peptide backbone exchange, this is no concern in this case, since the exchange occurs in the side-chain or at the C-alpha atom.

6.6 Computational tools allow rapid testing and generation of biological hypotheses

Several examples in this thesis show how computational tools can provide insight into biological problems. Automated domain annotation using hidden Markov model sequence profiles was used to test two hypotheses: (i) Eukaryotic proteases use histidine residues in their propeptides to sense organellar pH to

direct compartment-specific activation (Chapter 2) and (ii) secreted proteins select against small folded domains at their N-terminus (Chapter 5.4). In both cases HMM sequence profiles allowed me to test a hypothesis, formulated by experiments with a model protein, using thousands of protein sequences. While the results were consistent with the hypothesis in the first case, I was able to reject the latter hypothesis, which implied that secreted proteins must use other mechanisms to prevent premature folding of small N-terminal passenger domains.

Protein structure prediction using homology modeling was employed to obtain a structure of the furin propeptide, which was used to enable both biochemical and *in silico* experiments throughout this thesis. Homology modeling using a distant homolog was used to obtain a structure of the outside-open conformation of LdNT1.1, which enabled the design of experiments that defined the inside gate of LdNT1.1 (Chapter 5.2). Homology modeling was also fundamental to rationalize drug interactions of novel therapeutics against the kinase ROS1 (Chapter 5.5).

Protein structure prediction can be used not only for the prediction of tertiary structure, but also for prediction of quaternary structure. In Chapter 5.3 I used a specialized algorithm employing Monte-Carlo sampling of the conformational freedom of ubiquitin tethered to a lysine residue in OTUB1 to argue against the initial hypothesis that the interaction of OTUB1 to UbcH5b is strengthened by a direct interaction of the ubiquitin tether to OTUB1 with UbcH5b within a previously solved OTUB1/UbcH5b~Ub complex. The modeling

was then used to propose an alternative hypothesis for how mono-ubiquitination of OTUB1 is critical for its function. Based on a previous observation that ubiquitin can bind to the “backside” of UbcH5b I proposed that ubiquitin tethered to OTUB1 interacts with the “backside” of UbcH5b, which explains the increased interaction of monoubiquitinated OTUB1 with UbcH5b. Site-directed mutagenesis based on this model was then used to verify this hypothesis.

Molecular dynamics simulations were used to test the hypothesis that protonation of histidine residues is responsible for the pH-dependent conformational changes observed experimentally (Figure 3-3 and Figure 3-6). In the absence of experimental structures of the propeptides at different pH-values, molecular dynamic simulations and the application of knowledge-based pKa prediction using the available NMR structure of the PC1/3 propeptide (Figure 4-6) is the only way to gain insight into the activation mechanism at the atomic level. Together these data suggest that the unprotonated primary pH-sensor His₆₉/His₇₂ stabilizes the cleavage loop of the propeptide. Protonation causes the histidine to move out of a solvent accessible hydrophobic, so as to destabilize the loop, exposing the second cleavage site within the propeptide for proteolysis, which results in activation.

6.7 Future directions

Several developments in experimental approaches and as well as computational capabilities will drastically increase the use of combinations of experiment and computation. More and more experiments can be performed in a high-throughput manner, which necessitates the use of computational methods to

interpret their results. Moreover, results of these experiments will often be stored in publicly available databases that enable the benchmarking of computational predictions and also allow to test biological hypotheses. Examples of such databases include Proteomics DB [305], which archives results of proteomics experiments of different tissues, the STRING database [306], which archives protein-protein interactions, or PathwayCommons [307], which archives biological pathway information.

While computational approaches, especially simulations, have benefitted from specialized computers to increase both the system sizes and timescales that can be simulated [164], their greatest impact will probably be due to the increasing power of general-purpose computers. While the two-fold increase in computational power of computer chips doubles every 18 months, commonly referred to as Moore's law, seems to slow down [308] we are still seeing a rapid increase in available computational power. This is reflected in ambitious simulation goals that have recently been pursued such as simulation of the human brain [309] or bacterial cells [310].

Furthermore, improvements in the algorithms will increase the accuracy of computational predictions. For example in the field of molecular dynamics new force fields that take polarization effects into account have improved the accuracy of simulations [311]. Another exciting development is the use of machine learning to automatically detect patterns in large biological datasets and to generate predictions [312].

These advancements in computational tools could be used to further understand the pH-mediated activation of proprotein convertases. Many of the experiments described in this thesis were performed on isolated propeptides. A critical step in understanding the activation mechanism is to understand whether interaction of the propeptide with the protease modulates pH-dependent conformational changes and the pKa values of histidines. The fact that the propeptide is a substrate of the protease makes it difficult to perform these kinds of experiments *in vitro*, as the propeptide will be digested as soon as the pH is lowered. It is difficult to block proteolytic degradation *in vitro* without perturbing the interaction of propeptide and protease (data not shown), however, this could be achieved *in silico*. The advancements in timescales that can be simulated and improved force-fields described above, together with the cpHMD techniques described in Chapter 1.2.4 could be used to simulate the behavior of the propeptide in complex with the protease. In order to verify the accuracy of this simulation one could also simulate the propeptide by itself using the same approach and test whether the experimental pKa-values and pH-dependent conformational changes can be reproduced. Furthermore, this computational system could then be used to predict point mutations that should modulate pH-mediated activation *in vivo*, which can subsequently be tested experimentally, thus starting the next iteration of computation and experiment.

Software for analysis of histidine hydrogen-deuterium exchange

A1.1 Introduction

The analysis of the histidine hydrogen-exchange mass spectrometry experiments described in Chapter 4 requires analysis of a large number of liquid-chromatography mass spectrometry data files. Furthermore, conditions for fragmentation of the protein into single-histidine ions must be found. This requires iterations of experiments with modifications of the experimental parameters necessary for fragmentation. For efficient use of the available instrument time, these results must be easily and rapidly interpretable. Although software for the analysis of hydrogen-deuterium exchange already exists [313]–[315], it is optimized for the analysis of amide backbone exchange, which differs significantly in the both the kinetics of the exchange process and the experimental setup involved. Mainly, the back-exchange of deuterium with hydrogen during mass measurement, which is the major concern in backbone exchange experiments, can be almost neglected in histidine hydrogen-deuterium exchange due to the extremely long time scales involved, especially under conditions of acidic pH. Furthermore, the available software for experiments involving backbone hydrogen-deuterium exchange are not optimized to measure

exchange within specific fragments, but rather attempt to measure as many peptide fragments as possible. Peptide identification is either done just by precursor mass or more accurately by relying on proteomics software, like MASCOT or SEQUEST [316], [317]. While these proteomics software packages are very sophisticated and efficient for identification of proteins in complex mixtures they are not optimized to search for specific peptides in the digest of a purified protein.

These shortcomings of the existing software for the analysis of histidine hydrogen-deuterium exchange forced me to develop my own set of scripts that would initially allow for the rapid identification of fragment and peptide ions for optimizing a strategy for fragmentation, and subsequently integrate the many LC-MS experiments at different pH conditions into the desired pKa and k_{\max} values.

A1.2 Overall design

The software consists of three applications. The first two applications are used to support design of a protein fragmentation strategy to measure deuterium uptake in individual histidine residues. The first application, *Fragment_analyzer*, automatically examines MS2 or higher MSn spectra to identify fragments that can be reliably measured and the fragmentation parameters that provide optimal measurements of the isotopic distribution. The second application, *Peptide_analyzer*, is used to analyze proteolytic digests of proteins and is optimized for the detection of the maximal number of peptides derived from digest of a purified protein, in contrast with the commonly used proteomics

software that assumes a complex mixture of proteins. The last application, *Exchange_quantifier*, is used to analyze the data that is obtained by using the derived strategy, to quantify uptake and to obtain pK_a and k_{max} values after the exchange at different pH values has been performed.

All three application are implemented in python using the mass spectrometry library of the mMass application [249]. Parameters are specified in JSON format and output is performed in JSON format [318], which can be visualized using HTML and Javascript templates based on the D3.js library [319].

A1.3 Individual Applications

A1.3.1 Fragment_analyzer

The *Fragment_analyzer* program requires two input files: First the data from the liquid-chromatography mass spectrometry (LC-MS) experiment in the mzml format [320] and a JSON file with options. An example option file is shown in Figure A1-1. The user has to specify the sequence of the protein or peptide that is being fragmented as well as the type of ions that should be searched for (b and y for CID, c or z for ETD). Furthermore the user should provide the range of charge states as well as the approximate elution time. The elution time does not have to be precise, as the program will automatically select the scans with the most intense measurements.

The algorithm will calculate the theoretical isotopic distribution of all possible fragments, depending of the fragmentation method used and the selected range of charge states. It will then average all spectra with the same fragmentation parameters, indicated by the same filter string, that have more

than half-maximal ion intensities. This is a safe approach in case of top-down measurements where only one peak during the chromatography is expected, but in analysis of proteolytic digests care must be taken that not two different precursors with similar m/z values contribute. In this case the parameters allow for selection of an elution window to average only the desired data.

After averaging all spectra are checked for presence of any of the precalculated theoretical isotopic distributions. If at least half of the expected isotopic peaks are present the program will calculate the root mean square difference (RMSD) between the observed isotopic distribution and the expected isotopic distribution after normalization of the isotopic peak with the highest intensity to one.

Figure A1-2 shows an example output of this program. On top the sequence of the precursor is displayed and red and blue brackets indicate the presence of N-terminal or C-terminal fragments, respectively. Bold brackets indicate an RMSD of better than 10%, while thin brackets indicate an RMSD better than 20%. A table below lists for each of the identified fragments the RMSD, ion intensity, and the number of matched peaks. It further lists the charge state and the set of parameters that resulted in the best RMSD. The program also provides plots of the average m/z profile of the isotopic distribution to manually verify that this fragment has a good signal-to-noise ratio. Additionally, for every possible fragment the program provides RMSD, intensity, and number of peak parameters for every charge state and every set of parameters.

A1.3.2 Peptide_analyzer

Similar to the `Fragment_analyzer`, the `Peptide_analyzer` program accepts an `mzml` file and a parameter file as input. The `peptide_analyzer` program is optimized for analysis of standard data-dependent LC-MS experiments, but does not require it. However, the analysis of triggered MS2 spectra will help to confirm correct identification of peptides.

The parameter file (example shown in Figure A1-3) specifies the sequence of the protein as well as parameters that influence the range of peptides that will be searched for. The `"charge_states"` option specifies the minimal and maximal charge states that will be searched for. `"maximal_peptide_length"` specifies the maximal length of peptide that will be generated. Since the application is right now only tested for peptic digests and no enzymatic specificity is assumed, that is the only way to reduce the number of theoretically possible peptides. `"mz_range"` should equal to the m/z -range that is used in the LC-MS experiment. Only peptide ions that fall into this m/z -range will be considered.

The program initially calculates all potential peptides that could be generated from the protein sequences and calculates the theoretical isotopic distributions for all charge states of these peptides according to the above described parameters. Since this is the most computationally expensive part of the program, the result can be saved and reused in subsequent searches against the same protein sequence (option `"buffer_iso_dist"`). After calculation of the theoretical isotopic distributions the program iterates through all collected survey scans and checks for detection of any of the theoretical isotopic

distribution. Detection of the isotopic distribution is verified by calculating the RMSD between theoretical and observed isotopic distribution as in the `fragment_analyzer` application. The “rms_threshold” options provides the RMSD value that will be used to check if the program will consider this isotopic distribution as detected. If a sufficiently low RMSD is detected the intensity is saved together with the elution time of the survey scan to later construct a chromatogram for this ion. Furthermore, the measured m/z spectrum of the ion is saved for averaging in a following step.

After this initial search the collected data is rearranged to calculate for each peptide chromatograms of different charge states and the average of the detected m/z profiles. Furthermore, the program catalogs all MS2 scans that were triggered with a parent mass that could be consistent with this peptide. Since primary identification is just done by peptide mass, and often the peptide mass of multiple peptides is identical the program compiles a list for each peptide ion of other potential peptide ions that could show a similar isotopic distribution.

Figure A1-4 shows screenshots of the visualization of these results. Figure A1-4 A shows an overview of the results that is presented on the top of the page. Below the sequence logo every bar represents a peptide whose isotopic distribution has been detected in at least one scan. The bars are sorted and colored according to the overall intensity. High intensity ions likely are true identification, while low intensity ions are often just chance matches to noise. Every bar is a hyperlink to the detailed matching results of each fragment, which are described in more detail below. Since the number of potential matches can be

overwhelming and in this application one simply searches for peptides that contain specific residues, the application offers a rapid way to filter peptides. Clicking on residues results in display of only peptides that contain this residue, while clicking twice excludes all peptides that contain this residue. This way one can quickly display a list of all peptides that contain for example a unique histidine (Figure A1-4 B).

In the detail view of each peptide three different kinds of information are displayed for each charge state (Figure A1-4 C). First a chromatogram of the observed ion intensity during the liquid chromatography, where dashed lines indicates points at which MS2 scans were triggered against this ion. Second, the averaged m/z profile with indications of the theoretical m/z distribution, similar to the ones displayed in the Fragment_analyzer application. Lastly, a table list all peptides which could show a similar isotopic distribution (at least two isotopic peaks overlap). A click on these peptides links to their detail view, so the user can cross-check which peptide actually fits better. Especially useful to discern between two peptide, that share a similar mass is inspection of the MS2 scans if available. Clicking the dashed bars in the chromatogram will open a MS2 viewer (Figure A1-4 D). Peaks that can be linked to peptide fragments are colored and a calculated Xcorr score [321] allows quantification of the goodness of fit between two peptides with the same parent mass.

A1.3.3 Exchange_quantifier

The previous two applications are used to create a strategy of protein fragmentation that allows measuring of deuterium uptake in single histidine

residues. Once ions for all histidine residues have been identified, the actual experimental can be performed. The results of this experiment are multiple LC-MS data files, three per pH tested in this study. This application uses information about the strategy and the LC-MS files and automatically generates the final pH vs exchange rate plots.

The necessary information is passed to the application in three parameter files. The first (Figure A1-5 A) lists the fragments whose exchange should be quantified. These can be whole peptides or fragmentation ions, in which case the level of fragmentation, fragmentation type and precursor mass must be specified. The next file (Figure A1-5 B) list all mzml files and the conditions from which they were obtained. Most importantly they specify the pH and exchange time after which the measurement was made. Furthermore, control experiments, where the protein was never exposed to deuterium, can be specified. While this is not strictly necessary, it is highly recommended to exclude other chemical modifications that could introduce slightly different isotopic distributions compared to the theoretically derived one. For each condition, multiple files can be listed, which will be treated as replicas. The last file (Figure A1-5 C) assigns which fragments should be used to generate parameters for each histidine residue. Multiple fragments per histidine can be listed in order to have internal controls.

Data analysis is performed in several steps. Initially, the program constructs chromatograms for every fragment in all experiments. Figure A1-6 A shows an example. The chromatograms are normalized between replicas, and the

overall intensity is shown on the right. Replicas are color coded throughout the analysis. This is to verify that every fragment could be reliably detected in all mass spectrometry experiments. For every experiment the m/z profile of each fragment is average between all scans where the fragment had more than half-maximal intensity. The peak intensity module (Figure A1-6 B) shows this profile as well as a zoom-in of the I+1 peak. This is to verify that the mass defect between addition of a neutron to carbon or hydrogen does not result in a splitting of this peak, which would interfere with the analysis here. On the right the results of peak intensity quantification is shown. Peak intensity can be quantified by multiple algorithms. Here either the peak height is used (1) or the integral under the curve (2). In most cases these two methods are extremely consistent, but I found in most cases the integral to be more robust. Notably, the intensities are no longer associated with m/z values but with an integer I, that is 0 for the monoisotopic peaks and then denotes the number of added neutrons.

This array of intensities as a function of added neutrons is then used to quantify the uptake of neutrons at each pH (Figure A1-6 C). Three different algorithms are compared. First the ratio of the monoisotopic peak and the I+1 peak is used as described in [133] ("peak_ratio"), secondly the increase in average mass can be used as in [248] ("av_mass"), and as described in Chapter 4 the uptake can be quantified by fitting a linear combination of the theoretical isotopic distribution and the theoretical isotopic distribution after complete exchange ("fit"). Each of these algorithms requires an isotopic distribution of the fragment with no deuterium uptake, which is either derived from the theoretical

isotopic distribution or a control experiment after incubation in H₂O (“exp_cont”).

Finally the program calculates the rate constant of exchange from the observed uptake and fits equation 14 found in Section 1.2.4.4 to the data in order to obtain pKa and k_{\max} values(Figure A1-6 D).

A1.4 Discussion

The software described here is optimized to first support rapid evaluation of protein fragmentation by proteolytic or in-instrument methods, such as ETD or CID. Later it allows rapid integration of HD-exchange data collected by a thus derived fragmentation strategy into quantified deuterium uptake per histidine and finally the pKa.

The whole implementation is done in python, based on a library written in python with only a few pieces of native code to achieve rapid manipulation of the large amounts of data. That means it is easily portable and able to run on multiple operating systems. The output is performed as a JSON file, which has the advantage of being easy to parse by many platforms. This is especially useful for the visualization of the output in HTML and javascript. This allows rapid implementation of rich and interactive user experience, such as the peptide filtering tool. It also means that the results can be easily sent to customers in a core setting, who do not need specialized software to view the data.

One of the major innovations of this software is the method by which peptides are searched for. Previously, hydrogen-deuterium exchange software either identified fragments just on the parent mass or used identification of a

proteomics software such as Mascot or SEQUEST. These software packages identify peptides based on rapidly calculated cross-correlation scores between MS2 spectra and theoretical MS2 spectra, which are then statistically evaluated using a decoy database. They tend to have large false-negative rates, have limits on the minimal peptide length detectable, and are designed to work with fairly specific proteases, such as trypsin. All these shortcomings are of low concern for proteomics studies, but are critical for hydrogen-deuterium exchange. However, the two main issues that they are optimized to overcome, false positive identification and comparison to a large sequence library, are not as big of a concern when a purified protein is analyzed. Therefore, the approach here uses initially brute-force comparison of the isotopic distribution of all possible peptides to the spectra, and subsequently compiles them into an easily for the user interpretable fashion, which allows later verification of the peptide identification performed by MS2 spectra.

The usage of the RMSD value to compare expected and experimental isotopic distribution used here has no large benefit compared to simple comparison of the peptide mass, since two peptides with a similar mass have similar isotopic distributions. The reason why the RMSD value is used extensively in the software is that precise quantification of the isotopic distribution is critical for calculation of deuterium uptake and therefore restricting potential hits to ions whose isotopic distribution can be precisely measured is an efficient approach.

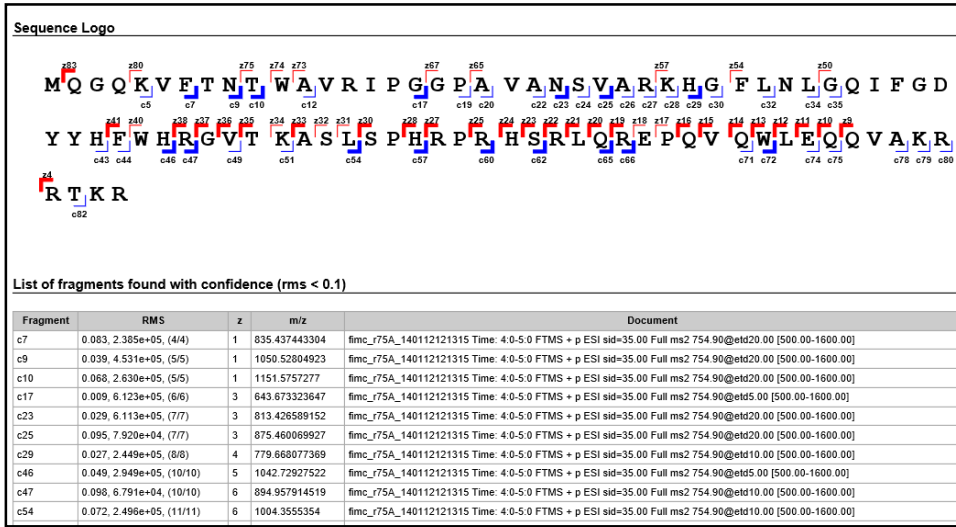
This software can be easily adopted for other experiments where sequence-specific quantification of isotope incorporation at different experimental conditions is needed. For example, it could be used to quantify the percentage of glycosylated or phosphorylated proteins after treatment with glycosidases or phosphatases in H_2^{18}O [322]. Also it has proven very useful in the analysis of the exchange into residues other than histidine that is described in Chapter 4.4.3.

Further development will focus on creating a statistically robust way to automatically verify peptide identification. This score could include the RMSD, the mass error, peptide intensity, the presence of other peptides that were cut at the same sites, and the number of other peptides in the same protein and a decoy database that could create the same score. Furthermore, the output of the results in HTML format suggest the integration of the software into server-based mass-spectrometry analysis, where the storage of data and calculation are performed on a server, requiring just a simple web browser on the user side.

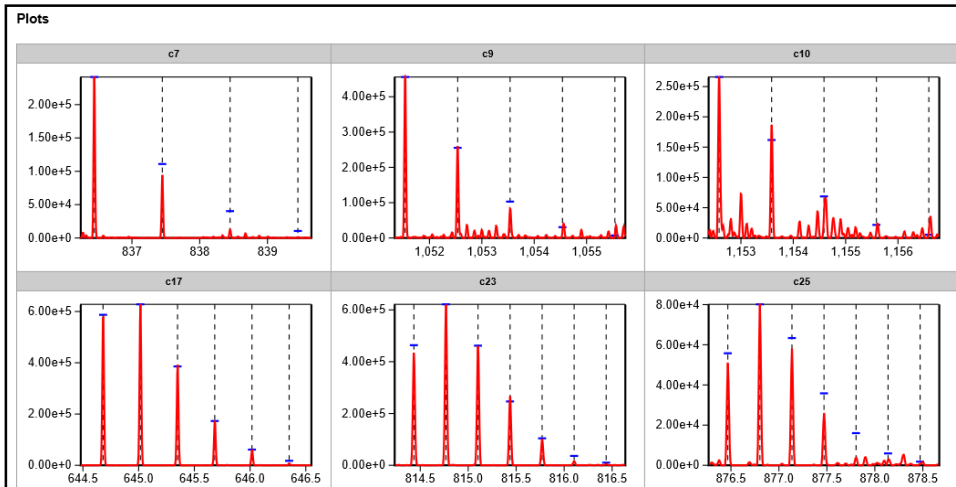

```
{ "sequence" : "MQGQKVFTNTWAVRIPPGVANSAVAVARKHGFLNL",  
  "fragmentseries" : ["c", "z"],  
  "charge_states" : [1, 10],  
  "elution_region" : [400, 800] }
```

Figure A1-1: Parameter file for Fragment_analyzer application

A



B



C

c16

MQGQKVFNTNTWAVRIP

Scan	1	2	3	4	5	6	7	8	9	10	11	12	13
fimc_r75A_140112121315 Time: 4.0-5.0 FTMS + p ESI sid=35.00 Full ms [400.00-1300.00]	0.297, 5.953e+04, (5/6)		0.684, 5.675e+04, (6/6)	0.521, 2.946e+03, (1/6)									
fimc_r75A_140112121315 Time: 4.0-5.0 FTMS + p ESI sid=35.00 Full ms2 754.90@etd5.00 [500.00-1600.00]	0.537, 2.004e+05, (6/6)		0.290, 6.304e+03, (4/6)										
fimc_r75A_140112121315 Time: 4.0-5.0 FTMS + p ESI sid=35.00 Full ms2 754.90@etd10.00 [500.00-1600.00]	0.634, 4.619e+05, (6/6)		0.271, 1.168e+04, (5/6)										
fimc_r75A_140112121315 Time: 4.0-5.0 FTMS + p ESI sid=35.00 Full ms2 754.90@etd20.00 [500.00-1600.00]	0.675, 7.489e+05, (6/6)		0.226, 1.290e+04, (5/6)										

c17

MQGQKVFNTNTWAVRIPG

Scan	1	2	3	4	5	6	7	8	9	10	11	12	13
fimc_r75A_140112121315 Time: 4.0-5.0 FTMS + p ESI sid=35.00 Full ms [400.00-1300.00]	0.295, 6.365e+04, (6/6)	0.437, 6.766e+04, (6/6)											
fimc_r75A_140112121315 Time: 4.0-5.0 FTMS + p ESI sid=35.00 Full ms2 754.90@etd5.00 [500.00-1600.00]	0.223, 2.667e+05, (6/6)	0.009, 6.123e+05, (6/6)											
fimc_r75A_140112121315 Time: 4.0-5.0 FTMS + p ESI sid=35.00 Full ms2 754.90@etd10.00 [500.00-1600.00]	0.117, 4.540e+05, (6/6)	0.009, 1.088e+06, (6/6)											
fimc_r75A_140112121315 Time: 4.0-5.0 FTMS + p ESI sid=35.00 Full ms2 754.90@etd20.00 [500.00-1600.00]	0.135, 5.743e+05, (6/6)	0.032, 1.189e+06, (6/6)											

Figure A1-2: Screenshots of Fragment_analyzer application

(A) Logo and overview (B) Average spectrum of isotopic distribution. Theoretical distribution is shown in dashed lines and blue crossbars (C) Overview of RMSD, intensity, and number of peaks for every charge state with all paramters

```
{ "sequence" : "MQGQKVFTNTWAVRIPPQVANSVAVARKHGFLNL",  
  "sequence_name" : "FIMC",  
  "mz_range" : [200,1800],  
  "charge_states" : [1,10],  
  "max_peptide_length" : 50,  
  "elution_region" : [300,900],  
  "rms_threshold" : 0.15,  
  "buffer_iso_dist" : true }
```

Figure A1-3: Parameter file for Peptide_analyzer application

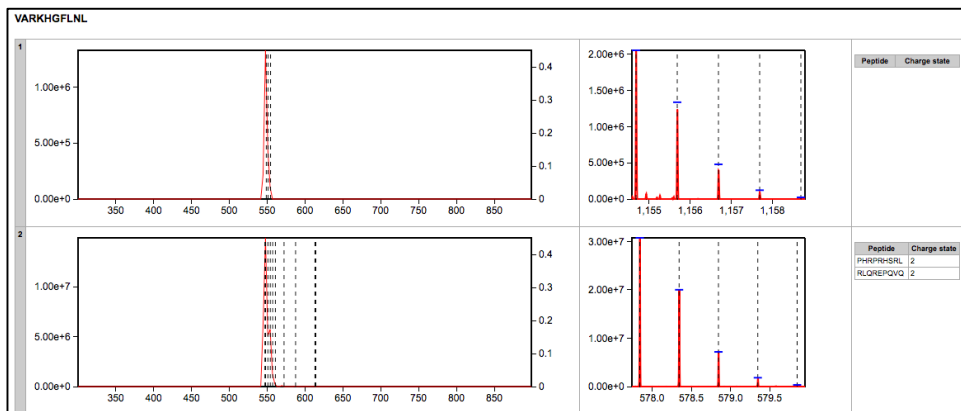
A



B



C



D

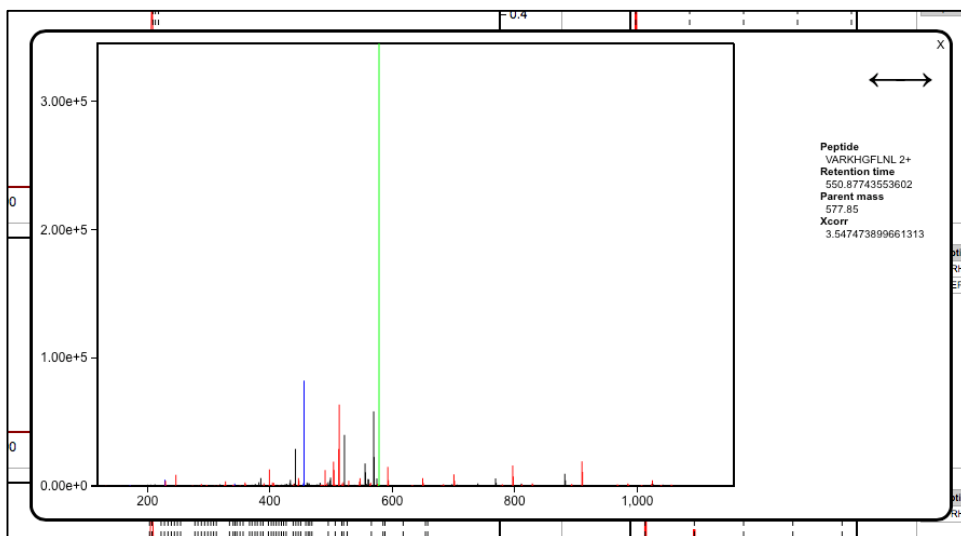


Figure A1-4: Screenshots of Peptide_analyzer application

(A) Sequence with bars indicating potentially identified peptides, sorted and colored by ion intensity. (B) Filtering peptides based on inclusion of specific residues (H66, green) and exclusion of other residues (H52 and H69, red) (C) Detail view for individual peptides. For every charge state a chromatogram, averaged m/z profile and a list of peptides with overlapping isotopic distribution is shown. (D) MS2 viewer can be invoked by clicking one of the dashed lines in (C). Peaks that can be associated to fragment ions are colored.

A

```
[{ "name": "YHF_1",
  "sequence": "YHF",
  "charge": 1,
  "spectrum": { "level": 1},
  "quantify_region": [602,635] },
{ "name": "KRSLSPHRPRHSRL_c7_1",
  "sequence": "KRSLSPHRPRHSRL",
  "fragment": true,
  "fragmentserie": "c",
  "fragmentsite": 7,
  "charge": 1,
  "spectrum": { "level": 2, "precursormz": 433.01 },
  "quantify_region": [465,535]}
```

B

```
[{ "name": "9.0",
  "control" : false,
  "ph": 9.0,
  "time": 72,
  "files" : [ "../9_0.mzML", "../9_0_rep2.mzML", "../9_0_rep3.mzML"
]},
{ "name": "3.5",
  "control" : false,
  "ph": 3.5,
  "time": 72,
  "files" : [ "../3_5.mzML", "../3_5_rep2.mzML",
  "../3_5_rep3.mzML"]},
{ "name": "H2O",
  "control" : true,
  "time": 0,
  "ph": 5.5,
  "files" : [ "../H2O_rep2.mzML", "../H2O_72_rep3.mzML",
  "../H2O.mzML"]}]
```

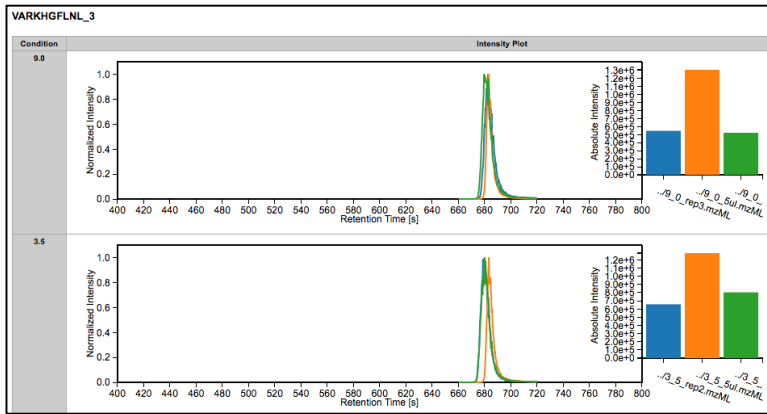
C

```
[{ "name": "H66",
  "fragments": ["YHF_1", "YHFWHRGVT_c2_1"] },
{ "name": "H80",
  "fragments": ["KRSLSPHRPRHSRL_c7_1"] }
]
```

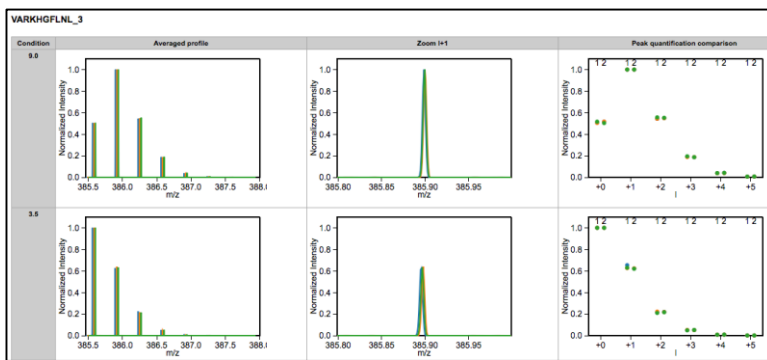
Figure A1-5: Parameter files for Exchange_quantifier application

(A) Fragment parameters (B) Experimental condition parameters (C) Histidine parameters

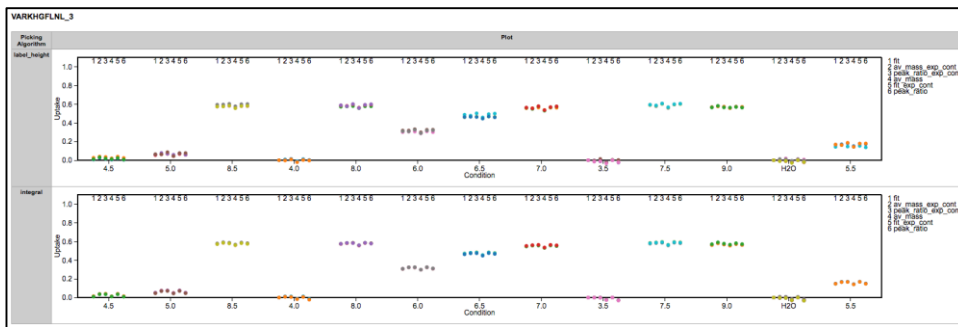
A



B



C



D

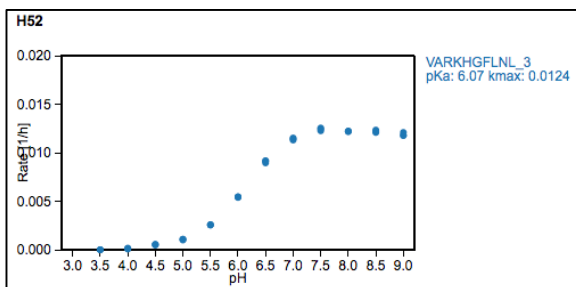


Figure A1-6: Screenshots of Exchange_quantifier application

(A) Chromatogram and intensity of every fragment to be analyzed in all experiments. Every pH condition is shown separately and replicates are superimposed in different colors.

(B) Averaged m/z profiles of fragments under all experimental conditions and results of peak quantification by two algorithms (1: Peak height, 2: Peak integral)

(C) Quantification of deuterium uptake by three different algorithms and two different non-uptake distributions

(D) Plot of exchange rate as a function of pH calculated from fragments according to parameter file.

A2

List of publications

This appendix lists the publications that were generated during my PhD.

A2.1 Publication included in this thesis

Dillon SL, Williamson DM, **Elferich J**, Radler D, Joshi R, Thomas G, Shinde U.

Propeptides are sufficient to regulate organelle-specific pH-dependent activation of furin and proprotein convertase 1/3. *J Mol Biol.* 2012 Oct 12;423(1):47-62.

Elferich J, Williamson DM, Krishnamoorthy B, Shinde U. Propeptides of eukaryotic proteases encode histidines to exploit organelle pH for regulation. *FASEB J.* 2013 Aug;27(8):2939-45.

Williamson DM*, **Elferich J***, Ramakrishnan P, Thomas G, Shinde U. The mechanism by which a propeptide-encoded pH sensor regulates spatiotemporal activation of furin. *J Biol Chem.* 2013 Jun 28;288(26):19154-65.

Davare MA, Saborowski A, Eide CA, Tognon C, Smith RL, **Elferich J**, Agarwal A, Tyner JW, Shinde UP, Lowe SW, Druker BJ. Foretinib is a potent inhibitor of oncogenic ROS1 fusion proteins. *Proc Natl Acad Sci U S A.* 2013 Nov 26;110(48):19519-24.

Li Y, Sun XX, **Elferich J**, Shinde U, David LL, Dai MS. Monoubiquitination is critical for ovarian tumor domain-containing ubiquitin aldehyde binding protein 1 (Oub1) to suppress UbcH5 enzyme and stabilize p53 protein. *J Biol Chem*. 2014 Feb 21;289(8):5097-108.

Valdés R, **Elferich J**, Shinde U, Landfear SM. Identification of the intracellular gate for a member of the equilibrative nucleoside transporter (ENT) family. *J Biol Chem*. 2014 Mar 28;289(13):8799-809.

Conti BJ, **Elferich J**, Yang Z, Shinde U, Skach WR. Cotranslational folding inhibits translocation from within the ribosome-Sec61 translocon complex. *Nat Struct Mol Biol*. 2014 Mar;21(3):228-35. doi: 10.1038/nsmb.2779. Epub 2014 Feb 23. PubMed PMID: 24561504; PubMed Central PMCID: PMC4351553.

Elferich J, Williamson DM, David LL, Shinde U. pKa determination of histidine residues in the propeptides of furin and PC1/3 using histidine hydrogen-deuterium exchange mass spectrometry. *Anal Chem*. 2015. Pending minor revisions.

A2.2 Publications not included in this thesis

Wang H, **Elferich J**, Gouaux E. Structures of LeuT in bicelles define conformation and substrate binding in a membrane-like context. *Nat Struct Mol Biol*. 2012 Jan 15;19(2):212-9.

Dikeakos JD, Thomas L, Kwon G, **Elferich J**, Shinde U, Thomas G. An interdomain binding site on HIV-1 Nef interacts with PACS-1 and PACS-2 on endosomes to down-regulate MHC-I. *Mol Biol Cell*. 2012 Jun;23(11):2184-97.

Nili M, David L, **Elferich J**, Shinde U, Rotwein P. Proteomic analysis and molecular modelling characterize the iron-regulatory protein haemojuvelin/repulsive guidance molecule c. *Biochem J*. 2013 May 15;452(1):87-95.

Tran KD, Rodriguez-Contreras D, Vieira DP, Yates PA, David L, Beatty W, **Elferich J**, Landfear SM. KHARON1 mediates flagellar targeting of a glucose transporter in *Leishmania mexicana* and is critical for viability of infectious intracellular amastigotes. *J Biol Chem*. 2013 Aug 2;288(31):22721-33.

Soysa R, Wilson ZN, **Elferich J**, Forquer I, Shinde U, Riscoe MK, Yates PA, Ullman B. Substrate inhibition of uracil phosphoribosyltransferase by uracil can account for the uracil growth sensitivity of *Leishmania donovani* pyrimidine auxotrophs. *J Biol Chem*. 2013 Oct 11;288(41):29954-64.

Williamson DM, **Elferich J**, Shinde U. Mechanism of Fine-tuning pH Sensors in Proprotein Convertases: Identification of a pH-sensing Histidine Pair in the

Propeptide of Proprotein Convertase 1/3. J Biol Chem. 2015 Jul 30. pii:
jbc.M115.665430.

Literature

- [1] R. J. Siezen and J. A. Leunissen, "Subtilases: the superfamily of subtilisin-like serine proteases.," *Protein Sci.*, vol. 6, no. 3, pp. 501–23, Mar. 1997.
- [2] U. Shinde and G. Thomas, "Insights from bacterial subtilases into the mechanisms of intramolecular chaperone-mediated activation of furin.," *Methods Mol Biol.*, vol. 768, pp. 59–106, Jan. 2011.
- [3] V. Ramírez, A. López, B. Mauch-Mani, M. J. Gil, and P. Vera, "An extracellular subtilase switch for immune priming in Arabidopsis.," *PLoS Pathog.*, vol. 9, no. 6, p. e1003445, Jan. 2013.
- [4] M. J. Blackman, H. Fujioka, W. H. L. Stafford, M. Sajid, B. Clough, S. L. Fleck, M. Aikawa, M. Grainger, and F. Hackett, "A Subtilisin-like Protein in Secretory Organelles of Plasmodium falciparum Merozoites," *J. Biol. Chem.*, vol. 273, no. 36, pp. 23398–23409, Sep. 1998.
- [5] M. C. McNutt, T. A. Lagace, and J. D. Horton, "Catalytic activity is not required for secreted PCSK9 to reduce low density lipoprotein receptors in HepG2 cells.," *J. Biol. Chem.*, vol. 282, no. 29, pp. 20799–803, Jul. 2007.
- [6] S.-I. Tanaka, H. Matsumura, Y. Koga, K. Takano, and S. Kanaya, "Four new crystal structures of Tk-subtilisin in unautoprocesed, autoprocessed and mature forms: insight into structural changes during maturation.," *J. Mol. Biol.*, vol. 372, no. 4, pp. 1055–69, Sep. 2007.
- [7] J. Kraut, "Serine Proteases: Structure and Mechanism of Catalysis," *Annu. Rev. Biochem.*, vol. 46, no. 1, pp. 331–358, Jun. 1977.
- [8] S. Henrich, I. Lindberg, W. Bode, and M. E. Than, "Proprotein convertase models based on the crystal structures of furin and kexin: explanation of their specificity.," *J. Mol. Biol.*, vol. 345, no. 2, pp. 211–27, Jan. 2005.
- [9] S. Henrich, A. Cameron, G. P. Bourenkov, R. Kiefersauer, R. Huber, I. Lindberg, W. Bode, and M. E. Than, "The crystal structure of the proprotein processing proteinase furin explains its stringent specificity.," *Nat. Struct. Biol.*, vol. 10, no. 7, pp. 520–6, Jul. 2003.
- [10] Y. Takeuchi, S. Noguchi, Y. Satow, S. Kojima, I. Kumagai, K. Miura, K. T. Nakamura, and Y. Mitsui, "Molecular recognition at the active site of subtilisin BPN': crystallographic studies using genetically engineered proteinaceous inhibitor SSI (Streptomyces subtilisin inhibitor).," *Protein Eng.*, vol. 4, no. 5, pp. 501–8, Jun. 1991.

- [11] T. Holyoak, C. A. Kettner, G. A. Petsko, R. S. Fuller, and D. Ringe, "Structural basis for differences in substrate selectivity in Kex2 and furin protein convertases," *Biochemistry*, vol. 43, no. 9, pp. 2412–21, Mar. 2004.
- [12] C. Withers-Martinez, M. Strath, F. Hackett, L. F. Haire, S. A. Howell, P. A. Walker, E. Christodoulou, C. Evangelos, G. G. Dodson, and M. J. Blackman, "The malaria parasite egress protease SUB1 is a calcium-dependent redox switch subtilisin," *Nat. Commun.*, vol. 5, p. 3726, Jan. 2014.
- [13] H. Ikemura, H. Takagi, and M. Inouye, "Requirement of pro-sequence for the production of active subtilisin E in *Escherichia coli*," *J. Biol. Chem.*, vol. 262, no. 16, pp. 7859–7864, Jun. 1987.
- [14] U. Shinde, Y. Li, S. Chatterjee, and M. Inouye, "Folding pathway mediated by an intramolecular chaperone," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 15, pp. 6924–8, Aug. 1993.
- [15] E. Subbian, Y. Yabuta, and U. P. Shinde, "Folding pathway mediated by an intramolecular chaperone: intrinsically unstructured propeptide modulates stochastic activation of subtilisin," *J. Mol Biol.*, vol. 347, no. 2, pp. 367–83, Mar. 2005.
- [16] S. C. Jain, U. Shinde, Y. Li, M. Inouye, and H. M. Berman, "The crystal structure of an autoprocessed Ser221Cys-subtilisin E-propeptide complex at 2.0 Å resolution," *J Mol Biol.*, vol. 284, no. 1, pp. 137–144, 1998.
- [17] E. Subbian, Y. Yabuta, and U. P. Shinde, "Folding pathway mediated by an intramolecular chaperone: intrinsically unstructured propeptide modulates stochastic activation of subtilisin," *J Mol Biol.*, vol. 347, no. 2, pp. 367–83, Mar. 2005.
- [18] E. D. Anderson, J. K. VanSlyke, C. D. Thulin, F. Jean, and G. Thomas, "Activation of the furin endoprotease is a multiple-step process: requirements for acidification and internal propeptide cleavage," *Embo J.*, vol. 16, no. 7, pp. 1508–1518, 1997.
- [19] N. G. Seidah, M. S. Sadr, M. Chrétien, and M. Mbikay, "The multifaceted proprotein convertases: Their unique, redundant, complementary, and opposite functions," *J. Biol. Chem.*, vol. 288, no. 30, pp. 21473–21481, 2013.
- [20] E. N. Hampton, M. W. Knuth, J. Li, J. L. Harris, S. A. Lesley, and G. Spraggon, "The self-inhibited structure of full-length PCSK9 at 1.9 Å reveals structural homology with resistin within the C-terminal domain," *Proc. Natl. Acad. Sci.*, vol. 104, no. 37, pp. 14604–14609, Sep. 2007.
- [21] S. Ohagi, J. LaMendola, M. M. LeBeau, R. Espinosa, J. Takeda, S. P. Smeekens, S. J. Chan, and D. F. Steiner, "Identification and analysis of the gene encoding human PC2, a prohormone convertase expressed in neuroendocrine tissues," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 89, no. 11, pp. 4977–81, Jun. 1992.

- [22] E. Jansen, T. A. Ayoubi, S. M. Meulemans, and W. J. Van de Ven, "Neuroendocrine-specific expression of the human prohormone convertase 1 gene. Hormonal regulation of transcription through distinct cAMP response elements.," *J. Biol. Chem.*, vol. 270, no. 25, pp. 15391–7, Jun. 1995.
- [23] S. Benjannet, N. Rondeau, L. Paquet, A. Boudreault, C. Lazure, M. Chrétien, and N. G. Seidah, "Comparative biosynthesis, covalent post-translational modifications and efficiency of prosegment cleavage of the prohormone convertases PC1 and PC2: glycosylation, sulphation and identification of the intracellular site of prosegment cleavage of PC1 and P," *Biochem. J.*, vol. 294 (Pt 3, pp. 735–43, Sep. 1993.
- [24] E. M. Bailyes, D. L. Bennett, and J. C. Hutton, "Proprotein-processing endopeptidases of the insulin secretory granule.," *Enzyme*, vol. 45, no. 5–6, pp. 301–13, Jan. 1991.
- [25] A. Bonic and R. B. Mackin, "Expression, purification, and PC1-mediated processing of human proglucagon, glicentin, and major proglucagon fragment.," *Protein Expr. Purif.*, vol. 28, no. 1, pp. 15–24, Mar. 2003.
- [26] Y. Rouillé, M. Bianchi, J. C. Irminger, and P. A. Halban, "Role of the prohormone convertase PC2 in the processing of proglucagon to glucagon.," *FEBS Lett.*, vol. 413, no. 1, pp. 119–23, Aug. 1997.
- [27] S. Benjannet, N. Rondeau, R. Day, M. Chretien, and N. G. Seidah, "PC1 and PC2 are proprotein convertases capable of cleaving proopiomelanocortin at distinct pairs of basic residues.," *Proc. Natl. Acad. Sci.*, vol. 88, no. 9, pp. 3564–3568, May 1991.
- [28] N. Scamuffa, F. Calvo, M. Chrétien, N. G. Seidah, and A.-M. Khatib, "Proprotein convertases: lessons from knockouts.," *FASEB J.*, vol. 20, no. 12, pp. 1954–63, Oct. 2006.
- [29] I. S. Farooqi, K. Volders, R. Stanhope, R. Heuschkel, A. White, E. Lank, J. Keogh, S. O'Rahilly, and J. W. M. Creemers, "Hyperphagia and Early-Onset Obesity due to a Novel Homozygous Missense Mutation in Prohormone Convertase 1/3," *J. Clin. Endocrinol. Metab.*, vol. 92, no. 9, pp. 3369–3373, Sep. 2007.
- [30] M. Mbikay, H. Tadros, N. Ishida, C. P. Lerner, E. De Lamirande, A. Chen, M. El-Alfy, Y. Clermont, N. G. Seidah, M. Chretien, C. Gagnon, and E. M. Simpson, "Impaired fertility in mice deficient for the testicular germ-cell protease PC4," *Proc. Natl. Acad. Sci.*, vol. 94, no. 13, pp. 6842–6846, Jun. 1997.
- [31] S. Basak, M. Chrétien, M. Mbikay, and A. Basak, "In vitro elucidation of substrate specificity and bioassay of proprotein convertase 4 using intramolecularly quenched fluorogenic peptides.," *Biochem. J.*, vol. 380, no. Pt 2, pp. 505–514, 2004.
- [32] Q. Qiu, A. Basak, M. Mbikay, B. K. Tsang, and A. Gruslin, "Role of pro-IGF-II processing by proprotein convertase 4 in human placental development.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 31, pp. 11047–52, Aug. 2005.

- [33] G. Thomas, "Furin at the cutting edge: from protein traffic to embryogenesis and disease," *Nat. Rev. Mol. Cell Biol.*, vol. 3, no. 10, pp. 753–66, Oct. 2002.
- [34] E. Rousselet, S. Benjannet, J. Hamelin, M. Canuel, and N. G. Seidah, "The proprotein convertase PC7: Unique zymogen activation and trafficking pathways," *J. Biol. Chem.*, vol. 286, no. 4, pp. 2728–2738, 2011.
- [35] N. Nour, G. Mayer, J. S. Mort, A. Salvas, M. Mbikay, C. J. Morrison, C. M. Overall, and N. G. Seidah, "The cysteine-rich domain of the secreted proprotein convertases PC5A and PACE4 functions as a cell surface anchor and interacts with tissue inhibitors of metalloproteinases," *Mol. Biol. Cell*, vol. 16, no. 11, pp. 5215–26, Nov. 2005.
- [36] G. Mayer, J. Hamelin, M. C. Asselin, A. Pasquato, E. Marcinkiewicz, M. Tang, S. Tabibzadeh, and N. G. Seidah, "The regulated cell surface zymogen activation of the proprotein convertase PC5A directs the processing of its secretory substrates," *J. Biol. Chem.*, vol. 283, no. 4, pp. 2373–2384, 2008.
- [37] P. B, M. G, W. W, H. M, D. F, and S. U, "'Shed' furin: mapping of the cleavage determinants and identification of its C-terminus," Mar. 2001.
- [38] K. Mori, S. Kii, A. Tsuji, M. Nagahama, A. Imamaki, K. Hayashi, T. Akamatsu, H. Nagamune, and Y. Matsuda, "A novel human PACE4 isoform, PACE4E is an active processing protease containing a hydrophobic cluster at the carboxy terminus," *J. Biochem.*, vol. 121, no. 5, pp. 941–8, May 1997.
- [39] R. Essalmani, A. Zaid, J. Marcinkiewicz, A. Chamberland, A. Pasquato, N. G. Seidah, and A. Prat, "In vivo functions of the proprotein convertase PC5/6 during mouse development: Gdf11 is a likely substrate," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 15, pp. 5750–5, Apr. 2008.
- [40] D. B. Constam and E. J. Robertson, "SPC4/PACE4 regulates a TGFbeta signaling network during axis formation," *Genes Dev.*, vol. 14, no. 9, pp. 1146–55, May 2000.
- [41] P. Villeneuve, S. Feliciangeli, G. Croissandeau, N. G. Seidah, M. Mbikay, P. Kitabgi, and A. Beaudet, "Altered processing of the neurotensin/neuromedin N precursor in PC2 knock down mice: a biochemical and immunohistochemical study," *J. Neurochem.*, vol. 82, no. 4, pp. 783–93, Aug. 2002.
- [42] A. J. Roebroek, L. Umans, I. G. Pauli, E. J. Robertson, F. van Leuven, W. J. Van de Ven, and D. B. Constam, "Failure of ventral closure and axial rotation in embryos lacking the proprotein convertase Furin," *Development*, vol. 125, no. 24, pp. 4863–4876, 1998.
- [43] D. Susan-Resiga, R. Essalmani, J. Hamelin, M.-C. Asselin, S. Benjannet, A. Chamberland, R. Day, D. Szumska, D. Constam, S. Bhattacharya, A. Prat, and N. G. Seidah, "Furin is the major processing enzyme of the cardiac-specific growth factor bone morphogenetic protein 10," *J. Biol. Chem.*, vol. 286, no. 26, pp. 22785–94, Jul. 2011.

- [44] B. Robertson, J. Moehring, and T. Moehring, "Defective processing of the insulin receptor in an endoprotease-deficient Chinese hamster cell strain is corrected by expression of mouse furin," *J. Biol. Chem.*, vol. 268, no. 32, pp. 24274–24277, Nov. 1993.
- [45] L. J. L. J., S. M. J., B. S., P. F., C. M., M. J., and S. N., "Endoproteolytic processing of integrin pro- α subunits involves the redundant function of furin and proprotein convertase (PC) 5A, but not paired basic amino acid converting enzyme (PACE) 4, PC5B or PC7," Feb. 2000.
- [46] H. Posthaus, C. M. Dubois, M.-H. Laprise, F. Grondin, M. M. Suter, and E. Müller, "Proprotein cleavage of E-cadherin by furin in baculovirus over-expression system: potential role of other convertases in mammalian cells," *FEBS Lett.*, vol. 438, no. 3, pp. 306–310, Nov. 1998.
- [47] H. Sato, T. Kinoshita, T. Takino, K. Nakayama, and M. Seiki, "Activation of a recombinant membrane type 1-matrix metalloproteinase (MT1-MMP) by furin and its interaction with tissue inhibitor of metalloproteinases (TIMP)-2," *FEBS Lett.*, vol. 393, no. 1, pp. 101–104, Sep. 1996.
- [48] E. Louagie, N. A. Taylor, D. Flamez, A. J. M. Roebroek, N. A. Bright, S. Meulemans, R. Quintens, P. L. Herrera, F. Schuit, W. J. M. Van de Ven, and J. W. M. Creemers, "Role of furin in granular acidification in the endocrine pancreas: identification of the V-ATPase subunit Ac45 as a candidate substrate.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 34, pp. 12319–24, Aug. 2008.
- [49] E. Decroly, M. Vandenbranden, J. Ruyschaert, J. Cogniaux, G. Jacob, S. Howard, G. Marshall, A. Kompelli, A. Basak, and F. Jean, "The convertases furin and PC1 can both cleave the human immunodeficiency virus (HIV)-1 envelope glycoprotein gp160 into gp120 (HIV-1 SU) and gp41 (HIV-1 TM)," *J. Biol. Chem.*, vol. 269, no. 16, pp. 12240–12247, Apr. 1994.
- [50] S. S. Molloy, P. A. Bresnahan, S. H. Leppla, K. R. Klimpel, and G. Thomas, "Human furin is a calcium-dependent serine endoprotease that recognizes the sequence Arg-X-X-Arg and efficiently cleaves anthrax toxin protective antigen.," *J. Biol. Chem.*, vol. 267, no. 23, pp. 16396–16402, Aug. 1992.
- [51] A. J. M. Roebroek, N. A. Taylor, E. Louagie, I. Pauli, L. Smeijers, A. Snellinx, A. Lauwers, W. J. M. Van de Ven, D. Hartmann, and J. W. M. Creemers, "Limited redundancy of the proprotein convertase furin in mouse liver.," *J. Biol. Chem.*, vol. 279, no. 51, pp. 53442–50, Dec. 2004.
- [52] D. E. Bassi, H. Mahloogi, R. Lopez De Cicco, and A. Klein-Szanto, "Increased furin activity enhances the malignant phenotype of human head and neck cancer cells.," *Am. J. Pathol.*, vol. 162, no. 2, pp. 439–47, Feb. 2003.
- [53] H. Turpeinen, E. Raitoharju, A. Oksanen, N. Oksala, M. Levula, L.-P. Lyytikäinen, O. Järvinen, J. W. M. Creemers, M. Kähönen, R. Laaksonen, M. Peltto-Huikko, T. Lehtimäki,

and M. Pesu, "Proprotein convertases in human atherosclerotic plaques: the overexpression of FURIN and its substrate cytokines BAFF and APRIL.," *Atherosclerosis*, vol. 219, no. 2, pp. 799–806, Dec. 2011.

- [54] T. Komiyama, J. A. Swanson, and R. S. Fuller, "Protection from anthrax toxin-mediated killing of macrophages by the combined effects of furin inhibitors and chloroquine.," *Antimicrob. Agents Chemother.*, vol. 49, no. 9, pp. 3875–82, Sep. 2005.
- [55] E. D. Anderson, L. Thomas, J. S. Hayflick, and G. Thomas, "Inhibition of HIV-1 gp160-dependent membrane fusion by a furin-directed alpha 1-antitrypsin variant.," *J. Biol. Chem.*, vol. 268, no. 33, pp. 24887–24891, Nov. 1993.
- [56] D. E. Bassi, R. Lopez De Cicco, H. Mahloogi, S. Zucker, G. Thomas, and A. J. Klein-Szanto, "Furin inhibition results in absent or decreased invasiveness and tumorigenicity of human cancer cells.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 18, pp. 10326–31, Aug. 2001.
- [57] J. W. M. Creemers, M. Vey, W. Schafer, T. A. Y. Ayoubi, A. J. M. Roebroek, H.-D. Klenk, W. Garten, and W. J. M. Van de Ven, "Endoproteolytic Cleavage of Its Propeptide Is a Prerequisite for Efficient Transport of Furin Out of the Endoplasmic Reticulum," *J. Biol. Chem.*, vol. 270, no. 6, pp. 2695–2702, Feb. 1995.
- [58] E. D. Anderson, S. S. Molloy, F. Jean, H. Fei, S. Shimamura, and G. Thomas, "The ordered and compartment-specific autoproteolytic removal of the furin intramolecular chaperone is required for enzyme activation," *J Biol Chem.*, vol. 277, no. 15, pp. 12879–12890, 2002.
- [59] S. F. Feliciangeli, L. Thomas, G. K. Scott, E. Subbian, C.-H. Hung, S. S. Molloy, F. Jean, U. Shinde, and G. Thomas, "Identification of a pH sensor in the furin propeptide that regulates enzyme activation.," *J. Biol. Chem.*, vol. 281, no. 23, pp. 16108–16, Jun. 2006.
- [60] G. J. Martens, J. A. Braks, D. W. Eib, Y. Zhou, and I. Lindberg, "The neuroendocrine polypeptide 7B2 is an endogenous inhibitor of prohormone convertase PC2.," *Proc. Natl. Acad. Sci.*, vol. 91, no. 13, pp. 5784–5787, Jun. 1994.
- [61] X. Zhu and I. Lindberg, "7B2 facilitates the maturation of proPC2 in neuroendocrine cells and is required for the expression of enzymatic activity.," *J. Cell Biol.*, vol. 129, no. 6, pp. 1641–50, Jun. 1995.
- [62] N. S. Lamango, E. Apletalina, J. Liu, and I. Lindberg, "The proteolytic maturation of prohormone convertase 2 (PC2) is a pH-driven process.," *Arch. Biochem. Biophys.*, vol. 362, no. 2, pp. 275–282, 1999.
- [63] a Basak, B. B. Touré, C. Lazure, M. Mbikay, M. Chrétien, and N. G. Seidah, "Enzymic characterization in vitro of recombinant proprotein convertase PC4.," *Biochem. J.*, vol. 343 Pt 1, pp. 29–37, 1999.
- [64] J. L. Sohl, S. S. Jaswal, and D. A. Agard, "Unfolded conformations of alpha-lytic protease are more stable than its native state.," *Nature*, vol. 395, no. 6704, pp. 817–9, Oct. 1998.

- [65] D. Rodriguez-Larrea, S. Minning, T. V Borchert, and J. M. Sanchez-Ruiz, "Role of solvation barriers in protein kinetic stability.," *J. Mol. Biol.*, vol. 360, no. 3, pp. 715–24, Jul. 2006.
- [66] S. Pietschmann, M. Fehn, G. Kaulmann, I. Wenz, B. Wiederanders, and K. Schilling, "Foldase function of the cathepsin S proregion is strictly based upon its domain structure.," *Biol. Chem.*, vol. 383, no. 9, pp. 1453–8, Sep. 2002.
- [67] E. D. Anderson, S. S. Molloy, F. Jean, H. Fei, S. Shimamura, and G. Thomas, "The ordered and compartment-specific autoproteolytic removal of the furin intramolecular chaperone is required for enzyme activation.," *J Biol Chem.*, vol. 277, no. 15, pp. 12879–90, Apr. 2002.
- [68] L. Muller, A. Cameron, Y. Fortenberry, E. V. Apletalina, and I. Lindberg, "Processing and sorting of the prohormone convertase 2 propeptide," *J. Biol. Chem.*, vol. 275, no. 50, pp. 39213–39222, 2000.
- [69] U. P. Shinde, J. J. Liu, and M. Inouye, "Protein memory through altered folding mediated by intramolecular chaperones.," *Nature*, vol. 389, no. 6650, pp. 520–2, Oct. 1997.
- [70] Y. Yasuda, T. Tsukuba, K. Okamoto, T. Kadowaki, and K. Yamamoto, "The role of the cathepsin E propeptide in correct folding, maturation and sorting to the endosome," *J. Biochem.*, vol. 138, no. 5, pp. 621–630, 2005.
- [71] Y. Jia, H. Liu, W. Bao, M. Weng, W. Chen, Y. Cai, Z. Zheng, and G. Zou, "Functional analysis of propeptide as an intramolecular chaperone for in vivo folding of subtilisin nattokinase.," *FEBS Lett.*, vol. 584, no. 23, pp. 4789–96, Dec. 2010.
- [72] A. Satomura, M. Nagayama, N. Miura, K. Kuroda, and M. Ueda, "Modification of enzymes by protein folding memory," *FASEB J*, vol. 27, no. 1_MeetingAbstracts, p. 784.1–, Apr. 2013.
- [73] A. Satomura, K. Kuroda, and M. Ueda, "Novel protein engineering of lipase by protein folding memory (567.1)," *FASEB J*, vol. 28, no. 1_Supplement, p. 567.1–, Apr. 2014.
- [74] R. L. Proia and E. Soravia, "Organization of the gene encoding the human beta-hexosaminidase alpha-chain.," *J. Biol. Chem.*, vol. 262, no. 12, pp. 5677–81, Apr. 1987.
- [75] F. S. Hagen, F. J. Grant, J. L. Kuijper, C. A. Slaughter, C. R. Moomaw, K. Orth, P. J. O'Hara, and R. S. Munford, "Expression and characterization of recombinant human acyloxyacyl hydrolase, a leukocyte enzyme that deacylates bacterial lipopolysaccharides.," *Biochemistry*, vol. 30, no. 34, pp. 8415–23, Aug. 1991.
- [76] W. Bode, P. Schwager, and R. Huber, "The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding," *Journal of Molecular Biology*, vol. 118, no. 1, pp. 99–112, 1978.

- [77] Y. Kitamoto, X. Yuan, Q. Wu, D. W. McCourt, and J. E. Sadler, "Enterokinase, the initiator of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains.," *Proc. Natl. Acad. Sci.*, vol. 91, no. 16, pp. 7588–7592, Aug. 1994.
- [78] L. A. Valls, C. P. Hunter, J. H. Rothman, and T. H. Stevens, "Protein sorting in yeast: the localization determinant of yeast vacuolar carboxypeptidase Y resides in the propeptide.," *Cell*, vol. 48, no. 5, pp. 887–897, 1987.
- [79] D. J. Klionsky, L. M. Banta, and S. D. Emr, "Intracellular sorting and processing of a yeast vacuolar hydrolase: proteinase A propeptide contains vacuolar targeting information.," *Mol. Cell. Biol.*, vol. 8, no. 5, pp. 2105–2116, 1988.
- [80] K. Müntener, R. Zwicky, G. Csucs, and A. Baici, "The alternative use of exons 2 and 3 in cathepsin B mRNA controls enzyme trafficking and triggers nuclear fragmentation in human cells.," *Histochem. Cell Biol.*, vol. 119, no. 2, pp. 93–101, Feb. 2003.
- [81] K. Müntener, R. Zwicky, G. Csucs, J. Rohrer, and A. Baici, "Exon skipping of cathepsin B: Mitochondrial targeting of a lysosomal peptidase provokes cell death," *J. Biol. Chem.*, vol. 279, no. 39, pp. 41012–41017, 2004.
- [82] C. Pop and G. S. Salvesen, "Human caspases: Activation, specificity, and regulation," *J. Biol. Chem.*, vol. 284, no. 33, pp. 21777–21781, 2009.
- [83] J. Srivastava, D. L. Barber, and M. P. Jacobson, "Intracellular pH sensors: design principles and functional significance.," *Physiology (Bethesda)*, vol. 22, pp. 30–9, Feb. 2007.
- [84] G. R. Grimsley, J. M. Scholtz, and C. N. Pace, "A summary of the measured pK values of the ionizable groups in folded proteins.," *Protein Sci.*, vol. 18, no. 1, pp. 247–51, Jan. 2009.
- [85] W. F. Boron, "Regulation of intracellular pH.," *Adv. Physiol. Educ.*, vol. 28, no. 1–4, pp. 160–179, 2004.
- [86] S. Matsuyama, J. Llopis, Q. L. Deveraux, R. Y. Tsien, and J. C. Reed, "Changes in intramitochondrial and cytosolic pH: early events that modulate caspase activation during apoptosis.," *Nat. Cell Biol.*, vol. 2, no. 6, pp. 318–25, Jun. 2000.
- [87] R. Schreiber, "Ca²⁺ signaling, intracellular pH and cell volume in cell proliferation.," *J. Membr. Biol.*, vol. 205, no. 3, pp. 129–137, 2005.
- [88] C.-H. Choi, B. A. Webb, M. S. Chimenti, M. P. Jacobson, and D. L. Barber, "pH sensing by FAK-His58 regulates focal adhesion remodeling.," *J. Cell Biol.*, vol. 202, no. 6, pp. 849–59, Sep. 2013.
- [89] G. Miesenböck, D. A. De Angelis, and J. E. Rothman, "Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins.," *Nature*, vol. 394, no. 6689, pp. 192–5, Jul. 1998.

- [90] P. Paroutis, N. Touret, and S. Grinstein, "The pH of the secretory pathway: measurement, determinants, and regulation.," *Physiology (Bethesda)*, vol. 19, no. 4, pp. 207–15, Aug. 2004.
- [91] C. Tanford, "Ionization-linked Changes in Protein Conformation. I. Theory," *J. Am. Chem. Soc.*, vol. 83, no. 7, pp. 1628–1634, Apr. 1961.
- [92] J. V. Kilmartin and L. Rossi-Bernardi, "Interaction of hemoglobin with hydrogen ions, carbon dioxide, and organic phosphates," *Physiol Rev*, vol. 53, no. 4, pp. 836–890, Oct. 1973.
- [93] B. Shaanan, "Structure of human oxyhaemoglobin at 2.1 Å resolution.," *J. Mol. Biol.*, vol. 171, no. 1, pp. 31–59, 1983.
- [94] G. Fermi, M. F. Perutz, B. Shaanan, and R. Fourme, "The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution.," *J. Mol. Biol.*, vol. 175, no. 2, pp. 159–174, 1984.
- [95] J. A. Lukin and C. Ho, "The structure--function relationship of hemoglobin in solution at atomic resolution.," *Chem. Rev.*, vol. 104, no. 3, pp. 1219–30, Mar. 2004.
- [96] L. J. Parkhurst, D. J. Goss, and M. F. Perutz, "Kinetic and equilibrium studies on the role of the .beta.-147 histidine in the Root effect and cooperativity in carp hemoglobin," *Biochemistry*, vol. 22, no. 23, pp. 5401–5409, Nov. 1983.
- [97] M. Berenbrink, "Evolution of vertebrate haemoglobins: Histidine side chains, specific buffer value and Bohr effect.," *Respir. Physiol. Neurobiol.*, vol. 154, no. 1–2, pp. 165–84, Nov. 2006.
- [98] I. M. Yu, W. Zhang, H. A. Holdaway, L. Li, V. A. Kostyuchenko, P. R. Chipman, R. J. Kuhn, M. G. Rossmann, and J. Chen, "Structure of the immature dengue virus at low pH primes proteolytic maturation," *Science (80-.)*, vol. 319, no. 5871, pp. 1834–1837, 2008.
- [99] J. A. T. Young and R. J. Collier, "Anthrax toxin: receptor binding, internalization, pore formation, and translocation.," *Annu. Rev. Biochem.*, vol. 76, pp. 243–65, Jan. 2007.
- [100] R. O. Blaustein, T. M. Koehler, R. J. Collier, and A. Finkelstein, "Anthrax toxin: channel-forming activity of protective antigen in planar phospholipid bilayers.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 86, no. 7, pp. 2209–2213, 1989.
- [101] J. C. Milne and R. J. Collier, "pH-Dependent permeabilization of the plasma membrane of mammalian cells by anthrax protective antigen," *Mol. Microbiol.*, vol. 10, no. 3, pp. 647–653, 1993.
- [102] C. Petosa, R. J. Collier, K. R. Klimpel, S. H. Leppla, and R. C. Liddington, "Crystal structure of the anthrax toxin protective antigen.," *Nature*, vol. 385, no. 6619, pp. 833–838, 1997.

- [103] M. Mourez, M. Yan, D. B. Lacy, L. Dillon, L. Bentsen, A. Marpoe, C. Maurin, E. Hotze, D. Wigelsworth, R.-A. Pimental, J. D. Ballard, R. J. Collier, and R. K. Tweten, "Mapping dominant-negative mutations of anthrax protective antigen by scanning mutagenesis.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 24, pp. 13803–13808, 2003.
- [104] J. Jiang, B. L. Pentelute, R. J. Collier, and Z. H. Zhou, "Atomic structure of anthrax protective antigen pore elucidates toxin translocation.," *Nature*, vol. advance on, Mar. 2015.
- [105] B. A. Krantz, A. Finkelstein, and R. J. Collier, "Protein translocation through the anthrax toxin transmembrane pore is driven by a proton gradient," *J. Mol. Biol.*, vol. 355, no. 5, pp. 968–979, 2006.
- [106] J. F. Andersen, X. D. Ding, C. Balfour, T. K. Shokhireva, D. E. Champagne, F. A. Walker, and W. R. Montfort, "Kinetics and Equilibria in Ligand Binding by Nitrophorins 1–4: Evidence for Stabilization of a Nitric Oxide–Ferriheme Complex through a Ligand-Induced Conformational Trap †," *Biochemistry*, vol. 39, no. 33, pp. 10118–10131, Aug. 2000.
- [107] A. Weichsel, J. F. Andersen, S. A. Roberts, and W. R. Montfort, "Nitric oxide binding to nitrophorin 4 induces complete distal pocket burial.," *Nat. Struct. Biol.*, vol. 7, no. 7, pp. 551–554, 2000.
- [108] J. F. Andersen, A. Weichsel, C. A. Balfour, D. E. Champagne, and W. R. Montfort, "The crystal structure of nitrophorin 4 at 1.5 Å resolution: transport of nitric oxide by a lipocalin-based heme protein.," *Structure*, vol. 6, no. 10, pp. 1315–1327, 1998.
- [109] R. E. Berry, M. N. Shokhirev, A. Y. W. Ho, F. Yang, T. K. Shokhireva, H. Zhang, A. Weichsel, W. R. Montfort, and F. A. Walker, "Effect of mutation of carboxyl side-chain amino acids near the heme on the midpoint potentials and ligand binding constants of nitrophorin 2 and its NO, histamine, and imidazole complexes.," *J. Am. Chem. Soc.*, vol. 131, no. 6, pp. 2313–27, Feb. 2009.
- [110] N. V Di Russo, D. A. Estrin, M. A. Martí, and A. E. Roitberg, "pH-Dependent conformational changes in proteins and their effect on experimental pK(a)s: the case of Nitrophorin 4.," *PLoS Comput. Biol.*, vol. 8, no. 11, p. e1002761, Jan. 2012.
- [111] N. V Di Russo, M. A. Martí, and A. E. Roitberg, "Underlying thermodynamics of pH-dependent allostery.," *J. Phys. Chem. B*, vol. 118, no. 45, pp. 12818–26, Nov. 2014.
- [112] G. J. A. Rainey, D. J. Wigelsworth, P. L. Ryan, H. M. Scobie, R. J. Collier, and J. A. T. Young, "Receptor-specific requirements for anthrax toxin delivery into cells.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 37, pp. 13278–83, Sep. 2005.
- [113] S. Jäger, I. Szundi, J. W. Lewis, T. L. Mah, and D. S. Kliger, "Effects of pH on rhodopsin photointermediates from lumirhodopsin to metarhodopsin II.," *Biochemistry*, vol. 37, no. 19, pp. 6998–7005, May 1998.

- [114] R. E. Hirsch and R. W. Noble, "Intrinsic fluorescence of carp hemoglobin: a study of the R- \rightarrow -T transition," *Biochim. Biophys. Acta*, vol. 914, no. 3, pp. 213–9, Aug. 1987.
- [115] A. WHITE, "Effect of pH on fluorescence of tryosine, tryptophan and related compounds.," *Biochem. J.*, vol. 71, no. 2, pp. 217–20, Feb. 1959.
- [116] A. Barth, "Infrared spectroscopy of proteins.," *Biochim. Biophys. Acta*, vol. 1767, no. 9, pp. 1073–101, Sep. 2007.
- [117] G. J. Thomas, "New structural insights from Raman spectroscopy of proteins and their assemblies.," *Biopolymers*, vol. 67, no. 4–5, pp. 214–25, Jan. 2002.
- [118] K. Gerwert, "Molecular Reaction Mechanisms of Proteins Monitored by Time-Resolved FTIR-Spectroscopy," *Biol. Chem.*, vol. 380, no. 7–8, pp. 931–935, Jan. 1999.
- [119] Q. Wu, F. Li, W. Wang, M. H. Hecht, and T. G. Spiro, "UV Raman monitoring of histidine protonation and H–2H exchange in plastocyanin," *J. Inorg. Biochem.*, vol. 88, no. 3–4, pp. 381–387, Feb. 2002.
- [120] G. Balakrishnan, A. A. Jarzecki, Q. Wu, P. M. Kozlowski, D. Wang, and T. G. Spiro, "Mode Recognition in UV Resonance Raman Spectra of Imidazole: Histidine Monitoring in Proteins," *J. Phys. Chem. B*, vol. 116, no. 31, pp. 9387–9395, Aug. 2012.
- [121] K. W. Hoffman, M. G. Romei, and C. H. Londergan, "A New Raman Spectroscopic Probe of Both the Protonation State and Noncovalent Interactions of Histidine Residues," *J. Phys. Chem. A*, vol. 117, no. 29, pp. 5987–5996, Jul. 2013.
- [122] G. Platzer, M. Okon, and L. P. McIntosh, "pH-dependent random coil (1)H, (13)C, and (15)N chemical shifts of the ionizable amino acids: a guide for protein pK_a measurements.," *J. Biomol. NMR*, vol. 60, no. 2–3, pp. 109–29, Nov. 2014.
- [123] N. Shimba, Z. Serber, R. Ledwidge, S. M. Miller, C. S. Craik, and V. Dötsch, "Quantitative identification of the protonation state of histidines in vitro and in vivo.," *Biochemistry*, vol. 42, no. 30, pp. 9227–34, Aug. 2003.
- [124] N. Baker, N. Baker, M. Holst, M. Holst, F. Wang, and F. Wang, "Adaptive multilevel finite element solution of the Poisson-Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems," *J. Comput. Chem.*, vol. 21, no. 15, pp. 1343–1352, 2000.
- [125] D. Bashford and K. Gerwert, "Electrostatic calculations of the pK_a values of ionizable groups in bacteriorhodopsin.," *J. Mol. Biol.*, vol. 224, no. 2, pp. 473–486, 1992.
- [126] J. Khandogin and C. L. Brooks, "Constant pH molecular dynamics with proton tautomerism.," *Biophys. J.*, vol. 89, no. 1, pp. 141–57, Jul. 2005.
- [127] J. Khandogin and C. L. Brooks, "Toward the accurate first-principles prediction of ionization equilibria in proteins.," *Biochemistry*, vol. 45, no. 31, pp. 9363–73, Aug. 2006.

- [128] M. H. M. Olsson, C. R. SØndergaard, M. Rostkowski, and J. H. Jensen, "PROPKA3: Consistent treatment of internal and surface residues in empirical pK_a predictions," *J. Chem. Theory Comput.*, vol. 7, pp. 525–537, 2011.
- [129] E. Alexov, E. L. Mehler, N. Baker, A. M. Baptista, Y. Huang, F. Milletti, J. Erik Nielsen, D. Farrell, T. Carstensen, M. H. M. Olsson, J. K. Shen, J. Warwicker, S. Williams, and J. M. Word, "Progress in the prediction of pK_a values in proteins," *Proteins: Structure, Function and Bioinformatics*, vol. 79, no. 12, pp. 3260–3275, 2011.
- [130] K. D. Rand, M. Zehl, and T. J. D. Jørgensen, "Measuring the Hydrogen / Deuterium Exchange of Proteins at High Spatial Resolution by Mass Spectrometry : Overcoming Gas-Phase Hydrogen / Deuterium Scrambling," 2014.
- [131] J. H. Bradbury, B. E. Chapman, and F. A. Pellegrino, "Hydrogen-deuterium exchange kinetics of the C-2 protons of imidazole and histidine compounds," *J. Am. Chem. Soc.*, vol. 95, no. 18, pp. 6139–6140, Sep. 1973.
- [132] J. D. Vaughan, Z. Mughrabi, and E. C. Wu, "Kinetics of deuteration of imidazole," *J. Org. Chem.*, vol. 35, no. 4, pp. 1141–1145, Apr. 1970.
- [133] M. Miyagi and T. Nakazawa, "Determination of pK_a values of individual histidine residues in proteins using mass spectrometry," *Anal. Chem.*, vol. 80, no. 17, pp. 6481–6487, Sep. 2008.
- [134] C. B. Anfinsen, "Principles that govern the folding of protein chains.," *Science*, vol. 181, no. 96, pp. 223–30, Jul. 1973.
- [135] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices.," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, 1999.
- [136] M. O. Dayhoff, W. C. Barker, and L. T. Hunt, *Establishing homologies in protein sequences*, vol. 91, no. C. Elsevier, 1983.
- [137] W. R. Pearson, "An introduction to sequence similarity ('homology') searching.," *Curr. Protoc. Bioinformatics*, vol. Chapter 3, p. Unit3.1, Jun. 2013.
- [138] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences.," *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, 1981.
- [139] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970.
- [140] W. Pearson, "Finding protein and nucleotide similarities with FASTA.," *Curr. Protoc. Bioinformatics*, vol. Chapter 3, p. Unit3.9, 2004.

- [141] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.
- [142] Y. Nakamura, G. Cochrane, and I. Karsch-Mizrachi, "The international nucleotide sequence database collaboration," *Nucleic Acids Res.*, vol. 41, no. D1, 2013.
- [143] The UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D71–5, 2012.
- [144] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.," *Nucleic Acids Res.*, vol. 38, no. Web Server issue, pp. W529–33, Jul. 2010.
- [145] S. W. Lockless and R. Ranganathan, "Evolutionarily conserved pathways of energetic connectivity in protein families.," *Science*, vol. 286, no. 5438, pp. 295–299, 1999.
- [146] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, "Protein 3D structure computed from evolutionary sequence variation," *PLoS One*, vol. 6, no. 12, 2011.
- [147] M. J. Harms and J. W. Thornton, "Analyzing protein structure and function using ancestral gene reconstruction.," *Curr. Opin. Struct. Biol.*, vol. 20, no. 3, pp. 360–6, Jun. 2010.
- [148] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bulletin of the American Mathematical Society*, vol. 73, no. 3, pp. 360–364, 1967.
- [149] S. R. Eddy, "Profile hidden Markov models.," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.
- [150] A. J. Viterbi, "A personal history of the Viterbi algorithm," *IEEE Signal Process. Mag.*, vol. 23, no. 4, p. 8, 2005.
- [151] L. R. Welch, "Hidden Markov Models and the Baum-Welch Algorithm," *IEEE Inf. Theory Soc. Newsl.*, vol. 53, no. 4, pp. 1,10–13, 2003.
- [152] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.
- [153] M. Punta, P. C. Coghill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, "The Pfam protein families database," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D290–301, 2012.
- [154] D. Wilson, M. Madera, C. Vogel, C. Chothia, and J. Gough, "The SUPERFAMILY database in 2007: Families and functions," *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, 2007.

- [155] J. G. Lees, D. Lee, R. A. Studer, N. L. Dawson, I. Sillitoe, S. Das, C. Yeats, B. H. Dessailly, R. Rentzsch, and C. A. Orengo, "Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis," *Nucleic Acids Res.*, vol. 42, no. D1, 2014.
- [156] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2004: refinements integrate structure and sequence family data.," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D226–D229, 2004.
- [157] C. A. Orengo, F. M. G. Pearl, and J. M. Thornton, "The CATH domain structure database.," *Methods Biochem. Anal.*, vol. 44, pp. 249–271, 2003.
- [158] F. Dong, B. Olsen, and N. A. Baker, "Computational Methods for Biomolecular Electrostatics," *Methods in Cell Biology*, vol. 84, pp. 843–870, 2008.
- [159] M. Pellegrini-Calace, T. Maiwald, and J. M. Thornton, "PoreWalker: A novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure," *PLoS Comput. Biol.*, vol. 5, no. 7, 2009.
- [160] S. M. Saberi Fathi and J. a Tuszynski, "A simple method for finding a protein's ligand-binding pockets.," *BMC Struct. Biol.*, vol. 14, no. 1, p. 18, 2014.
- [161] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 19, pp. 6679–6685, 2005.
- [162] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *J. Comput. Chem.*, vol. 4, no. 2, pp. 187–217, 1983.
- [163] H. Lei and Y. Duan, "Improved sampling methods for molecular simulation," *Current Opinion in Structural Biology*, vol. 17, no. 2, pp. 187–191, 2007.
- [164] D. E. Shaw, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Lerardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. M. Deneroff, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S. C. Wang, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, and K. J. Bowers, "Anton, a special-purpose machine for molecular dynamics simulation," *Communications of the ACM*, vol. 51, no. 7, p. 91, 2008.
- [165] D. E. Shaw, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. a Moraes, R. O. Dror, S. Piana, Y. Shan, B. Towles, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. a Bank, C. Young, M. M. Deneroff, and B. Batson, "Millisecond-scale molecular dynamics simulations on Anton," *Proc. Conf. High Perform. Comput. Netw. Storage Anal. SC 09*, no. c, p. 1, 2009.
- [166] R. O. Dror, H. F. Green, C. Valant, D. W. Borhani, J. R. Valcourt, A. C. Pan, D. H. Arlow, M. Canals, J. R. Lane, R. Rahmani, J. B. Baell, P. M. Sexton, A. Christopoulos, and D. E.

- Shaw, "Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs," *Nature*, vol. 503, no. 7475, pp. 295–9, 2013.
- [167] G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang, "Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics.," *Nature*, vol. 497, no. 7451, pp. 643–6, 2013.
- [168] U. HANSMANN and Y. OKAMOTO, "New Monte Carlo algorithms for protein folding," *Curr. Opin. Struct. Biol.*, vol. 9, no. 2, pp. 177–183, 1999.
- [169] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y. E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley, "Rosetta3: An object-oriented software suite for the simulation and design of macromolecules," *Methods Enzymol.*, vol. 487, no. C, pp. 545–574, 2011.
- [170] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker, "Protein Structure Prediction Using Rosetta," *Methods in Enzymology*, vol. 383, pp. 66–93, 2004.
- [171] F. Richter, A. Leaver-Fay, S. D. Khare, S. Bjelic, and D. Baker, "De novo enzyme design using Rosetta3," *PLoS One*, vol. 6, no. 5, 2011.
- [172] C. Wang, P. Bradley, and D. Baker, "Protein-Protein Docking with Backbone Flexibility," *J. Mol. Biol.*, vol. 373, no. 2, pp. 503–519, 2007.
- [173] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Atomic-level description of ubiquitin folding.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 15, pp. 5915–20, 2013.
- [174] A. Kryshchuk, K. Fidelis, and J. Moult, "CASP10 results compared to those of previous CASP experiments," *Proteins Struct. Funct. Bioinforma.*, vol. 82, no. SUPPL.2, pp. 164–174, 2014.
- [175] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction.," *Nat. Protoc.*, vol. 5, no. 4, pp. 725–38, Apr. 2010.
- [176] C. H. Tai, H. Bai, T. J. Taylor, and B. Lee, "Assessment of template-free modeling in CASP10 and ROLL," *Proteins Struct. Funct. Bioinforma.*, vol. 82, no. SUPPL.2, pp. 57–83, 2014.
- [177] H. H. Gan, R. A. Perlow, S. Roy, J. Ko, M. Wu, J. Huang, S. Yan, A. Nicoletta, J. Vafai, D. Sun, L. Wang, J. E. Noah, S. Pasquali, and T. Schlick, "Analysis of protein sequence/structure similarity relationships.," *Biophys. J.*, vol. 83, no. 5, pp. 2781–2791, 2002.

- [178] N. Eswar, D. Eramian, B. Webb, M.-Y. Shen, and A. Sali, "Protein structure modeling with MODELLER," *Methods Mol. Biol.*, vol. 426, pp. 145–159, 2008.
- [179] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, no. 5540, pp. 93–96, 2001.
- [180] T. M. Embley and W. Martin, "Eukaryotic evolution, changes and challenges," *Nature*, vol. 440, no. 7084, pp. 623–630, 2006.
- [181] J. R. Casey, S. Grinstein, and J. Orlowski, "Sensors and regulators of intracellular pH," *Nat Rev Mol Cell Biol.*, vol. 11, no. 1, pp. 50–61, 2010.
- [182] C. Lopez-Otin and J. S. Bond, "Proteases: multifunctional enzymes in life and disease," *J Biol Chem.*, vol. 283, no. 45, pp. 30433–30437, 2008.
- [183] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera—a visualization system for exploratory research and analysis," *J Comput Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.
- [184] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanthan, P. D. Thomas, C. H. Wu, C. Yeats, and S. Y. Yong, "InterPro in 2011: new developments in the family and domain prediction database," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D306–12, 2012.
- [185] I. Letunic and P. Bork, "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy," *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W475–8, 2011.
- [186] E. Paradis, J. Claude, and K. Strimmer, "APE: Analyses of Phylogenetics and Evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289–290, 2004.
- [187] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, 2012.
- [188] R. G. Newcombe, "Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: general issues and tail-area-based methods," *Stat Med.*, vol. 25, no. 4, pp. 543–557, 2006.
- [189] S. F. Feliciangeli, L. Thomas, G. K. Scott, E. Subbian, C.-H. Hung, S. S. Molloy, F. Jean, U. Shinde, and G. Thomas, "Identification of a pH sensor in the furin propeptide that regulates enzyme activation," *J Biol Chem.*, vol. 281, no. 23, pp. 16108–16, Jun. 2006.

- [190] S. L. Dillon, D. M. Williamson, J. Elferich, D. Radler, R. Joshi, G. Thomas, and S. Ujwal, "Propeptides are sufficient to regulate organelle-specific pH-dependent activation of furin and proprotein convertase 1/3," *J. Mol. Biol.*, vol. Accepted M, 2012.
- [191] S. L. Dillon, D. M. Williamson, J. Elferich, D. Radler, R. Joshi, G. Thomas, and U. Shinde, "Propeptides Are Sufficient to Regulate Organelle-Specific pH-Dependent Activation of Furin and Proprotein Convertase 1/3," *J Mol Biol.*, vol. 423, no. 1, pp. 47–62, 2012.
- [192] P. Carter and J. A. Wells, "Engineering enzyme specificity by 'substrate-assisted catalysis,'" *Science (80-.)*, vol. 237, no. 4813, pp. 394–399, 1987.
- [193] R. J. Siezen and J. A. Leunissen, "Subtilases: the superfamily of subtilisin-like serine proteases," *Protein Sci.*, vol. 6, no. 3, pp. 501–523, 1997.
- [194] A. Wlodawer, M. G. Li, H. Alla Oyama, B. M. Dunn, and K. Oda, "Structural and enzymatic properties of the sedolisin family of serine-carboxyl peptidases.," *Acta Biochim Pol.*, vol. 50, no. 1, pp. 81–102, 2003.
- [195] K. Oda, M. Sugitani, K. Fukuhara, and S. Murao, "Purification and properties of a pepstatin-insensitive carboxyl proteinase from a gram-negative bacterium," *Biochim Biophys Acta.*, vol. 923, no. 3, pp. 463–469, 1987.
- [196] H. Oyama, T. Hamada, S. Ogasawara, K. Uchida, S. Murao, B. B. Beyer, B. M. Dunn, and K. Oda, "A CLN2-related and thermostable serine-carboxyl proteinase, kumamolysin: cloning, expression, and identification of catalytic serine residue," *J Biochem.*, vol. 131, no. 5, pp. 757–765, 2002.
- [197] F. A. Gunkel and H. G. Gassen, "Proteinase K from *Tritirachium album* Limber. Characterization of the chromosomal gene and expression of the cDNA in *Escherichia coli*," *Eur J Biochem.*, vol. 179, no. 1, pp. 185–194, Jan. 1989.
- [198] N. G. Seidah, S. J. Mowla, J. Hamelin, A. M. Mamarbachi, S. Benjannet, B. B. Toure, A. Basak, J. S. Munzer, J. Marcinkiewicz, M. Zhong, J. C. Barale, C. Lazure, R. A. Murphy, M. Chretien, and M. Marcinkiewicz, "Mammalian subtilisin/kexin isozyme SKI-1: A widely expressed proprotein convertase with a unique cleavage specificity and cellular localization," *Proc Natl Acad Sci U S A.*, vol. 96, no. 4, pp. 1321–1326, 1999.
- [199] V. Turk, V. Stoka, O. Vasiljeva, M. Renko, T. Sun, B. Turk, and D. Turk, "Cysteine cathepsins: from structure, function and regulation to new frontiers," *Biochim Biophys Acta.*, vol. 1824, no. 1, pp. 68–88, 2012.
- [200] B. Turk, I. Dolenc, V. Turk, and J. G. Bieth, "Kinetics of the pH-induced inactivation of human cathepsin L," *Biochemistry*, vol. 32, no. 1, pp. 375–380, 1993.
- [201] Y. Nishimura, T. Kawabata, and K. Kato, "Identification of latent procathepsins B and L in microsomal lumen: characterization of enzymatic activation and proteolytic processing in vitro," *Arch Biochem Biophys.*, vol. 261, no. 1, pp. 64–71, 1988.

- [202] O. Quraishi, D. K. Nägler, T. Fox, J. Sivaraman, M. Cygler, J. S. Mort, and A. C. Storer, "The occluding loop in cathepsin B defines the pH dependence of inhibition by its propeptide," *Biochemistry*, vol. 38, no. 16, pp. 5017–5023, 1999.
- [203] C. Illy, O. Quraishi, J. Wang, E. Purisima, T. Vernet, and J. S. Mort, "Role of the occluding loop in cathepsin B activity," *J Biol Chem.*, vol. 272, no. 2, pp. 1197–1202, 1997.
- [204] E. M. Creagh, H. Conroy, and S. J. Martin, "Caspase-activation pathways in apoptosis and immunity," *Immunol Rev.*, vol. 193, pp. 10–21, 2003.
- [205] L. Tsiatsiani, F. Van Breusegem, P. Gallois, A. Zavialov, E. Lam, and P. V Bozhkov, "Metacaspases," *Cell Death Differ*, vol. 18, no. 8, pp. 1279–1288, 2011.
- [206] C. Degnin, F. Jean, G. Thomas, and J. L. Christian, "Cleavages within the prodomain direct intracellular trafficking and degradation of mature bone morphogenetic protein-4," *Mol Biol Cell.*, vol. 15, no. 11, pp. 5012–20, Nov. 2004.
- [207] N. G. Seidah and A. Prat, "The biology and therapeutic targeting of the proprotein convertases," *Nat Rev Drug Discov.*, vol. 11, no. 5, pp. 367–83, May 2012.
- [208] B. A. Webb, M. Chimenti, M. P. Jacobson, and D. L. Barber, "Dysregulated pH: a perfect storm for cancer progression," *Nat. Rev. Cancer*, vol. 11, no. 9, pp. 671–677, 2011.
- [209] M. Naghavi, R. John, S. Naguib, M. S. Siadaty, R. Grasu, K. C. Kurian, W. B. van Winkle, B. Soller, S. Litovsky, M. Madjid, J. T. Willerson, and W. Casscells, "pH Heterogeneity of human and rabbit atherosclerotic plaques; a new insight into detection of vulnerable plaque," *Atherosclerosis*, vol. 164, no. 1, pp. 27–35, 2002.
- [210] N. G. Seidah, G. Mayer, A. Zaid, E. Rousselet, N. Nassoury, S. Poirier, R. Essalmani, and A. Prat, "The activation and physiological functions of the proprotein convertases," *Int. J. Biochem. Cell Biol.*, vol. 40, no. 6–7, pp. 1111–1125, 2008.
- [211] K. Nakayama, "Furin: a mammalian subtilisin/Kex2p-like endoprotease involved in processing of a wide variety of precursor proteins," *Biochem. J.*, vol. 327 (Pt 3, pp. 625–635, 1997.
- [212] E. Subbian, Y. Yabuta, and U. Shinde, "Positive selection dictates the choice between kinetic and thermodynamic protein folding and stability in subtilases," *Biochemistry*, vol. 43, no. 45, pp. 14348–60, Nov. 2004.
- [213] M. Soskine and D. S. Tawfik, "Mutational effects and the evolution of new protein functions," *Nat. Rev. Genet.*, vol. 11, no. 8, pp. 572–582, Aug. 2010.
- [214] X. Fu, M. Inouye, and U. Shinde, "Folding pathway mediated by an intramolecular chaperone. The inhibitory and chaperone functions of the subtilisin propeptide are not obligatorily linked," *J. Biol. Chem.*, vol. 275, no. 22, pp. 16871–16878, 2000.

- [215] M. A. Tangrea, P. N. Bryan, N. Sari, and J. Orban, "Solution structure of the pro-hormone convertase 1 pro-domain from *Mus musculus*," *J Mol Biol.*, vol. 320, no. 4, pp. 801–12, Jul. 2002.
- [216] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1781–1802, 2005.
- [217] U. Shinde and M. Inouye, "Intramolecular chaperones: polypeptide extensions that modulate protein folding," *Semin. Cell Dev. Biol.*, vol. 11, no. 1, pp. 35–44, Feb. 2000.
- [218] Y. J. Chen and M. Inouye, "The intramolecular chaperone-mediated protein folding," *Current Opinion in Structural Biology*, vol. 18, no. 6, pp. 765–770, 2008.
- [219] S. Bhattacharjya, P. Xu, H. Xiang, M. Chrétien, N. G. Seidah, and F. Ni, "pH-induced conformational transitions of a molten-globule-like state of the inhibitory prodomain of furin: implications for zymogen activation," *Protein Sci.*, vol. 10, no. 5, pp. 934–42, May 2001.
- [220] A. L. Fink, L. J. Calciano, Y. Goto, T. Kurotsu, and D. R. Palleros, "Classification of acid denaturation of proteins: intermediates and unfolded states," *Biochemistry*, vol. 33, no. 41, pp. 12504–12511, 1994.
- [221] Y. Goto, N. Takahashi, and A. L. Fink, "Mechanism of acid-induced folding of proteins," *Biochemistry*, vol. 29, no. 14, pp. 3480–3488, 1990.
- [222] V. N. Uversky and Y. Goto, "Acid denaturation and anion-induced folding of globular proteins: multitude of equilibrium partially folded intermediates," *Curr. Protein Pept. Sci.*, vol. 10, no. 5, pp. 447–455, 2009.
- [223] P. E. Wright and H. J. Dyson, "Linking folding and binding," *Current Opinion in Structural Biology*, vol. 19, no. 1, pp. 31–38, 2009.
- [224] V. Daggett and M. Levitt, "A model of the molten globule state from molecular dynamics simulations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 89, no. 11, pp. 5142–5146, 1992.
- [225] V. Daggett and M. Levitt, "Protein unfolding pathways explored through molecular dynamics simulations," *J. Mol. Biol.*, vol. 232, no. 2, pp. 600–619, 1993.
- [226] C. L. Brooks, "Characterization of 'native' apomyoglobin by molecular dynamics simulation," *J. Mol. Biol.*, vol. 227, no. 2, pp. 375–380, 1992.
- [227] N. L. Salimi, B. Ho, and D. a Agard, "Unfolding simulations reveal the mechanism of extreme unfolding cooperativity in the kinetically stable alpha-lytic protease," *PLoS Comput. Biol.*, vol. 6, no. 2, p. e1000689, Feb. 2010.

- [228] W. Huang, A. P. Eichenberger, and W. F. van Gunsteren, "Molecular dynamics simulation of thionated hen egg white lysozyme.," *Protein Sci.*, vol. 21, no. 8, pp. 1153–61, 2012.
- [229] N. J. Greenfield, "Using circular dichroism spectra to estimate protein secondary structure.," *Nat. Protoc.*, vol. 1, no. 6, pp. 2876–2890, 2006.
- [230] N. J. Greenfield and N. J. Greenfield, "Analysis of the kinetics of folding of proteins and peptides using circular dichroism.," *Nat. Protoc.*, vol. 1, no. 6, pp. 2891–9, 2006.
- [231] S. K. Nair, T. J. Thomas, N. J. Greenfield, A. Chen, H. He, and T. Thomas, "Conformational dynamics of estrogen receptors alpha and beta as revealed by intrinsic tryptophan fluorescence and circular dichroism.," *J. Mol. Endocrinol.*, vol. 35, no. 2, pp. 211–223, 2005.
- [232] J. R. Lakowicz, B. P. Maliwal, H. Cherek, and A. Balter, "Rotational freedom of tryptophan residues in proteins and peptides.," *Biochemistry*, vol. 22, no. 8, pp. 1741–1752, 1983.
- [233] N. J. Greenfield, "Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism.," *Nat. Protoc.*, vol. 1, no. 6, pp. 2733–2741, 2006.
- [234] M. A. Tangrea, P. Alexander, P. N. Bryan, E. Eisenstein, J. Toedt, and J. Orban, "Stability and global fold of the mouse prohormone convertase 1 pro-domain," *Biochemistry*, vol. 40, pp. 5488–5495, 2001.
- [235] D. M. Williamson, J. Elferich, P. Ramakrishnan, G. Thomas, and U. Shinde, "The mechanism by which a propeptide-encoded pH sensor regulates spatiotemporal activation of furin.," *J. Biol. Chem.*, vol. 288, no. 26, pp. 19154–65, Jun. 2013.
- [236] A. Basak and C. Lazure, "Synthetic peptides derived from the prosegments of proprotein convertase 1/3 and furin are potent inhibitors of both enzymes.," *Biochem. J.*, vol. 373, no. Pt 1, pp. 231–239, 2003.
- [237] A. Basak, F. Jean, N. G. Seidah, and C. Lazure, "Design and synthesis of novel inhibitors of prohormone convertases.," *Int. J. Pept. Protein Res.*, vol. 44, no. 3, pp. 253–261, 1994.
- [238] R. Day, C. Lazure, A. Basak, A. Boudreault, P. Limperis, W. Dong, and I. Lindberg, "Prodynorphin processing by proprotein convertase 2. Cleavage at single basic residues and enhanced processing in the presence of carboxypeptidase activity," *J. Biol. Chem.*, vol. 273, no. 2, pp. 829–836, 1998.
- [239] S. Tanco, X. Zhang, C. Morano, F. X. Avilés, J. Lorenzo, and L. D. Fricker, "Characterization of the substrate specificity of human carboxypeptidase A4 and implications for a role in extracellular peptide processing," *J. Biol. Chem.*, vol. 285, no. 24, pp. 18385–18396, 2010.

- [240] A. Schönichen, B. A. Webb, M. P. Jacobson, and D. L. Barber, "Considering protonation as a posttranslational modification regulating protein structure and function," *Annu. Rev. Biophys.*, vol. 42, pp. 289–314, Jan. 2013.
- [241] J. Elferich, D. M. Williamson, B. Krishnamoorthy, and U. Shinde, "Propeptides of eukaryotic proteases encode histidines to exploit organelle pH for regulation," *FASEB J.*, vol. 27, no. 8, pp. 2939–45, Aug. 2013.
- [242] M. Chrétien, N. G. Seidah, A. Basak, and M. Mbikay, "Proprotein convertases as therapeutic targets," *Expert Opin. Ther. Targets*, vol. 12, no. 10, pp. 1289–1300, 2008.
- [243] S.-N. Lee, E. Prodhomme, and I. Lindberg, "Prohormone convertase 1 (PC1) processing and sorting: effect of PC1 propeptide and proSAAS," *J. Endocrinol.*, vol. 182, no. 2, pp. 353–64, Aug. 2004.
- [244] A. G. Woods, I. Sokolowska, A. G. Ngounou Wetie, K. Wormwood, R. Aslebagh, S. Patel, and C. C. Darie, "Mass spectrometry for proteomics-based investigation," *Adv. Exp. Med. Biol.*, vol. 806, pp. 1–32, Jan. 2014.
- [245] M. Miyagi and T. Nakazawa, "Research Determination of pK_a Values of Individual Histidine Residues in Proteins Using Mass Spectrometry," *Exch. Organ. Behav. Teach. J.*, vol. 80, no. 17, pp. 6481–6487, 2008.
- [246] V. Mullangi, X. Zhou, D. W. Ball, D. J. Anderson, and M. Miyagi, "Quantitative measurement of the solvent accessibility of histidine imidazole groups in proteins," *Biochemistry*, vol. 51, no. 36, pp. 7202–8, Sep. 2012.
- [247] V. Mullangi, S. Mamillapalli, D. J. Anderson, J. G. Bann, and M. Miyagi, "Long-range stabilization of anthrax protective antigen upon binding to CMG2," *Biochemistry*, vol. 53, no. 38, pp. 6084–91, Sep. 2014.
- [248] N. Hayashi, H. Kuyama, C. Nakajima, K. Kawahara, M. Miyagi, O. Nishimura, H. Matsuo, and T. Nakazawa, "Imidazole C-2 hydrogen/deuterium exchange reaction at histidine for probing protein structure and function with matrix-assisted laser desorption ionization mass spectrometry," *Biochemistry*, vol. 53, no. 11, pp. 1818–26, Mar. 2014.
- [249] M. Strohalm, D. Kavan, P. Novák, M. Volný, and V. Havlíček, "mMass 3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data," *Anal. Chem.*, vol. 82, no. 11, pp. 4648–4651, Jun. 2010.
- [250] J. Pan, J. Han, C. H. Borchers, and L. Konermann, "Hydrogen/deuterium exchange mass spectrometry with top-down electron capture dissociation for characterizing structural transitions of a 17 kDa protein," *J. Am. Chem. Soc.*, vol. 131, pp. 12801–12808, 2009.
- [251] T. J. D. Jørgensen, H. Gårdsvoll, M. Ploug, and P. Roepstorff, "Intramolecular migration of amide hydrogens in protonated peptides upon collisional activation," *J. Am. Chem. Soc.*, vol. 127, pp. 2785–2793, 2005.

- [252] H. Yang and R. A. Zubarev, "Mass spectrometric analysis of asparagine deamidation and aspartate isomerization in polypeptides," *Electrophoresis*, vol. 31, no. 11, pp. 1764–72, Jun. 2010.
- [253] I. Kheterpal, S. Zhou, K. D. Cook, and R. Wetzel, "Abeta amyloid fibrils possess a core structure highly resistant to hydrogen exchange," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 25, pp. 13597–601, Dec. 2000.
- [254] L. Huang, X. Lu, P. C. Gough, and M. R. De Felippis, "Identification of racemization sites using deuterium labeling and tandem mass spectrometry," in *Analytical Chemistry*, 2010, vol. 82, pp. 6363–6369.
- [255] R. Bachor, B. Setner, A. Kluczyk, P. Stefanowicz, and Z. Szewczuk, "The unusual hydrogen-deuterium exchange of α -carbon protons in N-substituted glycinecontaining peptides," *J. Mass Spectrom.*, vol. 49, pp. 43–49, 2014.
- [256] J. L. Radkiewicz, H. Zipse, S. Clarke, and K. N. Houk, "Accelerated racemization of aspartic acid and asparagine residues via succinimide intermediates: An ab initio theoretical exploration of mechanism," *J. Am. Chem. Soc.*, vol. 118, pp. 9148–9155, 1996.
- [257] B. Li, R. T. Borchardt, E. M. Topp, D. VanderVelde, and R. L. Schowen, "Racemization of an asparagine residue during peptide deamidation," *J. Am. Chem. Soc.*, vol. 125, no. 38, pp. 11486–7, Sep. 2003.
- [258] N. V Di Russo, M. A. Martí, and A. E. Roitberg, "Underlying Thermodynamics of pH-Dependent Allostery," *J. Phys. Chem. B*, vol. 118, no. 45, pp. 12818–12826, Nov. 2014.
- [259] O. P. Singh and S. Sundar, "Immunotherapy and targeted therapies in treatment of visceral leishmaniasis: current status and future prospects," *Front. Immunol.*, vol. 5, p. 296, Jan. 2014.
- [260] S. M. Landfear, B. Ullman, N. S. Carter, and M. A. Sanchez, "Nucleoside and Nucleobase Transporters in Parasitic Protozoa," *Eukaryot. Cell*, vol. 3, no. 2, pp. 245–254, Apr. 2004.
- [261] R. Valdés, S. Arastu-Kapur, S. M. Landfear, and U. Shinde, "An ab initio structural model of a nucleoside permease predicts functionally important residues," *J. Biol. Chem.*, vol. 284, no. 28, pp. 19067–19076, Jul. 2009.
- [262] R. Valdés, U. Shinde, and S. M. Landfear, "Cysteine cross-linking defines the extracellular gate for the *Leishmania donovani* nucleoside transporter 1.1 (LdNT1.1)," *J. Biol. Chem.*, vol. 287, no. 53, pp. 44036–45, Dec. 2012.
- [263] N. Yan, "Structural advances for the major facilitator superfamily (MFS) transporters," *Trends in Biochemical Sciences*, vol. 38, no. 3, pp. 151–159, 2013.

- [264] S. Dang, L. Sun, Y. Huang, F. Lu, Y. Liu, H. Gong, J. Wang, and N. Yan, "Structure of a fucose transporter in an outward-open conformation," *Nature*, vol. 467, no. 7316, pp. 734–738, Oct. 2010.
- [265] J. Söding, A. Biegert, and A. N. Lupas, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, 2005.
- [266] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, vol. 9, no. 2, pp. 173–175, 2011.
- [267] J. Wu and H. R. Kaback, "Helix proximity and ligand-induced conformational changes in the lactose permease of *Escherichia coli* determined by site-directed chemical crosslinking," *J. Mol. Biol.*, vol. 270, no. 2, pp. 285–293, 1997.
- [268] N. S. Green, E. Reisler, and K. N. Houk, "Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers," *Protein Sci.*, vol. 10, no. 7, pp. 1293–1304, 2001.
- [269] Y. Zhou, L. Guan, J. A. Freites, and H. R. Kaback, "Opening and closing of the periplasmic gate in lactose permease," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 10, pp. 3774–8, Mar. 2008.
- [270] R. Valdés, J. Elferich, U. Shinde, and S. M. Landfear, "Identification of the intracellular gate for a member of the equilibrative nucleoside transporter (ENT) family," *J. Biol. Chem.*, vol. 289, no. 13, pp. 8799–809, Mar. 2014.
- [271] W. R. Taylor, "Protein structure modelling from remote sequence similarity," in *Journal of Biotechnology*, 1994, vol. 35, no. 2–3, pp. 281–291.
- [272] L. H. Pearl and W. R. Taylor, "A structural model for the retroviral proteases," *Nature*, vol. 329, no. 6137, pp. 351–4, Jan. .
- [273] M. Miller, M. Jaskólski, J. K. Rao, J. Leis, and A. Wlodawer, "Crystal structure of a retroviral protease proves relationship to aspartic protease family," *Nature*, vol. 337, no. 6207, pp. 576–579, 1989.
- [274] B. P. Pedersen, H. Kumar, A. B. Waight, A. J. Risenmay, Z. Roe-Zurz, B. H. Chau, A. Schlessinger, M. Bonomi, W. Harries, A. Sali, A. K. Johri, and R. M. Stroud, "Crystal structure of a eukaryotic phosphate transporter," *Nature*, vol. 496, no. 7446, pp. 533–6, 2013.
- [275] S. Newstead, D. Drew, A. D. Cameron, V. L. G. Postis, X. Xia, P. W. Fowler, J. C. Ingram, E. P. Carpenter, M. S. P. Sansom, M. J. McPherson, S. A. Baldwin, and S. Iwata, "Crystal structure of a prokaryotic homologue of the mammalian oligopeptide-proton symporters, PepT1 and PepT2," *EMBO J.*, vol. 30, no. 2, pp. 417–426, 2011.

- [276] T. Hirai, J. A. W. Heymann, D. Shi, R. Sarker, P. C. Maloney, and S. Subramaniam, "Three-dimensional structure of a bacterial oxalate transporter.," *Nat. Struct. Biol.*, vol. 9, no. 8, pp. 597–600, 2002.
- [277] Y. Yin, X. He, P. Szewczyk, T. Nguyen, and G. Chang, "Structure of the multidrug transporter EmrD from *Escherichia coli*," *Science*, vol. 312, no. 5774, pp. 741–744, 2006.
- [278] T. Devine and M.-S. Dai, "Targeting the ubiquitin-mediated proteasome degradation of p53 for cancer therapy.," *Curr. Pharm. Des.*, vol. 19, no. 18, pp. 3248–62, 2013.
- [279] M. H. Kubbutat, S. N. Jones, and K. H. Vousden, "Regulation of p53 stability by Mdm2.," *Nature*, vol. 387, no. 6630, pp. 299–303, 1997.
- [280] X.-X. Sun, K. B. Challagundla, and M.-S. Dai, "Positive regulation of p53 stability and activity by the deubiquitinating enzyme Otubain 1.," *EMBO J.*, vol. 31, no. 3, pp. 576–92, Feb. 2012.
- [281] R. Wiener, X. Zhang, T. Wang, and C. Wolberger, "The mechanism of OTUB1-mediated inhibition of ubiquitination.," *Nature*, vol. 483, no. 7391, pp. 618–622, 2012.
- [282] Y. Li, X.-X. Sun, J. Elferich, U. Shinde, L. L. David, and M.-S. Dai, "Monoubiquitination is critical for ovarian tumor domain-containing ubiquitin aldehyde binding protein 1 (Otub1) to suppress UbcH5 enzyme and stabilize p53 protein.," *J. Biol. Chem.*, vol. 289, no. 8, pp. 5097–108, Feb. 2014.
- [283] R. Baker, S. M. Lewis, A. T. Sasaki, E. M. Wilkerson, J. W. Locasale, L. C. Cantley, B. Kuhlman, H. G. Dohlman, and S. L. Campbell, "Site-specific monoubiquitination activates Ras by impeding GTPase-activating protein function.," *Nat. Struct. Mol. Biol.*, vol. 20, no. 1, pp. 46–52, 2013.
- [284] S. Nakada, I. Tai, S. Panier, A. Al-Hakim, S.-I. Iemura, Y.-C. Juang, L. O'Donnell, A. Kumakubo, M. Munro, F. Sicheri, A.-C. Gingras, T. Natsume, T. Suda, and D. Durocher, "Non-canonical inhibition of DNA damage-dependent ubiquitination by OTUB1.," *Nature*, vol. 466, no. 7309, pp. 941–946, 2010.
- [285] Y. C. Juang, M. C. Landry, M. Sanches, V. Vittal, C. C. Y. Leung, D. F. Ceccarelli, A. R. F. Mateo, J. N. Pruneda, D. Y. L. Mao, R. K. Szilard, S. Orlicky, M. Munro, P. S. Brzovic, R. E. Kleivit, F. Sicheri, and D. Durocher, "OTUB1 Co-opts Lys48-Linked Ubiquitin Recognition to Suppress E2 Enzyme Function," *Mol. Cell*, vol. 45, no. 3, pp. 384–397, 2012.
- [286] P. S. Brzovic, A. Lissounov, D. E. Christensen, D. W. Hoyt, and R. E. Kleivit, "A UbcH5/ubiquitin noncovalent complex is required for processive BRCA1-directed ubiquitination," *Mol. Cell*, vol. 21, no. 6, pp. 873–880, 2006.
- [287] E. Sakata, T. Satoh, S. Yamamoto, Y. Yamaguchi, M. Yagi-Utsumi, E. Kurimoto, K. Tanaka, S. Wakatsuki, and K. Kato, "Crystal Structure of UbcH5b~Ubiquitin Intermediate: Insight

- into the Formation of the Self-Assembled E2~Ub Conjugates," *Structure*, vol. 18, no. 1, pp. 138–147, 2010.
- [288] A. Saha, S. Lewis, G. Kleiger, B. Kuhlman, and R. J. Deshaies, "Essential Role for Ubiquitin-Ubiquitin-Conjugating Enzyme Interaction in Ubiquitin Discharge from Cdc34 to Substrate," *Mol. Cell*, vol. 42, no. 1, pp. 75–83, 2011.
- [289] W. Song, D. Raden, E. C. Mandon, and R. Gilmore, "Role of Sec61alpha in the regulated transfer of the ribosome-nascent chain complex from the signal recognition particle to the translocation channel.," *Cell*, vol. 100, no. 3, pp. 333–343, 2000.
- [290] V. Goder, T. Junne, and M. Spiess, "Sec61p contributes to signal sequence orientation according to the positive-inside rule.," *Mol. Biol. Cell*, vol. 15, no. 3, pp. 1470–1478, 2004.
- [291] E. Perara, R. E. Rothman, and V. R. Lingappa, "Uncoupling translocation from translation: implications for transport of proteins across membranes.," *Science*, vol. 232, no. 4748, pp. 348–352, 1986.
- [292] P. K. Devaraneni, B. Conti, Y. Matsumura, Z. Yang, A. E. Johnson, and W. R. Skach, "Stepwise insertion and inversion of a type II signal anchor sequence in the ribosome-Sec61 translocon complex," *Cell*, vol. 146, no. 1, pp. 134–147, 2011.
- [293] B. J. Conti, J. Elferich, Z. Yang, U. Shinde, and W. R. Skach, "Cotranslational folding inhibits translocation from within the ribosome-Sec61 translocon complex.," *Nat. Struct. Mol. Biol.*, vol. 21, no. 3, pp. 228–35, Mar. 2014.
- [294] D. T. Rutkowski, V. R. Lingappa, and R. S. Hegde, "Substrate-specific regulation of the ribosome- translocon junction by N-terminal signal sequences.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 14, pp. 7823–7828, 2001.
- [295] M. Hojjat-Farsangi, "Small-molecule inhibitors of the receptor tyrosine kinases: promising tools for targeted cancer therapies.," *Int. J. Mol. Sci.*, vol. 15, no. 8, pp. 13768–801, Jan. 2014.
- [296] J. N. Spaans and G. D. Goss, "Trials to Overcome Drug Resistance to EGFR and ALK Targeted Therapies - Past, Present, and Future.," *Front. Oncol.*, vol. 4, p. 233, Jan. 2014.
- [297] K. Takeuchi, M. Soda, Y. Togashi, R. Suzuki, S. Sakata, S. Hatano, R. Asaka, W. Hamanaka, H. Ninomiya, H. Uehara, Y. Lim Choi, Y. Satoh, S. Okumura, K. Nakagawa, H. Mano, and Y. Ishikawa, "RET, ROS1 and ALK fusions in lung cancer," *Nature Medicine*, vol. 18, no. 3, pp. 378–381, 2012.
- [298] T. L. Gu, X. Deng, F. Huang, M. Tucker, K. Crosby, V. Rimkunas, Y. Wang, G. Deng, L. Zhu, Z. Tan, Y. Hu, C. Wu, J. Nardone, J. MacNeill, J. Ren, C. Reeves, G. Innocenti, B. Norris, J. Yuan, J. Yu, H. Haack, B. Shen, C. Peng, H. Li, X. Zhou, X. Liu, J. Rush, and M. J. Comb, "Survey of tyrosine kinase signaling reveals ROS kinase fusions in human cholangiocarcinoma," *PLoS One*, vol. 6, no. 1, 2011.

- [299] A. Charest, K. Lane, K. McMahon, J. Park, E. Preisinger, H. Conroy, and D. Housman, "Fusion of FIG to the receptor tyrosine kinase ROS in a glioblastoma with an interstitial del(6)(q21q21)," *Genes Chromosom. Cancer*, vol. 37, no. 1, pp. 58–71, 2003.
- [300] A. T. Shaw, D. R. Camidge, J. A. Engelman, B. J. Solomon, E. L. Kwak, J. W. Clark, G. Shapiro, Y. Bang, W. Tan, L. Tye, K. D. Wilner, P. Stephenson, and M. Varela-, "Clinical activity of crizotinib in advanced non-small cell lung cancer (NSCLC) harboring ROS1 gene rearrangement," *J Clin Oncol*, vol. 30, pp. 9–11, 2012.
- [301] M. A. Davare, A. Saborowski, C. A. Eide, C. Tognon, R. L. Smith, J. Elferich, A. Agarwal, J. W. Tyner, U. P. Shinde, S. W. Lowe, and B. J. Druker, "Foretinib is a potent inhibitor of oncogenic ROS1 fusion proteins.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 48, pp. 19519–24, Nov. 2013.
- [302] S. Zhang, F. Wang, J. Keats, X. Zhu, Y. Ning, S. D. Wardwell, L. Moran, Q. K. Mohemmad, R. Anjum, Y. Wang, N. I. Narasimhan, D. Dalgarno, W. C. Shakespeare, J. J. Miret, T. Clackson, and V. M. Rivera, "Crizotinib-resistant mutants of EML4-ALK identified through an accelerated mutagenesis screen," *Chem. Biol. Drug Des.*, vol. 78, no. 6, pp. 999–1005, 2011.
- [303] R. C. Doebele, A. B. Pilling, D. L. Aisner, T. G. Kutateladze, A. T. Le, A. J. Weickhardt, K. L. Kondo, D. J. Linderman, L. E. Heasley, W. A. Franklin, M. Varela-Garcia, and D. R. Camidge, "Mechanisms of resistance to crizotinib in patients with ALK gene rearranged non-small cell lung cancer," *Clin. Cancer Res.*, vol. 18, no. 5, pp. 1472–1482, 2012.
- [304] C. Wilson, R. V. Agafonov, M. Hoemberger, S. Kutter, A. Zorba, J. Halpin, V. Buosi, R. Otten, D. Waterman, D. L. Theobald, and D. Kern, "Using ancient protein kinases to unravel a modern cancer drug's mechanism.," *Science*, vol. 347, no. 6224, pp. 882–6, Mar. 2015.
- [305] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster, "Mass-spectrometry-based draft of the human proteome.," *Nature*, vol. 509, no. 7502, pp. 582–7, May 2014.
- [306] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, "STRING v10: protein-protein interaction networks, integrated over the tree of life.," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D447–52, Jan. 2015.
- [307] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander, "Pathway Commons, a web resource for biological pathway data.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D685–90, Jan. 2011.
- [308] J. Chen, "Analysis of Moore's Law on Intel Processors," in *Proceedings of the 2013 International Conference on Electrical and Information Technologies for Rail Transportation*

(EITRT2013)-Volume II SE - 42, vol. 288, L. Jia, Z. Liu, Y. Qin, M. Zhao, and L. Diao, Eds. Springer Berlin Heidelberg, 2014, pp. 391–400.

- [309] K. Amunts, A. Lindner, and K. Zilles, “The human brain project: neuroscience perspectives and German contributions,” *e-Neuroforum*, vol. 5, no. 2, pp. 43–50, 2014.
- [310] J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V Gutschow, J. M. Jacobs, B. Bolival, N. Assad-Garcia, J. I. Glass, and M. W. Covert, “A whole-cell computational model predicts phenotype from genotype,” *Cell*, vol. 150, no. 2, pp. 389–401, Jul. 2012.
- [311] K. Vanommeslaeghe and A. D. MacKerell, “CHARMM additive and polarizable force fields for biophysics and computer-aided drug design,” *Biochim. Biophys. Acta*, vol. 1850, no. 5, pp. 861–871, May 2015.
- [312] M. Helmstaedter, “The Mutual Inspirations of Machine Learning and Neuroscience,” *Neuron*, vol. 86, no. 1, pp. 25–28, Apr. 2015.
- [313] S. Liu, L. Liu, U. Uzuner, X. Zhou, M. Gu, W. Shi, Y. Zhang, S. Y. Dai, and J. S. Yuan, “HDX-analyzer: a novel package for statistical analysis of protein structure dynamics,” *BMC Bioinformatics*, vol. 12 Suppl 1, p. S43, 2011.
- [314] B. D. Pascal, M. J. Chalmers, S. A. Busby, and P. R. Griffin, “HD Desktop: An Integrated Platform for the Analysis and Visualization of H/D Exchange Data,” *J. Am. Soc. Mass Spectrom.*, vol. 20, no. 4, pp. 601–610, 2009.
- [315] B. D. Pascal, S. Willis, J. L. Lauer, R. R. Landgraf, G. M. West, D. Marciano, S. Novick, D. Goswami, M. J. Chalmers, and P. R. Griffin, “HDXWorkbench: Software for the analysis of H/D exchange MS data,” *J. Am. Soc. Mass Spectrom.*, vol. 23, no. 9, pp. 1512–1521, 2012.
- [316] J. K. Eng, B. Fischer, J. Grossmann, and M. J. MacCoss, “A fast SEQUEST cross correlation algorithm,” *J. Proteome Res.*, vol. 7, no. 10, pp. 4598–4602, 2008.
- [317] M. Brosch, L. Yu, T. Hubbard, and J. Choudhary, “Accurate and sensitive peptide identification with mascot percolator,” *J. Proteome Res.*, vol. 8, no. 6, pp. 3176–3181, 2009.
- [318] D. Crockford, “The application/json Media Type for JavaScript Object Notation,” *RFC-4627*, 2006. [Online]. Available: <http://www.ietf.org/rfc/rfc4627.txt?number=4627>.
- [319] M. Bostock, V. Ogievetsky, and J. Heer, “D3: Data-Driven Documents,” *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, 2011.
- [320] E. Deutsch, “mzML: A single, unifying data format for mass spectrometer output,” *Proteomics*, vol. 8, no. 14, pp. 2776–2777, 2008.
- [321] A. A. Klammer, C. Y. Park, and W. S. Noble, “Statistical calibration of the SEQUEST Xcorr function,” *J. Proteome Res.*, vol. 8, no. 4, pp. 2106–2113, 2009.

- [322] X. Ye, B. Luke, T. Andresson, and J. Blonder, "18O stable isotope labeling in MS-based proteomics," *Brief. Funct. Genomic. Proteomic.*, vol. 8, no. 2, pp. 136–44, Mar. 2009.