UNIVERSITY OF CALIFORNIA, SAN DIEGO

Deciphering Proteolytic Signaling

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Molecular Pathology

by

John C. Timmer

Committee in Charge:

Professor Guy Salvesen, Chair
Professor Victor Nizet, Co-Chair
Professor Lars Eckmann
Professor Elizabeth Komives
Professor Robert Liddington

2009

The Dissertation of John C. Timmer is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2009

DEDICATION

To my family

TABLE OF CONTENTS

LIST OF FIGURES & TABLES

Chapter II

Chapter III

ACKNOWLEDGEMENTS

Before coming to the Salvesen lab I had not even considered doing protease or apoptosis research, and now I am so entrenched in the questions and issues of the field that I have difficulty leaving my ideas untested and pet projects unfinished. I sincerely appreciate the opportunity Guy gave me when he took me into the lab, and all the time he spent teaching me and advising me on my long initiation into the protease and apoptosis fields. His passion for science and drive for sound experimentation are exemplary traits that I strive to emulate in my own life.

I would like to thank all of the people in the Salvesen lab that I had the chance to work with and be friends with. From the first day in the lab I remember thinking about how helpful, fun, and just plain nice Scott Snipas was. Since then I have been fortunate enough to become good friends with him, and have enjoyed many a beer with Scott at conferences late at night, after work at the Hilton, or even just on the weekend for a UFC event. Jean-Bernard Denault taught me how to Applescript, and entertained me with ‘is French-Canadian accent. Cristina Pop showed me the ropes of the Salvesen lab during my rotation, and is the epitome of hard work and dedication. I appreciate her caring, honest, and direct approach to life and friendship. Marcin Drag taught me about HPLC and entertained my novice chemistry questions. I won’t forget how he caught the grunion scouts and scared the masses all away. Brendan Eckelman was my classmate and fellow

inflammatory caspase enthusiast. His balanced and personal approach to life, made the lab entertaining, and helped me to appreciate the great camaraderie of the lab. I also thank all of the other lab members who I befriended during my Ph.D. I will not willingly forget all the jokes and fun we had, all the beers and the hung-over morning sessions at meetings, all the lunches together at the Burnham, the various lab parties, and all the interesting conversations.

To my parents, Anne and Bruce, I cannot thank you enough for all of your love, encouragement, and support. You always gave me the power to make my own choices, although I do not know if I always deserved it. Yet your confidence in me, and your support of my decisions enabled me to pursue my dreams without restrictions, even though my choices must have compromised some of your opportunities. Since having Caden, and a baby girl on the way, I realize all of the effort and compromise that being a parent demands. I hope I can teach my kids to pursue their own interests and to live a balanced life like you both did so well with Nick and I. Your involvement in Caden's life has been a great help for Anjuli and I, and a wonderful treat for Caden. Thanks you so very much.

I also have another family that has supported me during my Ph.D. Rakesh, Mom, and Subhash, Dad, thank you for taking me into your family and treating me as a son. You have expanded my perspectives and horizons, and challenged my most fundamental beliefs. It always feels like home when we make our semiannual trips back to Durham, and even though I still have not been

taken to Mount Major, I have been taken to India twice, and feel deeply privileged to have connected with your friends and family. Anuj, my little brother, I am so impressed by you in all that you do. Your fun-seeking and enthusiastic can-do approach to life continually inspires me to follow my dreams and make my life the way I want it to be. Thank you Minochas for your love and advice throughout my graduate studies.

My son, Caden, also deserves a huge thank you. Yes, he continues to make life more challenging than it was when I was the center of the universe, but he brings me a joy and purpose that permeates everything I do. I have grown more in the last 2 ½ years than ever before, and I have found a connection with the rest of humanity throughout history in the challenges and joys of raising a child. I am so thankful and proud of Caden everyday, because if it weren't for him and Anjuli, I would probably be working all of the time. Now my life is balanced, enjoying both family life and my professional pursuits.

I would like to thank my wife, Anjuli, without whom I probably never would have gotten into the Molecular Pathology Ph.D. program. Even though we were already planning our future together, it was in grad school that Anjuli and I got married (all three times). We also had our son Caden, and will soon have his baby sister. Not all of these changes were easy and fun. In fact, relegating my own wants to the needs of my family is probably the most difficult adjustment I have ever had to make. Luckily Anjuli was there through it all tempering my

frustration, sharing in the work, encouraging my hobbies, and supporting my aspirations. She showed me how to be persistent, to take responsibility, and to take control of my future. She is my voice of reason, my moral compass, and my guide when I lose sight of the big picture. Thank you for sharing your days, your cares, and your troubles with me. I love you and look forward to our future unfolding together.

*****

Chapter II was modified from the publication: Timmer JC, Enoksson M, Wildfang E, Zhu W, Igarashi Y, Denault JB, Ma Y, Dummitt B, Chang YH, Mast AE, Eroshkin A, Smith JW, Tao WA, Salvesen GS. Profiling constitutive proteolysis *in vivo*. *Journal of Biochemistry.* October 2007: 407(1):41-8, with permission from all co-authors. Chapter III was modified from the manuscript accdpted for publication: Timmer JC, Zhu W, Pop C, Snipas SJ, Eroshkin AM, Riedl SJ, Salvesen GS. Structural and kinetic determinants of protease substrates. *Nature Structure and Molecular Biology.* Accepted, with permission from all co-authors.

CURRICULUM VITAE

EDUCATION

2004 to 2009 Molecular Pathology PhD Program
University of California, San Diego, and Burnham Institute for Medical research

1996 to 2000 California Polytechnic State University, San Luis Obispo
Major: Microbiology

PUBLICATIONS

Timmer JC, Zhu W, Pop C, Regan T, Snipas SJ, Eroshkin AM, Riedl SJ, Salvesen GS. Structural and kinetic determinants of protease substrates. *Nat Struct Mol Biol.* Accepted..

Timmer AM**, Timmer JC**, Pence MA, Hsu LC, Ghochani M, Frey TG, Karin M, Salvesen GS, and Nizet V. Group A *Streptococcus* immune evasion by cytolysin-mediated accelerated macrophage apoptosis**.** *J Biol Chem*. January 2009: 284(2):862-71

Timmer JC, Enoksson M, Wildfang E, Zhu W, Igarashi Y, Denault JB, Ma Y, Dummitt B, Chang YH, Mast AE, Eroshkin A, Smith JW, Tao WA, Salvesen GS. Profiling constitutive proteolytic events *in vivo*. *Biochem J.* October 2007: 407(1):41-8.

Enoksson M, Li J, Ivancic MM, **Timmer JC**, Wildfang E, Eroshkin A, Salvesen GS, Tao WA. Identification of proteolytic cleavage sites by quantitative proteomics. *J Proteome Res.* July 2007: 6(7):2850-8.

Timmer JC, Salvesen GS. Caspase substrates. *Cell Death Differ.* January 2007: 14(1):66-72.

Pop C, **Timmer J,** Sperandio S, Salvesen GS. The apoptosome activates caspase-9 by dimerization. *Mol Cell.* April 2006: 22(2):269-75.
.

RESEARCH EXPERIENCE:

- **Graduate student in the Molecular Pathology PhD Program, University of California, San Diego:** Research in the lab of Dr. Guy Salvesen investigating protease substrate identification and structural

features of cleavage-sites. Developed a method to identify protease substrates and cleavage-sites using proteomics, and applied this and biochemical techniques to determine what sequences and structures natural substrates utilize to drive proteolysis *(March 2005 to present).*

- **Rotation student in the Molecular Pathology PhD Program, University of California, San Diego:** Lab of Dr. Richard Gallo exploring the promoter elements regulating the cathelicidin antimicrobial peptide gene (*January to March 2005).*

- **Associate Scientist II at Applied Molecular Evolution, San Diego, CA (now a wholly owned subsidiary of Eli Lilly Co.):** In the Research group under Dr. Craig Dickinson on several protein engineering projects (proprietary). Responsible for molecular biology, tissue culture, cell-based assays, recombinant protein express, and purification (*January 2003 to Auguest 2004).*

- **Research Associate I at Applied Molecular Evolution, San Diego, CA (now a wholly owned subsidiary of Eli Lilly Co.):** In the DNA sequencing core under Franz Triana. Responsible for DNA sequence analysis and oligonucleotide synthesis. Implemented logistical and policy changes to reduce sequencing turn around time to next day results (*October 2000 to December 2002).*

- **Intern at Genentech, South San Francisco, CA:** In the Quality Control DNA group under Dr. Judy Helder and Dr. Joyce Eldering. Tested non-radioactive detection methods for visualizing DNA during agarose electrophoresis, and for Southern Blotting. Performed DNA sequencing with premixed wild type and mutant plasmids to determine the sensitivity of the system to detect point mutations (*July to September 1996 & 1997).*

FELLOWSHIPS AND HONORS

- National Cancer Institute Research Training Fellowship (*September 2007 to present*)
- International Proteolysis Society Conference Travel Award Recipient *(October 2007)*
- International Proteolysis Society Conference Travel Award Recipient *(October 2005)*

PRESENTATIONS:

- 6th Annual Pacific Coast Protease Conference, Warner Springs, CA April 2009. Oral presentation: *Deciphering proteolyisis in signaling: What it takes to be a protease substrate*.

- The Centers for Networks and Pathways, National Institutes for Health, MD March 2008. Poster Presentation: *Finding natural protease substrates by N-terminal proteomics and disordered cleavage-site prediction*.

- 5th Annual Pacific Coast Protease Conference, Borrego Springs, CA April 2008. Oral presentation: *The structural constraints of proteolysis*.

- International Proteolysis Society Conference Patras, Greece October 2007. Poster Presentation: *Finding natural protease substrates by N-terminal proteomics and disordered cleavage-site prediction*.

- 4th Annual Pacific Coast Protease Conference, Borrego Springs, CA March 2007. Oral presentation: *Profiling proteolysis in vivo by mass spectrometry*.

- American Society of Mass Spectrometry, Seattle, WA May 2006. Poster Presentation: *The"N-terminome": Fingerprinting the in vivo activity of processing peptidases in* E. coli.

- 3th Annual Pacific Coast Protease Conference, Desert Hot Springs, CA April 2006. Oral presentation: *The* E. coli *"N-terminome": Fingerprinting the activity of processing peptidases in vivo*.

- 23rd Winter School Conference on Proteases and Their Inhibitors, Tiers Italy, March 2006. Oral presentation: The Recognition of Natural Substrates by Proteases: *Caspases on* E. coli *and human proteins as a model system*.

- International Proteolysis Society Conference Quebec City, Canada October 2005. Poster Presentation: *Predicting caspase substrates: How flexible do you need to be?*

- 2nd Annual Pacific Coast Protease Conference, Half Moon Bay May 2005. Oral presentation: *Why caspases kill E. coli*?

UNIVERSITY SERVICE:

- Teacher's assistant for beginning and intermediate pottery on the wheel *(2004 to current).*

ABSTRACT OF THE DISSERTATION

Deciphering Proteolytic Signaling

by

John C. Timmer

Doctor of Philosophy in Molecular Pathology
University of California, San Diego, 2009

Professor Guy Salvesen, Chair
Professor Victor Nizet, Co-Chair

Proteases are important enzymes involved in biological and pathological systems. We developed a proteomics-based method to identify protease substrates and their cleavage-sites directly. The principle of identification relies on labeling the new N-terminus generated by proteolysis with an affinity tag, that serves to enrich cleavage-site peptides for analysis by liquid chromatography coupled tandem mass spectrometry (LC-MS/MS). Spectra collected are searched against a relevant proteome database to identify the protein identities

as well as the location of the labeled peptides. Each peptide is annotated based on the location and sequence of the peptide within the full-length protein sequence. In this way we assessed the endogenous proteolytic activity in *E. coli*, *S. cerevisiae*, several mouse tissues, a human cell line, and human serum. A substantial proportion of the N-terminal peptides identified from all samples were full-length protein N-termini, as well as N-termini corresponding to co-translational proteolytic processing by well-characterized proteases. Yet, the majority of all N-termini corresponded to internal cleavage-sites, suggesting that proteolysis *in vivo* is much more prevalent than had previously been appreciated.

A longstanding dogma in the protease field states that proteases only cleave substrates in regions lacking secondary structure. We reasoned that N-terminomics could challenge this dogma by identifying human caspase-3 and Staphylococcal glutamyl endopeptidase cleavage-sites from the structurally diverse and well-characterized folded protein library contained within *E. coli* lysate without bias in regard to the structures cleaved. Our analysis revealed that both proteases cleaved *E. coli* proteins in unstructured loops as well as α-helices, but almost never in β-strands. Many *E. coli* substrates of caspase-3 were recombinantly expressed and purified, and kinetically analyzed by *in vitro* assay, revealing that *E. coli* substrates were kinetically inferior to natural caspase-3 substrates. This kinetic deficiency was successfully overcome by engineering a poor *E. coli* substrate cleavage-site with the optimal amino acid sequence within an extended flexible loop. These results show that although helices can be cleaved by proteases, the natural substrates of caspase-3 have co-evolved with

this protease resulting in efficient cleavage of near-optimal amino acid sequences positioned within flexible loop structures.

Chapter I

Introduction:
Protease Activity And Function

INTRODUCTION

Proteases in biology and pathology. Initially characterized as protein degrading enzymes, proteases are themselves proteins that catalyze the hydrolysis of peptide bonds linking adjacent amino acids together to form peptides and proteins. Proteases are encoded by genes found in nearly every organism, and play diverse roles in physiological processes, such as food digestion, blood coagulation, bone homeostasis, wound healing, blood pressure regulation, immune function, fertilization, cholesterol synthesis, embryonic development, and blood glucose maintenance (Jackson and Nemerson 1980) (Toriseva and Kahari 2009) (Brown and Goldstein 1999) (Reilly, Tewksbury et al. 1982) (Studdy, Lapworth et al. 1983) (Black, Kronheim et al. 1988) (Thornberry, Bull et al. 1992) (Lilja 1990) (Holst and Deacon 1998) (Vaux and Strasser 1996). In addition, proteases are frequently involved in various pathologies by contributing to viral, fungal, and bacterial infection and immune evasion (reference); and have been shown to contribute to cancer progression and metastasis (Anderson, Schiffer et al. 2009) (Gross, Poeck et al. 2009) (Rasmussen and Bjorck 2002) (Zinkernagel, Timmer et al. 2008) (Gocheva and Joyce 2007). Studies have validated the role of proteases in these settings by showing that the biological or pathological process in question is compromised when the protease is inactivated by chemical inhibition or genetic deletion. These numerous and varied contributions of proteases attest to their versatility and general utility in nature.

Proteolytic signaling. From a cellular perspective, proteases function after chromosomally encoded genes are transcribed from DNA into mRNA, and then translated by ribosomes into proteins. Some proteases then function to degrade protein substrates into individual amino acids, which are recycled for new protein synthesis or cellular metabolism. Other proteases are players in signal transduction pathways transmitting cellular signals through limited cleavage of protein substrates. Substrate cleavage serves to functionally alter substrates in some manner that perpetuates the information transfer culminating in a cellular or physiological change. The functional outcome of substrate cleavage is unique to each substrate, and can be different for the same substrate cleaved at distinct sites, providing a differential response to converging signaling pathways with divergent outcomes. Yet, substrate cleavage in signaling generally causes a change in substrate activity (activation or inhibition), a change in substrate regulation, or a change in substrate localization (Liu, Zou et al. 1999) (Timmer and Salvesen 2007) (Scott, Fuchs et al. 2008).

Protease specificity. Proteases can be sub-divided by their activity on protein substrates: aminopeptidases remove amino acids from the amino-terminus (N-terminus), while carboxypeptidases remove amino acids form the carboxy-terminus (C-terminus), and endopeptidase cleave proteins at intermediate sites in proteins, and are not restricted to the N- or C-termini. Nearly all proteases maintain some degree of specificity in terms of the amino acid composition that they prefer to cleave. However, unlike the exquisite specificity of

many common restriction endonucleases that cut DNA at particular nucleotide sequences, most proteases will tolerate the presence of similar amino acids within the cleavage-site. This flexibility in specificity usually translates into differential rates of substrate cleavage depending on the degree to which the substrate satisfies the protease's amino acid preferences.

Substrate features. Other factors besides protease specificity can also influence the kinetics of substrate cleavage. Proteases must physically interact with substrates to catalyze cleavage; therefore, both components must be present at the same time and in the same physical or sub-cellular location. Although many proteases are thought to limit their interactions with substrates to the active site binding the cleavage-site, a large number of proteases utilize distinct secondary interactions to bind substrates (Mikolajczyk, Drag et al. 2007) (Stubbs and Bode 1995) (Overall 2002). These exosites provide additional interactions that can modulate protease specificity, and enhance substrate catalysis for exosite containing substrates. The structural composition of substrates also has bearing on which sites are susceptible to proteolysis. It has long been thought that proteases cleave substrates in flexible and unstructured regions, while rigid structured regions are protected from cleavage. All of these factors contribute to the ability of proteases to cleave substrates, and to the catalytic rate at which substrates are cleaved.

Essential signaling substrates. Although signaling proteases are often surrounded by and have physical access to a plethora of potential substrates, they only cleave a discrete minority of proteins, and usually at a single site per substrate. However, not all substrates cleaved by proteases are critical to drive forward the signaling pathway in order to elicit a phenotypic response. These non-signaling and non-essential substrates are termed bystanders, and pose a non-trivial problem for scientists as they dramatically complicate signal transduction pathways. Bystander substrates need to be partitioned from the essential signaling substrates for scientists investigating proteolytic signaling to realize the ultimate goal: to define the minimal substrate repertoire necessary to achieve the biological outcome of interest.

REFERENCES

Anderson, J., C. Schiffer, et al. (2009). "Viral protease inhibitors." Handb Exp Pharmacol(189): 85-110.

Black, R. A., S. R. Kronheim, et al. (1988). "Generation of biologically active interleukin-1 beta by proteolytic cleavage of the inactive precursor." J Biol Chem **263**(19): 9437-42.

Brown, M. S. and J. L. Goldstein (1999). "A proteolytic pathway that controls the cholesterol content of membranes, cells, and blood." Proc Natl Acad Sci U S A **96**(20): 11041-8.

Gocheva, V. and J. A. Joyce (2007). "Cysteine cathepsins and the cutting edge of cancer invasion." Cell Cycle **6**(1): 60-4.

Gross, O., H. Poeck, et al. (2009). "Syk kinase signalling couples to the Nlrp3 inflammasome for anti-fungal host defence." Nature.

Holst, J. J. and C. F. Deacon (1998). "Inhibition of the activity of dipeptidyl-peptidase IV as a treatment for type 2 diabetes." Diabetes **47**(11): 1663-70.

Jackson, C. M. and Y. Nemerson (1980). "Blood coagulation." Ann. Rev. Biochem. **49**: 765-811.

Lilja, H. (1990). "Cell biology of semenogelin." Andrologia **22 Suppl 1**: 132-41.

Liu, X., H. Zou, et al. (1999). "Activation of the apoptotic endonuclease DFF40 (caspase-activated DNase or nuclease). Oligomerization and direct interaction with histone H1." J Biol Chem **274**(20): 13836-40.

Mikolajczyk, J., M. Drag, et al. (2007). "Small Ubiquitin-related Modifier (SUMO)-specific Proteases: PROFILING THE SPECIFICITIES AND ACTIVITIES OF HUMAN SENPs." J Biol Chem **282**(36): 26217-24.

Overall, C. M. (2002). "Molecular determinants of metalloproteinase substrate specificity: matrix metalloproteinase substrate binding domains, modules, and exosites." Mol Biotechnol **22**(1): 51-86.

Rasmussen, M. and L. Bjorck (2002). "Proteolysis and its regulation at the surface of Streptococcus pyogenes." Mol Microbiol **43**(3): 537-44.

Reilly, C. F., D. A. Tewksbury, et al. (1982). "Rapid conversion of angiotensin I to angiotensin II by neutrophil and mast cell proteinases." J Biol Chem **257**: 8619-8622.

Scott, F. L., G. J. Fuchs, et al. (2008). "Caspase-8 cleaves histone deacetylase 7 and abolishes its transcription repressor function." J Biol Chem **283**(28): 19499-510.

Stubbs, M. T. and W. Bode (1995). "The clot thickens: clues provided by thrombin structure." Trends Biochem Sci **20**(1): 23-8.

Studdy, P. R., R. Lapworth, et al. (1983). "Angiotensin-converting enzyme and its clinical significance--a review." J Clin Pathol **36**(8): 938-47.

Thornberry, N. A., H. G. Bull, et al. (1992). "A novel heterodimeric cysteine protease is required for interleukin-1 beta processing in monocytes." Nature **356**(6372): 768-74.

Timmer, J. C. and G. S. Salvesen (2007). "Caspase substrates." Cell Death Differ **14**(1): 66-72.

Toriseva, M. and V. M. Kahari (2009). "Proteinases in cutaneous wound healing." Cell Mol Life Sci **66**(2): 203-24.

Vaux, D. L. and A. Strasser (1996). "The molecular biology of apoptosis." Proc Natl Acad Sci U S A **93**(6): 2239-44.

Zinkernagel, A. S., A. M. Timmer, et al. (2008). "The IL-8 protease SpyCEP/ScpC of group A Streptococcus promotes resistance to neutrophil killing." Cell Host Microbe **4**(2): 170-8.

Chapter II

Profiling constitutive proteolytic events *in vivo*

ABSTRACT

Nearly every organism known to science encodes proteases that perform essential functions to maintain cellular and organismal homeostasis. The activity of these proteases on cellular substrates forms the steady-state level of proteolysis in a cell or tissue. In this chapter, we describe a method to define these constitutive proteolytic events in diverse proteomes, from *Escherichia coli* to humans. The method takes advantage of specific N-terminal biotinylation of protein samples, followed by affinity enrichment and conventional LC (liquid chromatography)–MS/MS (tandem mass spectrometry) analysis. The method is simple, uses conventional and easily obtainable reagents, and is applicable to most proteomics facilities. As proof of principle, we demonstrate profiles of proteolytic events that reveal exquisite *in vivo* specificity of methionine aminopeptidase in *E. coli* and unexpected processing of mitochondrial transit peptides in yeast, mouse and human samples. Taken together, our results demonstrate how to rapidly distinguish real proteolysis that occurs *in vivo* from the predictions based on *in vitro* experiments.

INTRODUCTION

A substantial fraction of nearly all proteomes encode genes for proteases, accounting for approximately 1-5% of genomes, depending on the species (Rawlings, Morton et al. 2006) (Lopez-Otin and Overall 2002). Proteases were originally identified over a century ago as protein-degrading enzymes in digestive juices and tissue homogenates, leading to the concept that they are protein destructors. Although many proteases are indeed required for digestion and protein degradation, the protease field now appreciates the diverse signaling functions of these enzymes. These signaling proteases carry out limited proteolysis of specific substrates to play pivotal roles in most biological processes (Salvesen and Dixit 1997). The function of proteolysis *in vivo* is to catalyze the irreversible cleavage of a set of protein substrates that alters their function, and leads to the changes required for the downstream biological event. The sum total of the proteases and their target substrates operating in a physiological pathway therefore defines the global proteolytic signature of that pathway (Overall and Dean 2006). Dysregulation of proteolysis in organisms is deleterious, and many developmental problems and diseases are attributed to aberrant proteolytic activity (Lopez-Otin and Overall 2002). Since the job of proteases is to cleave protein substrates, then a vital part of understanding the role of proteolysis in health and disease is to determine the products of this proteolysis.

Efforts to understand protease and substrate interactions have primarily focused on defining a protease's cleavage-site specificity. Protease specificity has been investigated using synthetic peptide libraries and phage display technology (Ding, Coombs et al. 1995) (Smith, Shi et al. 1995) (Thornberry, Rano et al. 1997) (Stennicke, Renatus et al. 2000) (Deng, Bickett et al. 2000) (Harris, Backes et al. 2000) (Turk, Huang et al. 2001). Although this information is important, it does not directly lead to natural substrate identification, and (paradoxically) has sometimes produced substrate predictions that do not occur naturally (Ding, Coombs et al. 1995). Long lists of potential cleavage-sites are not uncommon results when using a protease's specificity to identify substrates *in silico*. Attempts to validate these predicted cleavage-sites often fail, most likely due to structural restrictions of these sites. The ultimate goal is a direct approach to identify naturally substrates cleaved *in vivo*, but is technically challenging. This approach to identify biologically relevant protease substrates delivers information not on what a protease can do, but on what it does. We reasoned that substrates and cleavage-sites could be determined by identifying the N-terminal sequence of proteolytic products by proteomics.

Proteolytic events can be considered to be either constitutive or regulated. In contrast with regulated proteolytic events that depend on a specific trigger, constitutive proteolytic events remove portions of proteins following translation. Examples of constitutive proteolytic events are methionine removal by MetAPs (methionine aminopeptidases) (Bradshaw, Brickey et al. 1998), signal peptide removal by signal peptidases (Paetzel, Karla et al. 2002) and mitochondrial

transit peptide removal by mitochondrial peptidases (Ito 1999). Protein databases contain most of the predicted proteins expressed in a given species and frequently contain sites predicted to be targeted by these proteases. However, only a fraction of these predictions have been determined experimentally. The challenge is to develop a technology that will determine the magnitude and specificity of these constitutive proteolytic events *in vivo*.

In search of protease activity *in vivo*, we have developed a proteomics approach that combines specific N-terminal tagging of proteins with affinity enrichment and LC (liquid chromatography)–MS/MS (tandem MS) detection. This technique uses readily available reagents and is amenable for use in nearly any proteomics laboratory. In this chapter we present extensive validation of the approach, and demonstrate a series of conventional and unusual proteolytic events as a result of the action of proteases on natural substrates.

RESULTS & DISCUSSION

Method outline and validation. This method takes advantage of the fact that a single proteolytic event generates a new N-terminal amine (see **Fig. 2.1** for method outline). This new amine and the original N-terminal amine of the protein(s) can be specifically labeled by an amine-reactive tag, provided that other amines (e.g. lysine side chains) are blocked. Lysine guanidination using *o*-methylisourea has long been used to aid in the derivatization of amines for

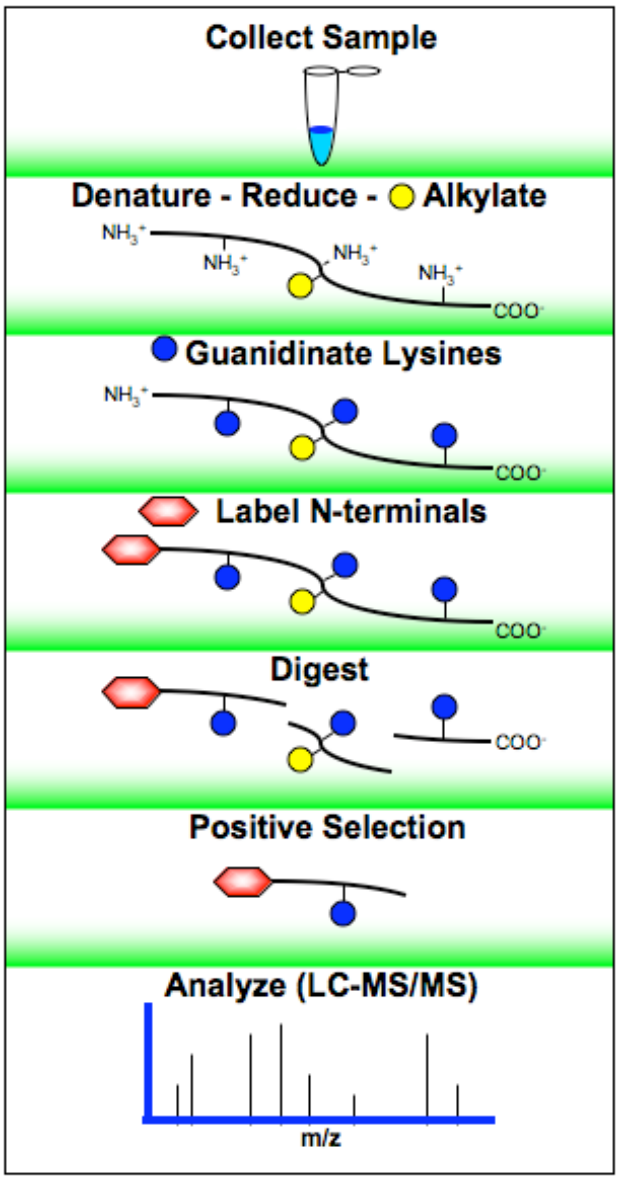


Figure 2.1 Method outline. A protein sample is immediately denatured and reduced to prevent further protease activity. Selective labeling of N-terminals is ensured by Cys alkylation and Lys guanidination. NHS-SS-biotin is covalently coupled to exposed N-terminals. Excess reagent is quenched with glycine and excluded by buffer exchange. The labeled proteins are digested into peptides, which are subsequently captured and enriched by immobilized streptavidin. Positively selected peptides are eluted by DTT, and analyzed by LC-MS/MS. MS/MS spectra are searched against SwissProt using SEQUEST SorcererTM. Peptides with probability scores of at least 0.95 and Xcorr of 2.0 or greater are annotated using SwissProt.

protein MS (Kimmel 1967; Beardsley, Karty et al. 2000; Warwood, Mohammed et al. 2006). We optimized guanidination conditions using purified aprotinin (**Fig. 2.2**, **2.3**). Treatment of the test protein aprotinin with *o*-methylisourea resulted in a mass increase corresponding to guanidination of the four lysine residues (**Fig. 2.2**). Since the lysine signal in amino acid analysis was almost completely suppressed (**Fig. 2.3**), we conclude that only the lysine ε-amines were modified, and any guanidination of the aprotinin N-terminal amine was undetectable. We can not rule out that unwanted guanidination of proteins at the N-terminus may occur, preventing subsequent detection, but our validation experiments with aprotinin suggest that this would be a rare occurrence and would not have a substantial impact on data acquisition. Indeed, previous studies of protein and peptide guanidination revealed very little, if any, N-terminal modification under conditions similar to those of the present study (Kimmel 1967) (Beardsley, Karty et al. 2000) (Warwood, Mohammed et al. 2006). The efficiency of lysine side-chain guanidination was >94%, tested on a mixture of nine purified proteins (results not shown).

Strategic N-terminal labeling using a disulfide-cleavable biotinylated tag enabled enrichment and selection of the tagged peptides following digestion of the protein sample. Potential side reactions of the biotin tag with the side chains of lysine, serine, threonine and histidine residues were found to constitute only 6.6% of all N-terminally tagged peptides (**Fig. 2.4, 2.5**). We identified peptides by

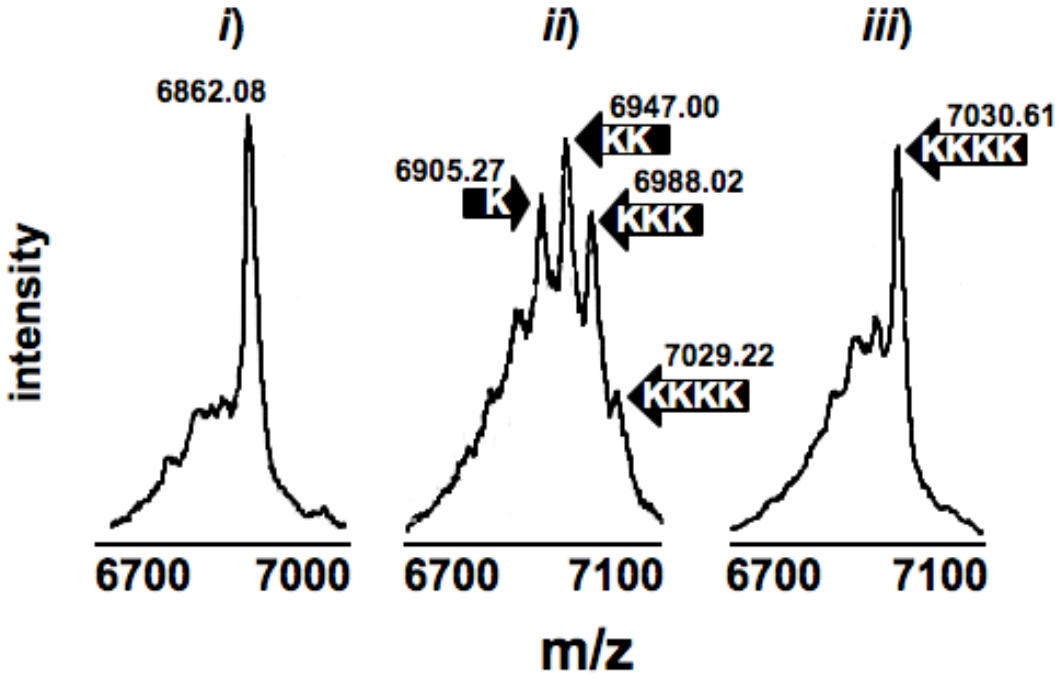


Figure 2.2 Guanidination of aprotinin. Aprotinin was denatured, alkylated and guanidinated (+42Da/Lys residue) at 4ºC for 1h (ii) or overnight (iii). Panel (i) shows MALDI spectra of aprotinin prior to addition of *o*-methylisourea hemisulfate. The main MALDI peak in i) demonstrates the m/z for unmodified aprotinin, while in panel iii), the main peak corresponds to successful modification of all four ε-NH_2 groups of the protein. Note the time-dependent shift in the main mass peak during the guanidination reaction.

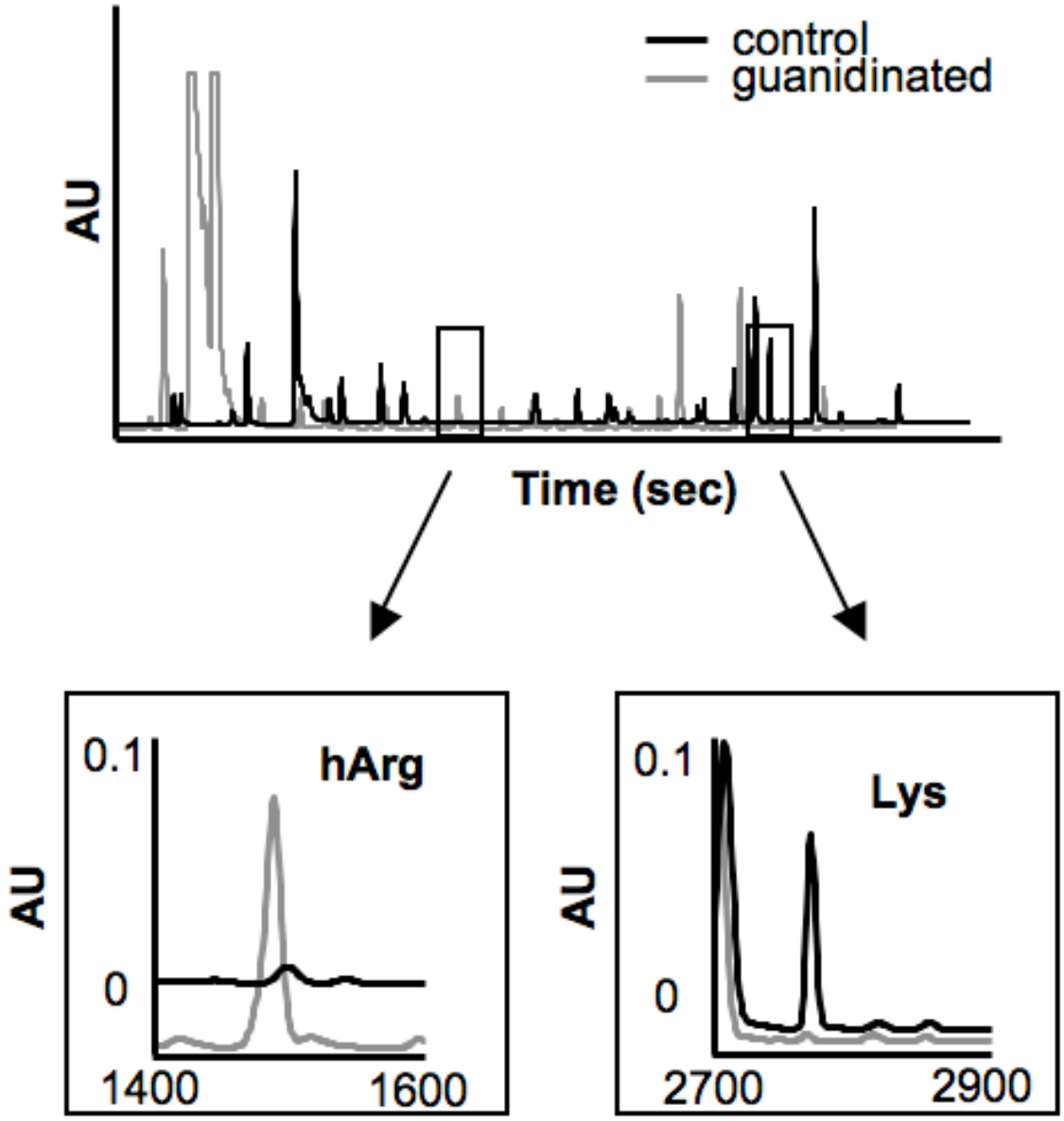
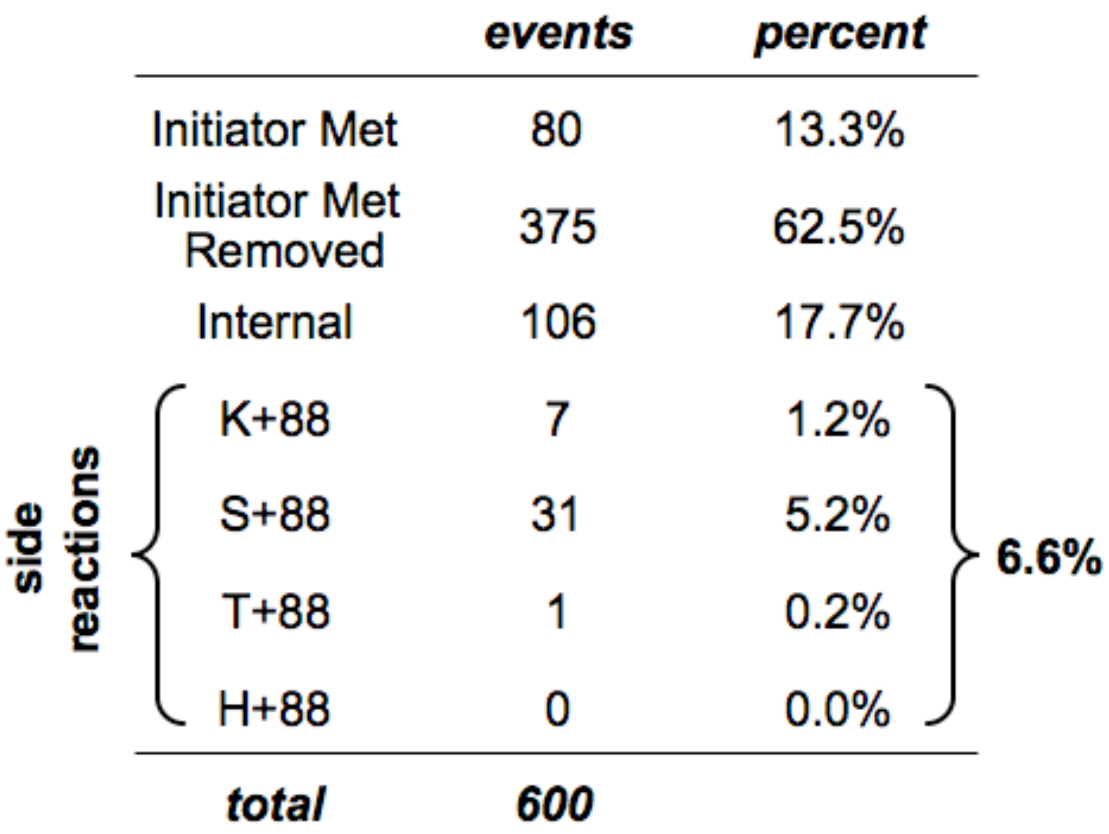


Figure 2.3 Amino acid analysis of guanidinated aprotinin. Guanidinated aprotinin was exchanged into HCl for hydrolysis to amino acids, and analyzed following pre-column derivatization and RPLC. The amino acid analysis revealed a composition characteristic of the protein, with individual peaks matching for untreated (bold line) and guanidinylated samples (light line), with the exception of the Lys peak. Lower panels show blowups of the homoArg (hArg) peak region and the Lys peak region. Note the almost complete suppression of Lys and an accompanying formation of hArg.



		events	percent	
	Initiator Met	80	13.3%	
	Initiator Met Removed	375	62.5%	
	Internal	106	17.7%	
side reactions	K+88	7	1.2%	6.6%
	S+88	31	5.2%	
	T+88	1	0.2%	
	H+88	0	0.0%	
	total	***600***		

Figure 2.4 Quantitation of side-reactions using a purified protein mixture. Peptides were identified from spectra by searching with a fixed modification of all peptide N-terminal amino acids. Side reactions of the biotin tag with unintended amino acid side chains were identified using a differential modification on lysine (K), serine (S), threonine (T), or histidine (H) residues. Less than 7% of peptides identified had an undesired side chain modification with the biotin tag.

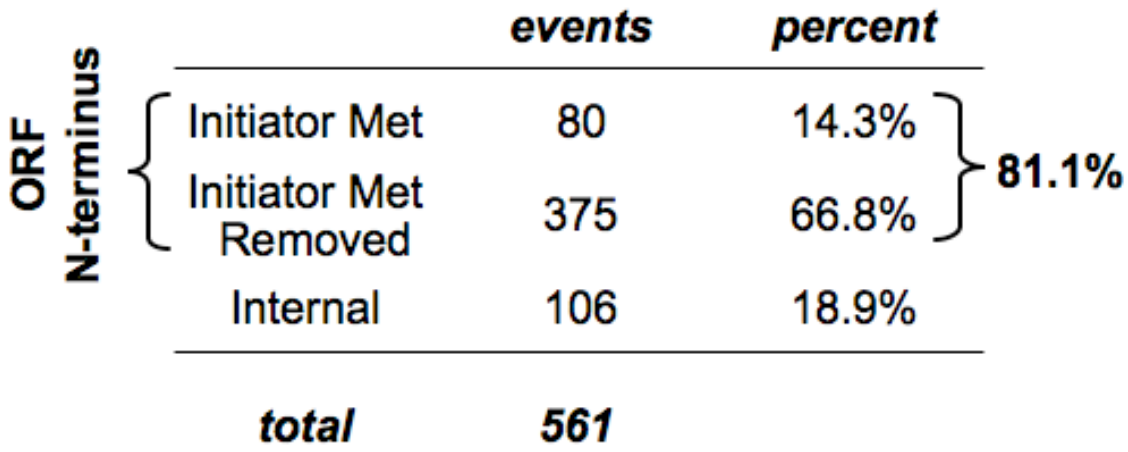


Figure 2.5 Cleavage-site validation using a purified protein mixture. Over 80% of the peptides identified without side chain adducts corresponded to the expected protein N-terminus.

LC–MS/MS and parsed the data by filtering for peptides with the chemically modified N-terminus. Like in other proteomic analyzes of posttranslational modifications, we expected to identify a single peptide per protein, and so the inclusion criteria must be rigorous. To this end, we included peptides with a PeptideProphet probability score of ≥0.95 and cross-correlation (Xcorr) value of ≥2.0, which is extremely stringent (**Fig. 2.6**). Specifically, searching a representative *E. coli* dataset against a forward and reverse database, using these criteria, resulted in a false positive peptide identification rate of only 0.53%. Thus we demonstrate that the accepted spectra were of high quality, and the resulting peptide identifications constituted sufficient evidence to confidently identify proteins of interest. We analyzed each individual sample three times to increase our confidence in detected peptides (owing to higher spectral counts), as well as enhancing the possibility of detecting peptides of low abundance. We also analyzed repeated preparations of each sample type and used both trypsin and Glu-C digestion to increase the total number of peptides and thus again increase spectral counts and our confidence in the MS data. Depending on the origin of the sample analyzed, we obtained an overlap of 50–70% when comparing the peptides found in the same sample run three times on two different occasions, which is comparable with that reported previously (~70% reproducibility in LTQ-MS/MS analysis of yeast samples (Elias, Haas et al. 2005)).

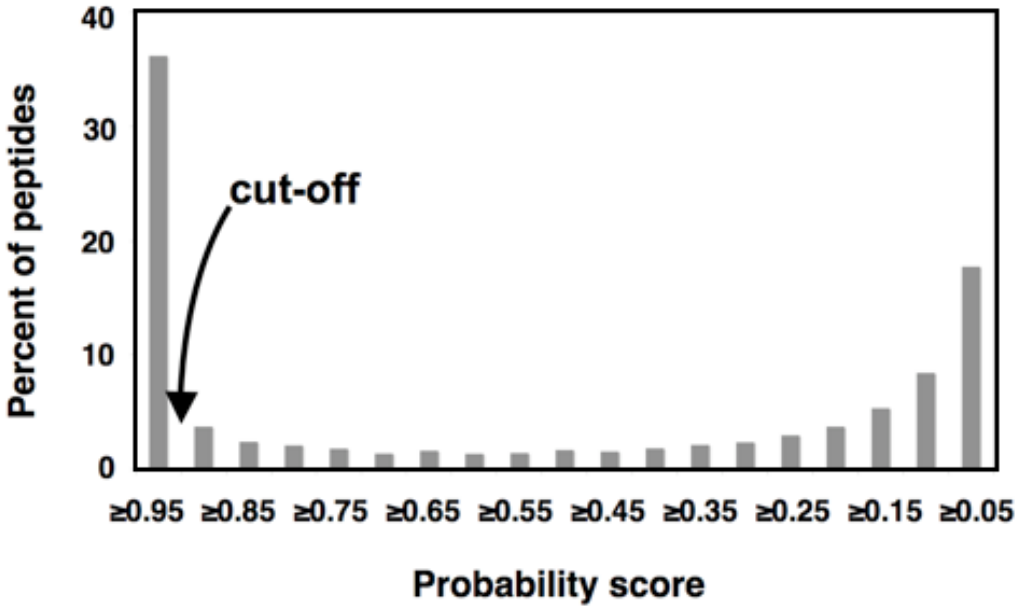


Figure 2.6 Probability score distribution. Distribution of probability score of all observed peptides in a typical *E. coli* sample. The arrow indicates the cut-off (0.95) for peptide assignment. A cross-correlation cut-off at 2.0 was also included.

Identification of N-termini from proteolytically processed and unprocessed proteins validates our methodology. Indeed, our analysis of *E. coli* revealed peptides corresponding to 8.7% of all predicted protein N-termini (365 native protein N-termini and 28 with the signal peptide removed, from 4506 predicted genes). The method confirms predicted proteolytic events *in vivo* and elucidates previously uncharacterized cleavage events. Examples of constitutive proteolytic events are methionine removal by MetAPs (Bradshaw, Brickey et al. 1998), signal peptide removal by signal peptidases (Paetzel, Karla et al. 2002) and mitochondrial transit peptide removal by mitochondrial peptidases (Ito 1999). We set out to profile these constitutive proteolytic events in several biological samples.

Analysis of *E. coli*, yeast, mouse and human proteomes. Peptides were characterized using Swiss-Prot features corresponding to annotated proteolytic events and grouped into functional categories (**Fig. 2.7, 2.8**). Spectra were collected by the mass spectrometer using dynamic exclusion criteria; however, abundant proteins were often identified by multiple spectra corresponding to the same N-terminal peptide. The majority of N-terminal peptides in *E. coli* were derived from open reading frames with or without the initiator methionine, in contrast with eukaryotes, where N-acetyltransferases block approx. 50% of yeast cytosolic proteins and upwards of 80% of mammalian ones (Polevoda and Sherman 2003). Analysis of 365 N-terminal *E. coli* peptides provided a profile for the *in vivo* specificity of MetAP, which demonstrated a strict

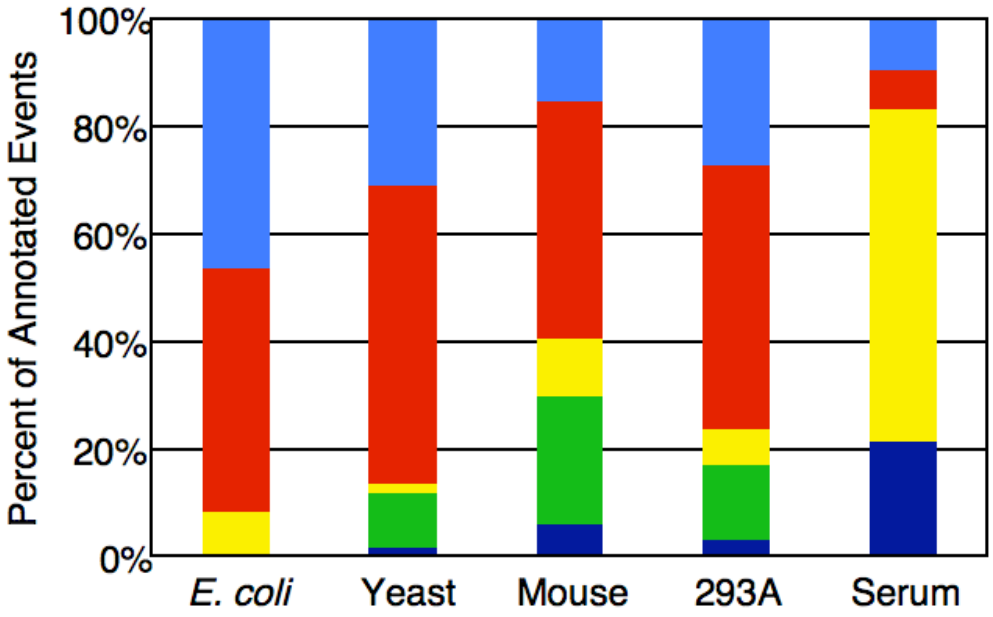


Figure 2.7 Summary of annotated events. Distribution of original N-terminals and proteolytic events from different species. ■ initiator Met, ■ initiator Met removed, ■ signal peptide removed, ■ mitochondrial transit peptide removed, and ■ propeptide removed.

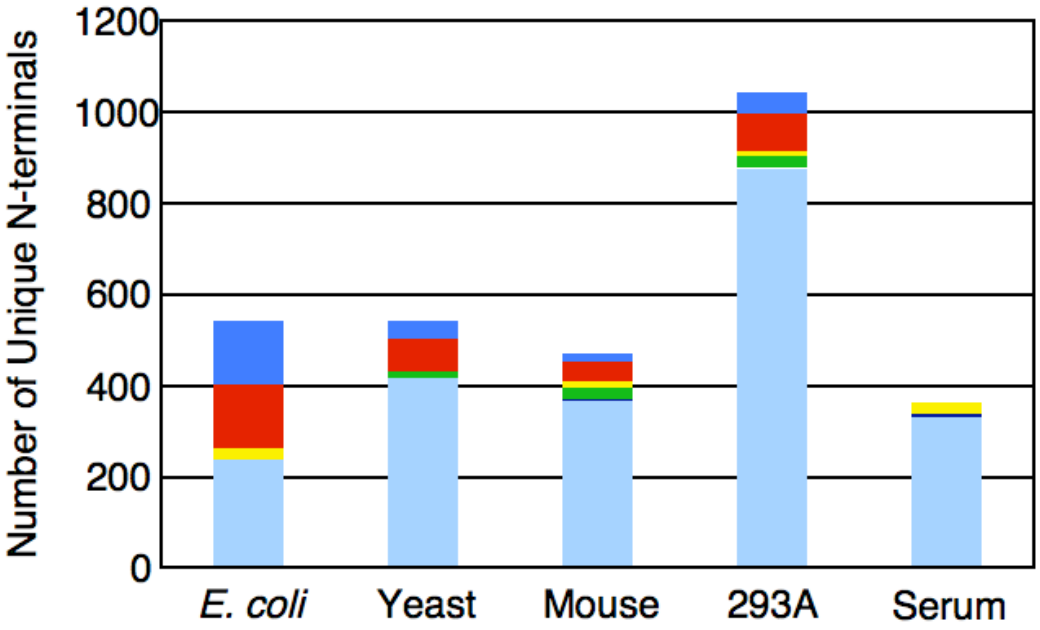


Figure 2.8 Summary of unique N-terminals. Distribution of unique N-terminals observed from different species. ■ initiator Met, ■ initiator Met removed, ■ signal peptide removed, ■ mitochondrial transit peptide removed, ■ propeptide removed, and ■ unascribed peptides.

preference for small and uncharged amino acids in the P1' position (**Fig. 2.9**). Thus our *ex vivo* analysis of natural substrates cleaved *in vivo* supports the *in vitro* analysis of MetAP specificity using synthetic substrates (Hirel, Schmitter et al. 1989) (Frottin, Martinez et al. 2006).

Protein trafficking across membranes is governed by N-terminal sequences that are proteolytically removed upon translocation. Signal peptidases remove the signal peptides required to drive translocation of proteins through the cell membrane in prokaryotes, or the secretion apparatus in eukaryotes. Hallmark features of signal peptides are (i) a basic N-terminus, (ii) a hydrophobic membrane-spanning stretch, and (iii) a C-terminal polar region terminating in an Ala-Xaa-Ala motif (Bendtsen, Nielsen et al. 2004). We analyzed peptides that corresponded to a cleavage site between residues 15 and 50 as a search criterion for signal peptides. In *E. coli*, we observed 28 signal peptide events, of which 21 had been previously determined experimentally and seven predicted. Our sampling of *E. coli* signal peptidase cleavage sites confirmed the accuracy of predictions for this prokaryote. However, we found evidence for miss-annotations of signal peptidase cleavage sites in mammalian samples. Thus predictions of signal peptidase are more accurate in *E. coli* than in mammals. Mitochondrial proteins are imported into the matrix following transit peptide removal by MPP (mitochondrial processing peptidase). Transit peptides are heterogeneous in length, generally contain arginine in the P3 or P2 position, and the P1' residue is

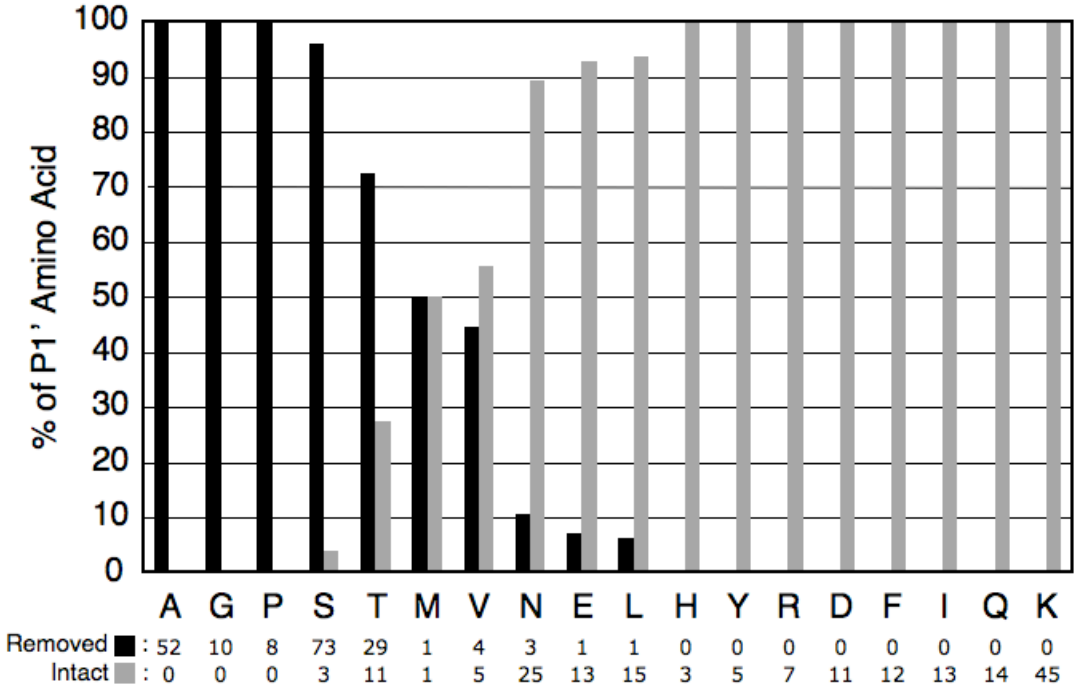


Figure 2.9 *E. coli* MetAP specificity *in vivo*. Extent of MetAP processing depending on the residue that follows the initiator Met, shown as percentage per total for each residue. Black bars show removal of Met, and gray bars show retained Met at protein N-terminals. The total number of unique N-terminals observed per residue are shown below. *E. coli* MetAP displays a strict and efficient processing preference for small and uncharged amino acids in the P1' position.

.

often aromatic or hydrophobic (Gakh, Cavadini et al. 2002). The matrix-localized MIP (mitochondrial intermediate peptidase) can subsequently process imported peptides by removing the N-terminal eight amino acids (Gakh, Cavadini et al. 2002). Additionally, matrix proteins can be translocated to the intermembrane space by IMP (inner membrane protease) 1 and 2, which recognize and process N-terminal transit peptides (Luo, Chen et al. 2003). Many mitochondrial proteins are processed by these proteases; however, the exact cleavage sites follow very loose consensuses, making prediction problematic (Gakh, Cavadini et al. 2002), and are often unknown or inferred from a few well-studied examples. In the yeast, mouse and human datasets, we find N-termini confirming annotated and predicted cleavage sites of mitochondrial transit peptides. Strikingly, we find substantial discrepancies between the predicted and observed limits of mitochondrial transit peptides (**Table 2.1**). Representative high-quality spectra identify new transit peptide-cleavage sites of mitochondrial proteins (**Figure 2.10**). Alternatively, heterogeneity in processing can occur through aminopeptidase activity. In sampling these proteomes, we present a strategy that can be pivotal in refining the authentic cleavage sites used *in vivo* by the several MPPs.

In human serum, we observed hallmarks of specific limited proteolysis of the blood clotting cascade, and considerable trimming of N-termini (**Fig. 2.11**). N-terminal trimming of serum proteins has been observed previously (Nedelkov, Kiernan et al. 2005), and we detected a series of nested N-terminal peptides

Table 2.1 Mitochondrial transit peptides. Disagreements between our data (P1 observed) and the corresponding SwissProt annotations (P1 annotated) were most severe when we analyzed proteins annotated in SwissProt as mitochondrial proteins. In each case the observed cleavage site was consistent with prior removal of the transit peptides that help sort the proteins to mitochondria. In addition to our newly discovered transit peptide cleavage sites (NOVEL) we observed a substantial portion of different cleavage sites (NEW TRANSIT), sometimes consistent with the action of mitochondrial intermediate peptidases (MIP) that generally remove additional octapeptides from proteins localized in the mitochondrial matrix. Occasionally we observed trimming (TRIMMING) of single residues, presumably by aminopeptidases. Proteins with two distinct peptides are highlighted grey. Spectral counts indicate the number of times the unique peptide was identified, and guanidinated lysine, which gives a homoarginine derivative is abbreviated as K#.

YEAST

SwissProt ID	Peptide sequence	Spectral counts	P1 observed	P1 annotated	Description
IDH2_YEAST	F.LATVK#QPSIGR.Y	5	14	15	NEW TRANSIT
IDH2_YEAST	L.ATVK#QPSIGR.Y	26	15	15	ANNOTATED
DHSB_YEAST	G.MATATTAAATHTPR.L	2	18	20	NEW TRANSIT
DHSB_YEAST	M.ATATTAAATHTPR.L	4	19	20	NEW TRANSIT
GLYM_YEAST	G.LLTSGAQSLVSK#PVSEGDPEMFDILQQER.H	3	19	20	NEW TRANSIT
NFU1_YEAST	L.IHIK#TLTTPNE.N	3	21	unknown	NOVEL
ETFA_YEAST	Y.ASTLAFIESSK#DGSVSR.S	12	22	unknown	NOVEL
GCSH_YEAST	N.SSGNALNK#NK#LPFLYSSQGPQAVR.Y	23	22	47	NEW TRANSIT
UCRI_YEAST	L.ISQSLLASK#STYR.T	2	22	30	NEW TRANSIT
ODO2_YEAST	F.K#STSIEVPPMAESLTEGSLK#EYTK#.N	1	71	unknown	NOVEL

MOUSE

SwissProt ID	Peptide sequence	Spectral counts	P1 observed	P1 annotated	Description
KBL_MOUSE	A.HSALAQLR.C	4	17	unknown	NOVEL
ALDH2_MOUSE	L.SAAATSAVPAPNHQPE.V	13	19	19	ANNOTATED
ALDH2_MOUSE	S.AAATSAVPAPNHQPE.V	2	20	19	ANNOTATED + TRIMMING
HMGCL_MOUSE	A.VSTSSMGTLPK#QVK#.I	3	20	27	NEW TRANSIT
OTC_MOUSE	H.FWCGK#PVQSQVQLK#GR.D	2	24	32	NEW TRANSIT
CISY_MOUSE	H.ASASSTNLK#DVLSNLIPK#EQAR.I	3	25	27	NEW TRANSIT
SARDH_MOUSE	L.ATEARPTTEK#SVPYQR.T	4	28	unknown	NOVEL TRANSIT @ 20 + MIP
NDUB8_MOUSE	A.FHMTK#DMLPGSYPR.T	1	29	28	ANNOTATED + TRIMMING
CP27A_MOUSE	A.K#ATIPAALQAQESTEGPGTGQDRPR.L	7	30	32	NEW TRANSIT
DHSA_MOUSE	K.ASAK#VSDAISTQYPVVDHE.F	5	42	43	NEW TRANSIT
DHSA_MOUSE	A.SAK#VSDAISTQYPVVDHE.F	9	43	43	ANNOTATED
DHSA_MOUSE	S.AK#VSDAISTQYPVVDHE.F	1	44	43	ANNOTATED + TRIMMING
ATPA_MOUSE	L.QK#TGTAEMSSILEER.I	17	43	43	ANNOTATED
ATPA_MOUSE	Q.K#TGTAEMSSILEER.I	11	44	43	ANNOTATED + TRIMMING
PRDX3_MOUSE	H.TPAVTQHAPYFK#.G	1	62	63	NEW TRANSIT
DNJA3_MOUSE	A.K#DDYYGQILGVPR.N	5	90	unknown	NOVEL

HUMAN

SwissProt ID	Peptide sequence	Spectral counts	P1 observed	P1 annotated	Description
AA55_HUMAN	G.LHHK#AVLAVR.R	1	19	32	NEW TRANSIT
ATPO_HUMAN	P.FAK#LVRPPVQVYGIEGR.Y	1	23	23	ANNOTATED
ATPO_HUMAN	F.AK#LVRPPVQVYGIEGR.Y	2	24	23	TRANSIT + TRIMMING
CISY_HUMAN	H.ASASSTNLK#DILADLIPK#.E	24	25	27	NEW TRANSIT
DHSA_HUMAN	G.FHFTVDGNK#R.A	1	32	43	NEW TRANSIT
ECH1_HUMAN	L.TGSSAQEEASGVALGEAPDHSYESLR.V	1	33	unknown	NOVEL
EFTS_HUMAN	S.ASASSK#ELLMK#LR.R	2	41	45	NEW TRANSIT
DNJA3_HUMAN	A.K#EDYYQILGVPR.N	13	90	unknown	NOVEL

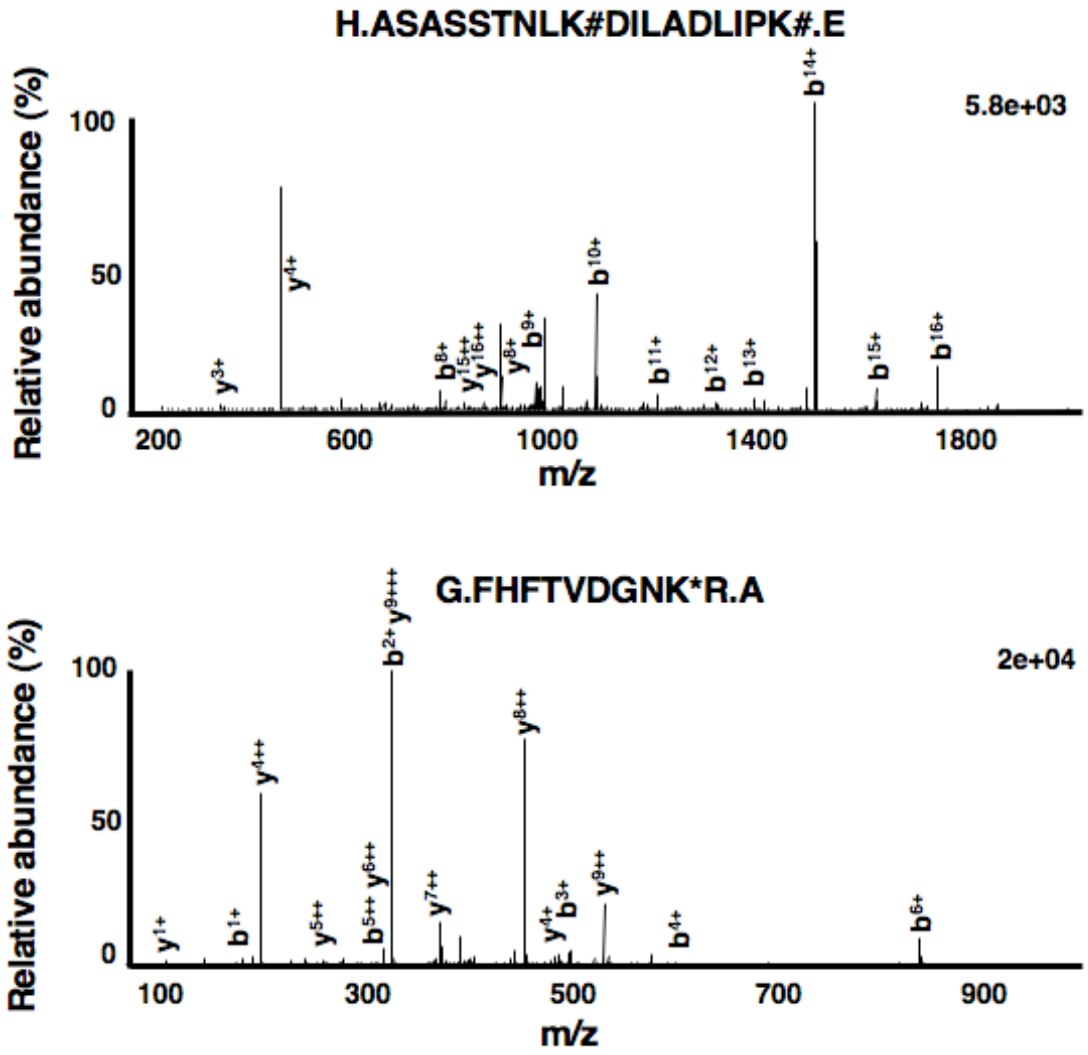


Figure 2.10 Representative MS/MS spectra. MS/MS spectra of two N-terminally labeled peptides, identifying previously unobserved mitochondrial protease cleavage-sites in human 293A cells. Identification of new or novel transit cleavage-sites in CISY_HUMAN (top) and DHSA_HUMAN (bottom). Guanidination of Lys is shown by #.

protein	event	peptide
Alpha-1-antichymotrypsin	signal removed	HPNSPLDEENLTQENQDR↓G…
	trimmed	NSPLDEENLTQENQDR↓G…
Hemopexin	signal removed	TPLPPTSAHGNVAEGETK#PDPDVTER↓C…
	trimmed	PPTSAHGNVAEGETK#PDPDVTER↓C…
		HGNVAEGETK#PDPDVTER↓C…
Alpha-2-HS-glycoprotein	signal removed	APHGPGLIYR↓Q…
	trimmed	HGPGLIYR↓Q …
Alpha-2-macroglobulin	signal removed	SVSGK#PQYMVLVPSLLHTE↓T…
	trimmed	K#PQYMVLVPSLLHTE↓T…
Apolipoprotein-C-II	signal removed	GTQQPQQDEMPSPTFLTQVK#↓E…
	trimmed	TQQPQQDEMPSPTFLTQVK#↓E…
CD5 antigen like	signal removed	ASPSGVRLVGGLHRCE↓G…
	trimmed	SPSGVRLVGGLHRCE…*
		SGVRLVGGLHRCE↓G…
Fibrinogen alpha chain	signal removed	ADSGEGDFLAEGGGVR↓G…
	trimmed	DSGEGDFLAEGGGVR↓G…
		SGEGDFLAEGGGVR↓G…
		EGDFLAEGGGVR↓G…
Haptoglobin	signal removed	VDSGNDVTDIADDGCPK#PPEIAHGYVEHSVR↓Y…
	trimmed	GNDVTDIADDGCPK#PPEIAHGYVEHSVR↓Y…
		DVTDIADDGCPK#PPEIAHGYVEHSVR↓Y…

Figure 2.11 N-terminal trimming of human blood serum proteins. Blood serum contained several proteins with ragged N-terminals. The longest derivative of each corresponds to the N-terminus following signal peptide removal, as annotated in SwissProt, and the trimming is consistent with the activity of cell-associated aminopeptidases and dipeptidyl peptidases. Guanidinated lysine, which gives a homoarginine derivative is abbreviated as K#. * The annotated signal peptide cleavage site for CD5 antigen like protein in SwissProt.

consistent with the action of ectopic cell-surface proteases, including enzymes that remove one residue at a time (aminopeptidase N), or two residues at a time [FAPα (fibroblast activation protein-α) and DPPIV (dipeptidyl peptidase IV)]. These cell-surface proteases have been reported to modify tumor cell behavior (Turner 2004) (Kelly 2005). DPPIV releases a dipeptide from circulating glucagon-like peptides, gastric inhibitory polypeptide and members of the enteroglucagon/GRF (growth-hormone-releasing factor) superfamily, resulting in their biological inactivation. Consequently, DPPIV inhibitors are under clinical trials for therapeutic potential to enhance insulin secretion and overcome Type 2 diabetes (reviewed in (Wiedeman and Trevillyan 2003)). We do not know whether the trimming of the N-termini by FAPα or DPPIV-like activity has a biological consequence, but we are struck by the potential that our discovery may have as a diagnostic biomarker of the efficacy of therapeutic treatment. As expected, cellular proteins are very low in the blood samples, whereas these samples are enriched in signal peptide-cleavage products corresponding to secreted proteins.

Peptides that cannot be readily ascribed to the constitutive proteolytic events described above comprise a substantial portion of all datasets (**Fig. 2.8**). Because they are not annotated in protein sequence databases, these events probably originate from previously undocumented cleavages. They may represent biologically relevant peptides, trimming by aminopeptidases or results

of natural protein turnover. However, it is unlikely that many of them are artifacts arising from sample preparation procedures. In our samples of mixtures of purified proteins, which model a complex proteome, we observed 81% of peptides that represent native protein N-termini (**Fig 2.5**), and therefore >80% of the unascribed peptides in our biological samples are likely to represent the results of *in vivo* proteolysis. With the exception of well-characterized pro-peptides, internal cleavages in proteins are rarely detected by techniques used previously. The frequency of these unascribed cleavage events is unknown, but from our data, one can predict that internal proteolysis may be far more common than appreciated previously. However, in the absence of more direct data, we do not yet wish to ascribe these events to specific biologies.

Characterizing proteolytic profiles using positive selection. The approach we have described to identify N-termini of proteins is closely related in outcome to N-terminal analysis by Edman degradation. Indeed, the N-terminal sequences of 223 *E. coli* proteins were deduced following two-dimensional PAGE followed by Edman degradation of the excised protein spots (Link, Robison et al. 1997). However, this procedure would take months to complete, and is extremely reagent-intensive. Our methodology is complementary to two other MS-based techniques based on negative selection of modified tryptic peptides (Gevaert, Goethals et al. 2003) (McDonald, Robertson et al. 2005). Our more direct strategy, in contrast, enriches N-termini by positive selection. Positive selection has certain advantages over negative selection: (i) we use it as a filter

to simplify datasets, because only N-terminally labeled peptides are counted in our analysis, (ii) datasets are simplified further because all N-acetylated proteins are discarded, and (iii) if the biotin N-terminal probe is replaced by fluoresceinated amine-reactive dyes, it is possible to utilize the protocol to assess differences in constitutive (or even regulated) proteolysis by differential gel electrophoresis (Van den Bergh, Clerens et al. 2003). Although we have focused on constitutive proteolytic events, we realize that simply replacing the N-terminal label with one that is isotope-coded would allow us to quantify regulated proteolytic events, much as described by the COFRADIC (combined fractional diagonal chromatography) negative-selection strategy that employs identification of N-terminally labeled peptides by digestion-mediated incorporation of ^{18}O into the C-terminus (Van Damme, Martens et al. 2005).

Proteolytic cleavage event annotations are scarce in protein databases. This is primarily due to the difficulty in identifying cleavage events, but compounded because often precise cleavage sites are unknown, the acting protease is not defined or the biological relevance is not established. The method of the present study is simple, uses conventional and easily obtainable reagents and is applicable to most proteomics facilities. It substantially facilitates genuine proteolytic event identification in biological samples and reveals the cleavage site location and amino acid sequence. Profiling protease activity *in vivo* can implicate distinct proteases as pathological targets for therapeutics. In addition, specific substrates can be used as biomarkers for diagnosis and early detection

of pathology. Our approach addresses the limitations hindering the current understanding of proteolysis as a post-translational modification in health and disease.

MATERIALS & METHODS

Analysis of aprotinin guanidination. Aprotinin (10.3 mg) was denatured (in 100 mM Hepes, pH 7.5, 50 mM NaCl and 6 M guanidine) and reduced [in 10 mM DTT (dithiothreitol)] at 50ºC for 60 min. Following alkylation (in 50 mM iodoacetamide), sample pH was raised to approx. 10.3 by addition of NaOH. *o*-Methylisourea was added to a final concentration of 0.5 M, the pH was readjusted to 10.3, and the sample was incubated at 4ºC for 1 or 16 h. Samples were prepared for MS using C_{18} ZipTips (Millipore) according to the manufacturer's protocol and spotted on to the MALDI (matrix-assisted laser-desorption ionization) target in 2 μl of matrix solution [10 mg/ml α-cyano-4-hydroxycinnamic acid in 50% acetonitrile/0.1% TFA (trifluoroacetic acid)]. MALDI–MS was performed on an Applied Biosystems Voyager-DE PRO Biospectrometry Workstation. Total amino acid analysis was performed with 250 μg (38.5 nmol) of guanidinated aprotinin. Protein digestion was performed in a vacuum hydrolysis tube (Pierce) in 6 M HCl at 110ºC overnight. Samples were evaporated, dissolved in coupling buffer (acetonitrile/pyridine/trietylamine/water, 10:5:2:3, by vol.) and derivatized with phenylisothiocyanate (Pierce) for 5 min at room temperature (23ºC). Samples were evaporated and dissolved in 50 mM ammonium acetate (pH 6.8). Aliquots (20 μl) were injected on to a Varian

Microsorb C_{18} reverse-phase HPLC column, equilibrated to 52◦C and eluted with a gradient of 100 mM ammonium acetate (pH 6.8) and 50% acetonitrile using a Beckman System Gold HPLC system.

Characterizing chemical derivatization. The labeling reagent NHS-SS-biotin [sulfosuccinimidyl-2-(biotinamido) ethyl-1,3-dithiopropionate] can react with side chains of serine, threonine, histidine and unreacted lysine residues. Since the database search only allows for modification of the N-terminal, MS/MS spectra corresponding to a peptide with a labeled side chain will be unassigned and not detected as a false-positive event. However, major side reaction could possibly cause saturation of the streptavidin and limit detection by MS/MS. To characterize N-terminal peptide recovery and quantify side reactions of the biotin tag, we prepared and analyzed a defined test sample. We expressed and purified nine recombinant proteins with anticipated N-terminal peptides of favorable size after tryptic digestion (see Figures 2C and 2D). The proteins used were human caspase 3 C285A, human caspase 3 D9A/C285A, human caspase 7 C285A, baculoviral p35 C2A, human wild-type and three N-terminal mutants (SGPI, MVPI and ANPR) of Smac (second mitochondrial activator of caspases), and human FADD (Fas-associated protein with death domain). These proteins were combined into a single tube at 1 μM each in a total volume of 500 μl and derivatized and analyzed as described below. The spectra were analyzed for peptides with a fixed N-terminal adduct, as well as potential side reactions.

Sample preparation. *Escherichia coli* strain MG1655 cultures were grown in 2xYT medium [1.6% (w/v) tryptone, 1% (w/v) yeast extract and 0.5% (w/v) NaCl] in baffled flasks with shaking at 200 rev./min at 37ºC to a D_{600} of 0.8 for exponential-phase cultures, or left overnight for stationary-phase cultures. FVB/N wild-type mice were housed and bred in compliance with the NIH (National Institutes of Health) guidelines and the Burnham Institute Animal Research Committee. Female mouse liver, kidney, heart and skeletal muscle were surgically removed following killing, briefly washed in PBS solution and snap-frozen in liquid nitrogen. C57/BL6 mouse peritoneal macrophages were elicited by thioglycolate injection and collected 3 days later, washed in PBS and kept on ice. Yeast strain BY4741 map1Δ (map1::KanMX) was obtained from A.T.C.C. (Manassas, VA, U.S.A.). The map1Δ slow-growth phenotype was rescued by transformation with a single-copy plasmid containing the MAP1 gene under control of 1 kb of the endogenousUAS (upstream activating sequence) (pRS415MAP1) (Zuo, Guo et al. 1995). Yeast transformation was performed with lithium acetate. For analysis, yeast cultures (100 ml) were grown to a D_{600} of 1.0 in YPD [1% (w/v) yeast extract, 2% (w/v) peptone and 2% (w/v) glucose]. Cells were pelleted by centrifugation at 2000 g for 5 min, the pellets were washed once with water and then frozen at −80ºC until analysis.

HEK-293A (human embryonic kidney) cells were grown in DMEM (Dulbecco's modified Eagle's medium) supplemented with 10% fetal bovine serum, 100 units/ml penicillin, 100 µg/ml streptomycin and 2 mM glutamine at

37ºC in a humidified atmosphere containing 5% CO2. The cells were harvested by scraping in cold PBS and washed twice in cold PBS. Cytosolic HEK-293A cell extracts were prepared using hypotonic buffer (20 mM Pipes, pH 7.4, 10 mM KCl, 5 mM EDTA, 2 mM $MgCl_2$ and 4 mM DTT), essentially as described in (Ellerby, Martin et al. 1997). Briefly, the cells were harvested by scraping in cold PBS, washed twice in cold PBS, and incubated in hypotonic buffer for 30 min on ice to induce cell swelling. Extracts were prepared by cell membrane shearing using 20- and, subsequently, 27-gauge needles followed by centrifugation at 1000 g for 20 min. The supernatants were centrifuged a second time and the resulting supernatants were collected and stored at −80ºC.

Human serum samples were prepared from whole blood drawn from the anticubital vein into vacutainer tubes containing either no additive or acid citrate dextrose solution (Becton Dickinson). Whole blood containing no additives was allowed to clot for 15 min at room temperature. Samples were then centrifuged at 1500 g for 15 min at 4ºC to pellet cellular blood components. Serum, obtained from the tube containing no additives, and plasma, obtained from the tube containing acid citrate dextrose solution, were divided into aliquots and stored at −80ºC. The samples were filtered (0.45 µm Whatman filter). Albumin and IgG were removed in some of the samples using ProteoExtractTM albumin/IgG removal kit (Calbiochem), according to the manufacturer's instructions.Guanidine (6 M) and DTT (10 mM) were added immediately after depletion of IgG and

albumin (depleted samples, 60 µl) or after filtering (nondepleted samples, 350 µl).

Labeling procedure. Yeast, *E. coli*, mouse, HEK-293A cell and blood samples were immediately denatured and reduced in 6 M guanidine with 10 mM DTT, and boiled for 10 min to inactivate cellular proteases. Iodoacetamide (30 mM) was added to alkylate cysteine side chains. The pH of each sample was increased to 10.3 with NaOH before adding 0.5 M o-methylisourea. The pH was readjusted to 10.3, and the lysine guanidination reaction was carried out at 4 ◦C for 20 h. The proteins were desalted by buffer exchange (PD-10, Amersham Biosciences) into urea buffer (8 M urea, 50 mM Hepes, pH 7.8, and 50 mM NaCl). The urea stock solution was made fresh, deionized using AG 501-X8 resin (Bio-Rad) and filtered before use. The proteins were subsequently labeled by 5 mM EZ-Link sulfo-NHS-SS-biotin (Pierce Biotechnology) at room temperature for approx. 1 h. This NHS-reactive reagent is specific for the N-terminal of the proteins, i.e. native N-termini and proteolytic cleavage sites, owing to the previous blocking of cysteine and lysine side chains, and the biotin tag of the molecule allows for positive selection by immobilized streptavidin. Eventually, unreacted biotin reagent was quenched by the addition of 50 mM glycine for 30 min and subsequently excluded by buffer exchange into 4 M urea, 100 mM Hepes, pH 7.8, and 100 mM NaCl. The samples were diluted 1:2 with distilled water before digestion overnight by sequencing-grade modified trypsin (Promega) or endoproteinase Glu-C (Roche). Further enzymatic activity was

inhibited by boiling the samples for 5 min. The samples were centrifuged at 10,000 g for 5 min, and the resulting supernatant was added to immobilized streptavidin (Pierce Biotechnology) for 1 h at room temperature for selection and enrichment of the biotinylated N-terminal peptides. The streptavidin beads were washed extensively with AmmBic buffer (50 mM triethylammonium bicarbonate, pH 7.8), high-salt AmmBic buffer (AmmBic buffer containing 1 M NaCl) and finally by AmmBic buffer again. The flow-through and first AmmBic wash were collected and allowed to bind new streptavidin beads to minimize eventual loss of labeled peptides. The labeled peptides were eluted by the addition of 50 mM DTT, which cleaves the disulfide-linked biotin tag, leaving an 88 Da addition to the N-termini of the labeled peptides. The peptide elution was dried by vacuum to reduce sample volume, and desalted using C_{18} OMIX tips (Varian), according to the manufacturer's instructions. The peptide solution was again dried by vacuum and re-dissolved in 0.1% TFA for analysis by LC–MS/MS.

Sample analysis by nano LC–MS. The automated NanoLC-LTQ system consists of an Eksigent Nano-2D LC autosampler, a switch valve, a C18 trap column (Agilent), a capillary separation column (100 μm internal diameter x 10 cm length, packed with Synergi 4 μm C_{18}), and an LTQ ion-trap mass spectrometer (Thermo Electron). The separation column is mounted into the Finnigan Nanospray II ion source (Thermo Electron) and used as the electrospray tip as well. First, trypsin- or Glu-C-digested peptides (5–9 μl) were loaded by autosampler on to the trap column in 100% solvent A [2% acetonitrile

and 0.1% methanoic (formic) acid] using a flow rate of 10 µl/min for 4 min. After sample loading and washing, the valve was switched, and the gradient was delivered to the trap and separation column at 500 nl/min. Peptides were separated with a 100–120 min linear gradient of 10–60% solvent B (80% acetonitrile and 0.1% methanoic acid), and was then eluted directly into the LTQ spectrometer. The fully automated NanoLC LTQ was operated via an Instrument Method of Xcalibur.MS/MS spectra were collected automatically during the LC–MS runs. Each scan was set to acquire a full MS scan followed by four MS/MS scans of the four most intense ions from the preceding MS scan.

Database searching. After data acquisition, MS/MS spectra were then extracted and searched against the corresponding protein database (Swiss-Prot) using SEQUEST SorcererTM (SageN). For *E. coli* samples, a non-enzymatic peptide database was used. For human, mouse and yeast samples, semi-tryptic or Glu-C peptide databases were used. A molecular mass of 88 Da was added to the static search of all N-termini to account for NHS-SS-biotin modification. A molecular mass of 57 Da was added to all cysteine residues to account for carboxyamidomethylation. A differential search of amino acids includes methionine +16 Da for oxidation and lysine +42 Da for guanidination. In some searches, we included +88 Da for possible side reactions of the biotinylation probe of serine, threonine, histidine or lysine residues. After SEQUEST searching, the results were filtered automatically, organized and displayed by PeptideProphet and ProteinProphet (ISB), which are installed in SorcererTM. A

minimum probability score of 0.95 was set to assure low errors in peptide identification (see the Results and discussion section). All peptides must have cross-correlation values (Xcorr) of at least 2.0. The false-positive peptide identification rate was quantified using forward and reverse database searching, and was found to be only 0.53%.

N-terminal peptide annotation. Bioinformatic analysis was essential to interpret the thousands of high quality peptides identified from the relevant SEQUEST database searches. N-terminal peptides corresponding to protein N-termini or protease cleavage-sites were annotated with scripts written in-house. The workflow proceeded as follows: spectra satisfying the threshold criteria (probability score ≥ 0.95, xcorr ≥ 2.0), which identified a unique N-terminus were counted (spectral counts) and condensed to a non-redundant list, while keeping track of the frequency distribution of spectra per N-terminus per sample. In this way, the relative abundance of an N-terminus could be compared between different samples. This master list of all N-termini was then annotated by referencing the Swiss-Prot/Uniprot database. The protein name, sequence, and relevant cleavage-site features were retrieved and input into the master list. The script then identified the N-terminal position of the identified peptide sequence in the full-length protein, and compared this site to the features from Swiss-Prot/Uniprot. In this way N-termini were annotated as “initiator Met”, “initiator Met removed”, “signal peptide removed”, “mitochondrial transit peptide removed”, or propeptide removed”. If no annotation matched the cleavage-site, then these

sites were termed “unascribed cleavage-sites”. The script concluded by counting the total number of spectra and the number of non-redundant N-termini, and tallying the breakdown of N-termini for each annotated category as well as unascribed cleavage-sites. This analysis facilitated rapid assessment of the proteomic results and the functional breakdown of N-termini.

ACKNOWLEDGEMENTS

The work in this chapter was supported by NIH (National Institutes of Health) grant RR19752, and by NIH Roadmap Initiative National Biotechnology Resource Center grant RR20843 for the Center on Proteolytic Pathways, CA69381 from the NCI. This chapter was reproduced with permission from co-authors, modified from the publication: Timmer JC, Enoksson M, Wildfang E, Zhu W, Igarashi Y, Denault JB, Ma Y, Dummitt B, Chang YH, Mast AE, Eroshkin A, Smith JW, Tao WA, Salvesen GS. Profiling constitutive proteolysis *in vivo*. *Journal of Biochemistry.* October 2007: 407(1):41-8.

REFERENCES

Beardsley, R. L., J. A. Karty, et al. (2000). "Enhancing the intensities of lysine-terminated tryptic peptide ions in matrix-assisted laser desorption/ionization mass spectrometry." Rapid Commun Mass Spectrom 14(23): 2147-53.

Bendtsen, J. D., H. Nielsen, et al. (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol 340(4): 783-95.

Bradshaw, R. A., W. W. Brickey, et al. (1998). "N-terminal processing: the methionine aminopeptidase and N alpha-acetyl transferase families." Trends Biochem Sci 23(7): 263-7.

Deng, S. J., D. M. Bickett, et al. (2000). "Substrate specificity of human collagenase 3 assessed using a phage-displayed peptide library." J Biol Chem 275(40): 31422-7.

Ding, L., G. S. Coombs, et al. (1995). "Origins of the specificity of tissue-type plasminogen activator." Proc Natl Acad Sci USA 92: 7627-7631.

Elias, J. E., W. Haas, et al. (2005). "Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations." Nat Methods 2(9): 667-75.

Ellerby, H. M., S. J. Martin, et al. (1997). "Establishment of a cell-free system of neuronal apoptosis: comparison of premitochondrial, mitochondrial, and postmitochondrial phases." J Neurosci 17(16): 6165-78.

Frottin, F., A. Martinez, et al. (2006). "The proteomics of N-terminal methionine cleavage." Mol Cell Proteomics.

Gakh, O., P. Cavadini, et al. (2002). "Mitochondrial processing peptidases." Biochim Biophys Acta 1592(1): 63-77.

Gevaert, K., M. Goethals, et al. (2003). "Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides." Nat Biotechnol 21(5): 566-9.

Harris, J. L., B. J. Backes, et al. (2000). "Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries." Proc Natl Acad Sci U S A 97(14): 7754-9.

Hirel, P. H., M. J. Schmitter, et al. (1989). "Extent of N-terminal methionine excision from Escherichia coli proteins is governed by the side-chain

length of the penultimate amino acid." Proc Natl Acad Sci U S A 86(21): 8247-51.

Ito, A. (1999). "Mitochondrial processing peptidase: multiple-site recognition of precursor proteins." Biochem Biophys Res Commun 265(3): 611-6.

Kelly, T. (2005). "Fibroblast activation protein-alpha and dipeptidyl peptidase IV (CD26): cell-surface proteases that activate cell signaling and are potential targets for cancer therapy." Drug Resist Updat 8(1-2): 51-8.

Kimmel, J. R. (1967). "Guanidination of proteins." Meth. Enzymol. 11: 584-9.

Link, A. J., K. Robison, et al. (1997). "Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12." Electrophoresis 18(8): 1259-313.

Lopez-Otin, C. and C. M. Overall (2002). "Protease degradomics: a new challenge for proteomics." Nat Rev Mol Cell Biol 3(7): 509-19.

Luo, W., X. Chen, et al. (2003). "Factors governing nonoverlapping substrate specificity by mitochondrial inner membrane peptidase." J Biol Chem 278(7): 4943-8.

McDonald, L., D. H. Robertson, et al. (2005). "Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides." Nat Methods 2(12): 955-7.

Nedelkov, D., U. A. Kiernan, et al. (2005). "Investigating diversity in human plasma proteins." Proc Natl Acad Sci U S A 102(31): 10852-7.

Overall, C. M. and R. A. Dean (2006). "Degradomics: systems biology of the protease web. Pleiotropic roles of MMPs in cancer." Cancer Metastasis Rev 25(1): 69-75.

Paetzel, M., A. Karla, et al. (2002). "Signal peptidases." Chem Rev 102(12): 4549-80.

Polevoda, B. and F. Sherman (2003). "N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins." J Mol Biol 325(4): 595-622.

Rawlings, N. D., F. R. Morton, et al. (2006). "MEROPS: the peptidase database." Nucleic Acids Res 34(Database issue): D270-2.

Salvesen, G. S. and V. M. Dixit (1997). "Caspases: intracellular signaling by proteolysis." Cell 91(4): 443-6.

Smith, M., L. Shi, et al. (1995). "Rapid identification of highly active and selective substrates for stromelysin and matrilysin using bacteriophage peptide display libraries." J Biol Chem 270: 6440-6449.

Stennicke, H. R., M. Renatus, et al. (2000). "Internally quenched fluorescent peptide substrates disclose the subsite preferences of human caspases 1, 3, 6, 7 and 8." Biochem J 350(Pt 2): 563-568.

Thornberry, N. A., T. A. Rano, et al. (1997). "A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis." J Biol Chem 272(29): 17907-11.

Turk, B. E., L. L. Huang, et al. (2001). "Determination of protease cleavage site motifs using mixture-based oriented peptide libraries." Nat Biotechnol 19(7): 661-7.

Turner, A. J. (2004). Membrane alanyl aminopeptidase. Handbook of Proteolytic Enzymes. A. J. Barrett, N. D. Rawlings and J. F. Woessner. London, Elsevier: 289-284.

Van Damme, P., L. Martens, et al. (2005). "Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis." Nat Methods 2(10): 771-7.

Van den Bergh, G., S. Clerens, et al. (2003). "Fluorescent two-dimensional difference gel electrophoresis and mass spectrometry identify age-related protein expression differences for the primary visual cortex of kitten and adult cat." J Neurochem 85(1): 193-205.

Warwood, S., S. Mohammed, et al. (2006). "Guanidination chemistry for qualitative and quantitative proteomics." Rapid Commun Mass Spectrom 20(21): 3245-56.

Wiedeman, P. E. and J. M. Trevillyan (2003). "Dipeptidyl peptidase IV inhibitors for the treatment of impaired glucose tolerance and type 2 diabetes." Curr Opin Investig Drugs 4(4): 412-20.

Zuo, S., Q. Guo, et al. (1995). "Evidence that two zinc fingers in the methionine aminopeptidase from Saccharomyces cerevisiae are important for normal growth." Mol Gen Genet 246(2): 247-53.

Chapter III

Deciphering Proteolysis in Signaling

ABSTRACT

The structural repertoire and kinetic threshold distinguishing legitimate signaling substrates are fundamental questions in proteolytic networks and pathways. We used N-terminal proteomics to address these issues by identifying cleavage-sites within the *E. coli* proteome driven by the apoptotic signaling protease caspase-3 and the bacterial protease GluC. Defying the dogma that proteases cleave primarily in natively unstructured loops, we found that caspase-3 and GluC cleave in α-helices nearly as frequently as extended loops. Strikingly, biochemical and kinetic characterization revealed that *E. coli* caspase-3 substrates were greatly inferior to natural substrates, suggesting protease/substrate co-evolution. Engineering an *E. coli* substrate to match natural catalytic rates defined a kinetic threshold depicting a signaling event. This unique combination of proteomics, biochemistry, kinetics and substrate engineering reveals new insights into the structure-function relationship of protease targets and their validation from large-scale approaches.

INTRODUCTION

Proteases are prominent components of biological systems in health and disease, often functioning as signaling molecules in networks and pathways – reviewed in (Puente, Sanchez et al. 2003; Salvesen and Abrams 2004). In these settings, proteases signal through limited proteolysis of discrete substrate pools, transmitting information by altering substrate localization, regulation, or activity. Identifying endogenous protease substrates and their cleavage-sites is paramount to delineating their downstream molecular signaling pathways (Gevaert, Impens et al. 2007; Timmer and Salvesen 2007). However, the critical substrate repertoire of most proteases is incompletely known or in many cases completely lacking. Unlike the stringent specificity of restriction endonucleases, proteases are enzymes with varying degrees of specificity and selectivity not solely influenced by a substrate's cleavage-site amino acid sequence.

There are four main determinants influencing which substrates a protease cleaves and where cleavage-sites are. *Spaciotemporal co-localization*: proteases and substrates must be simultaneously present in the same sub-cellular compartment or physical space to be biologically relevant. *Exosite interactions*: some proteases and substrates utilize surfaces distinct from the active site and cleavage-site to drive affinity and selectivity. *Sub-site specificity*: many proteases display some degree of amino acid specificity in positions adjacent to the scissile bond, which provide varying levels of substrate cleavage-site selectivity. Motifs have been proposed for many proteases, based on information from synthetic

peptide libraries or phage display technology (Ding, Coombs et al. 1995; Smith, Shi et al. 1995; Thornberry, Rano et al. 1997; Deng, Bickett et al. 2000; Harris, Backes et al. 2000; Stennicke, Renatus et al. 2000; Nazif and Bogyo 2001; Turk, Huang et al. 2001). One of the best examples of this is the caspase family that maintains a strict requirement for aspartate in the P1 position, with distinctions in the P4, P3, P2, and P1' positions (Thornberry, Rano et al. 1997; Stennicke, Renatus et al. 2000). Structural presentation: proteases are thought to cleave substrates in flexible solvent exposed loops (Hubbard, Campbell et al. 1991; Coombs, Bergstrom et al. 1998), and have been used historically to map protein domain limits. Crystal structures of protease inhibitors with prominent cleavage-site loops have also reinforced this view (Gettins 2002; Kelly, Laskowski et al. 2005); however, disproportionately few substrate structures have been solved that include residues encompassing the cleavage-site. The likely explanation for this is that cleavage-sites are in flexible and unstructured regions of proteins, also known as disordered regions, and are inherently difficult to crystallize. Thus the key is to identify protease cleavage-sites from a library that samples both structural conformation and amino acid distribution to discriminate between sub-site amino acid preference and cleavage-site structure.

Based on these current concepts, a clear hypothesis emerges stating that limited proteolysis of any proteome should be dominated by cleavages in unstructured regions. We endeavored to assess the contributions of both amino acid composition and structural presentation to substrate proteolysis using N-

terminal proteomics (N-terminomics) on the non co-evolved, structurally intact, and well-characterized *E. coli* proteome, which allowed us to identify cleavage-sites on a global scale. We established two experimental paradigms. In the unbiased paradigm – the *E. coli* proteome challenged with the human protease caspase-3 and the Staphylococcal protease glutamyl endopeptidase (GluC) - there could be no pre-selection of substrates and proteases during evolution. In the biased paradigm - specific human caspase substrates challenged with human caspase-3 – we expect co-evolution of substrate sequences and structural requirements to fit the biological role of the caspase. An additional advantage of the *E. coli* proteome is that many of the potential substrates have previously reported structures, which we used to elucidate the structural conformation of the cleavage-sites found. Finally, we were able to propose threshold criteria of protease/substrate co-evolution. These studies offer fundamental insights into the structure-function relationship of proteases and substrates, and reveal a critical catalytic threshold distinctive of natural substrates, which non-evolved substrates fail to meet.

RESULTS

Experimental approach & workflow. To elucidate the thee-dimensional conformation of protease cleavage-sites, we assembled an *in vitro* system comprising two purified proteases: human caspase-3 and Staphylococcal GluC; and an evolutionarily unbiased library of protein substrates composed of soluble *E. coli* lysate. Many measures were taken to maintain the structural integrity of

proteins within the *E. coli* lysate, yet we saw no indication at any point in the process that proteins were denatured. We screened for cleavage-sites by combining our proteases-of-interest with the *E. coli* lysate, and assayed for cleavage-site peptides using N-terminomics, a method we recently reported to identify protease cleavage-sites in complex biological samples (Timmer, Enoksson et al. 2007). Briefly, proteases cleave substrates to expose free N-terminal a-amines, which can be specifically labeled with a cleavable biotin tag. After digestion with trypsin, biotinylated peptides are enriched on neutravidin-linked agarose, eluted, and analyzed by LC-MS/MS. The proteome-wide coverage and structurally unbiased nature of N-terminomics suits it perfectly to identify protease cleavage-sites, which can then be mapped back onto the structures of substrates residing in the Protein Data Bank (PDB), thereby elucidating the structure encompassing the cleavage-site (**Fig. 3.1**). Experiments were conducted using a range of protease concentrations spanning two orders of magnitude.

N-terminomic results & cleavage-site identification. High quality peptides from each experiment were determined by searching a concatenated forward and reverse *E. coli* semi-tryptic database, resulting in peptide false discovery rate of less than 2% (Elias, Haas et al. 2005). The resulting peptide lists corresponded to protease cleavage-sites in *E. coli* proteins as well as unblocked protein N-termini. We used a three-pronged approach to discern

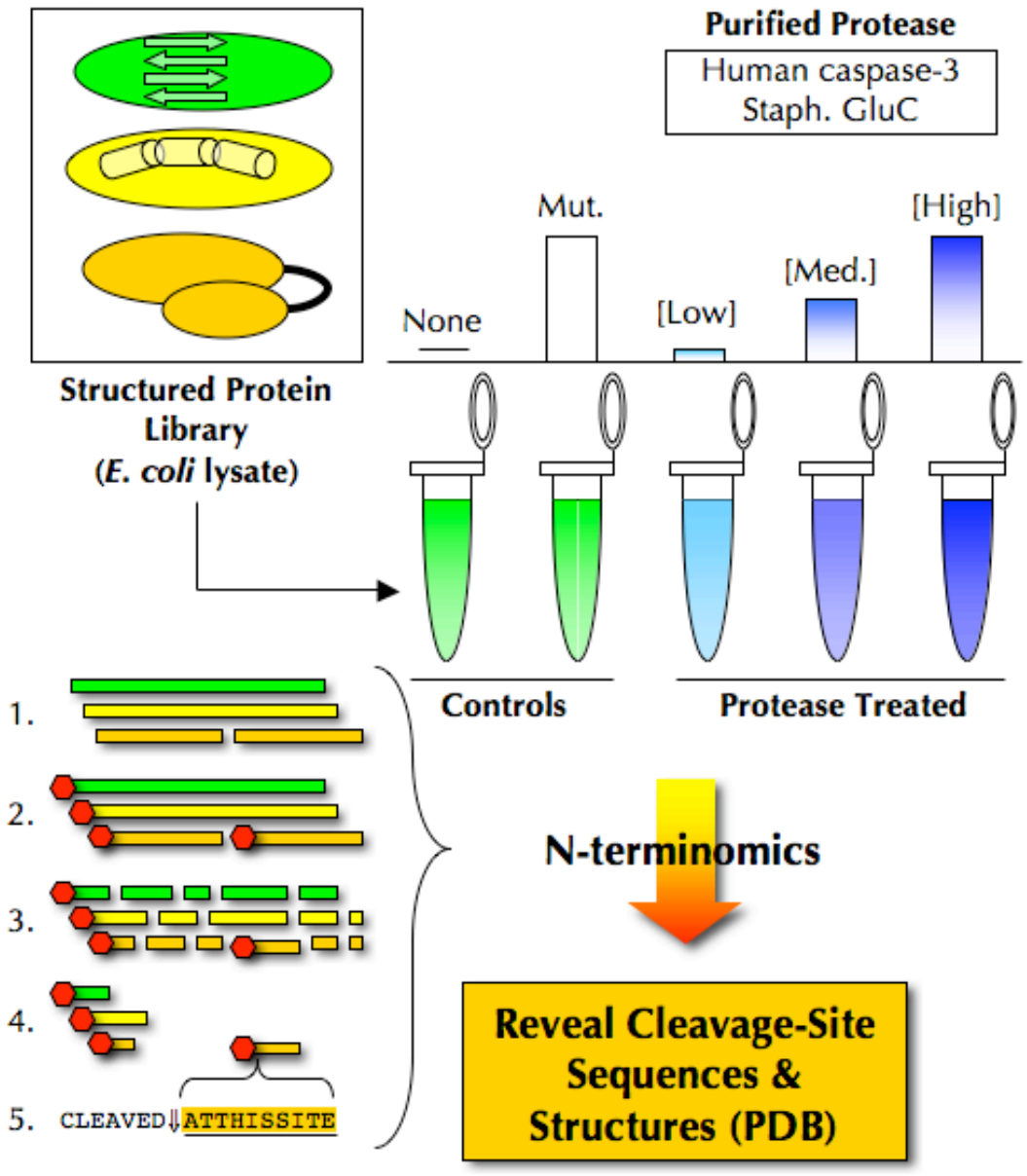


Figure 3.1 Experimental approach. Samples of a library of structured proteins (E. coli soluble lysate) is treated with a range of exogenous protease concentrations (see text for details), and screened for cleavage-sites using N-terminomics. The location of cleavage sites is determined, and compared to available structures deposited in the Protein Data Bank, revealing the relationship between proteolysis and secondary structure.

cleavage-sites of the purified proteases from the substantial background of endogenous proteolytic events (**Fig. 3.2, Table 3.1**). Cleavage-sites produced from the addition of exogenous proteases to *E. coli* lysate would differ from annotated cleavage-sites of endogenous proteases, falling into the category of unascribed cleavage-sites. This step was primarily used to exclude cleavage-sites originating from co-translational protein processing, such as initiator methionine excision and signal peptide removal. Protease-of-interest cleavage-sites were also excluded if they were concurrently found in control samples representing lysate alone, or in the case of caspase-3, treated with a catalytically inactive mutant of the protease. The last stipulation for identifying protease-of-interest cleavage-sites was that the strict P1 amino acid specificity for aspartate (Asp) for caspase-3, or glutamate (Glu) for GluC, must be preserved. Indeed, the stringent primary (P1) specificity of these proteases greatly simplifies the task of discerning true protease-of-interest cleavage-sites from other unascribed cleavage-sites not found in control samples due to incomplete sampling. Lists of caspase-3 and GluC cleavage-sites in *E. coli* proteins are shown in **Tables 3.2, 3.3**.

To verify the purity and activity of the proteases added to *E. coli* lysate, we assessed the frequency of amino acids in the P1 position for all unascribed cleavage-sites comparing control and protease treated samples (**Fig. 3.3, 3.4**). These data confirm the P1 specificity of human caspase-3 is enriched for Asp, and GluC is enriched for Glu with no other obvious enrichments. The frequency

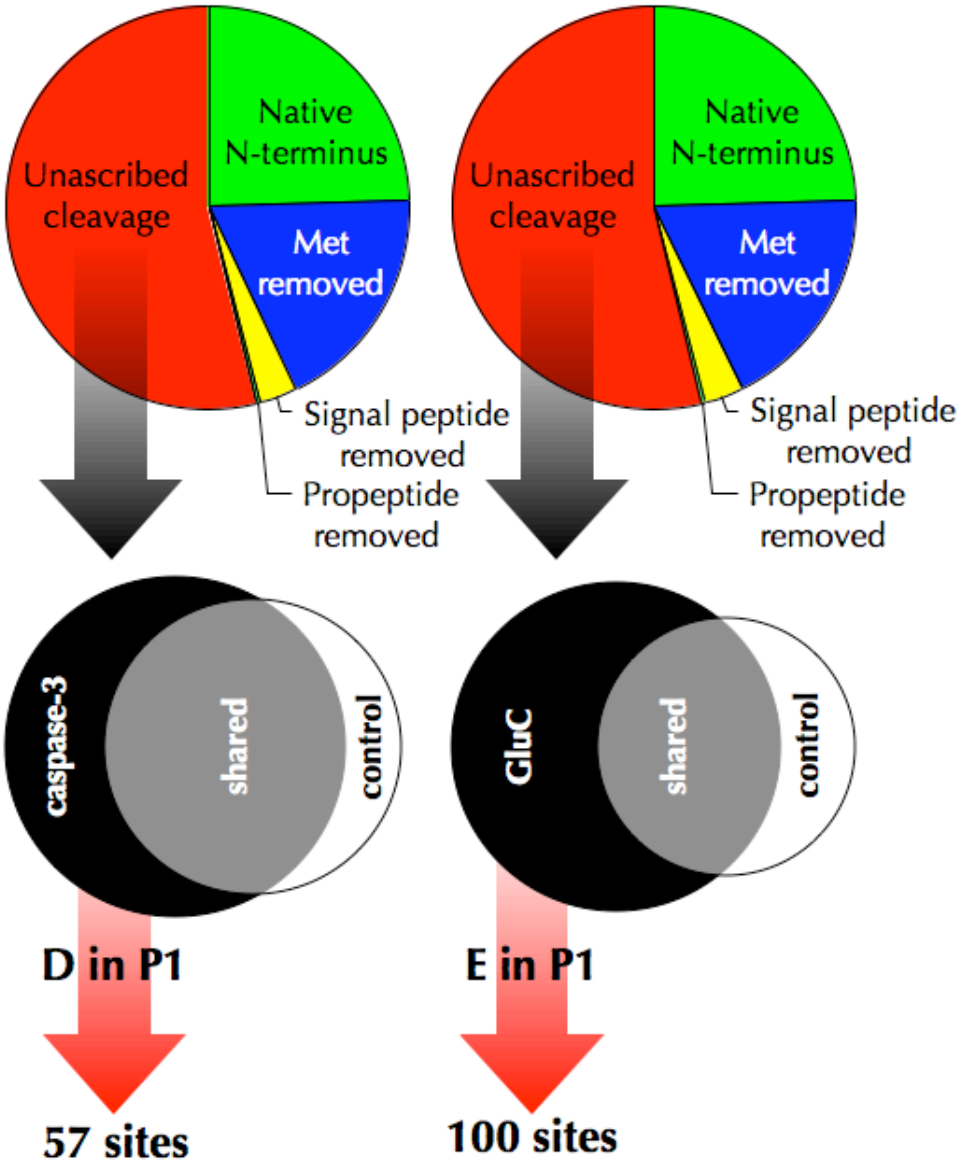


Figure 3.2 N-terminomics reveals protease specific cleavage-sites. N-terminomic analysis of *E. coli* lysate treated with (a) human caspase-3 and (b) GluC reveals specific substrates and corresponding cleavage-sites. A tripartite criterion was used to identify genuine cleavage-sites of exogenous proteases from background proteolytic events inherent in *E. coli* lysate: Cleavage-sites must (1) not already be annotated as an endogenous site of proteolysis, (2) be only found in protease treated samples, and never in control samples, and (3) maintain the P1 aspartate/glutamate characteristic of each protease. See **Table 3.3** for definitions.

Table 3.1 Total and non-redundant peptide spectra identified by N-terminomics. N-terminal peptides corresponding to protease cleavage-sites as well as unblocked protein N-termini are sub-divided based on their proteolytic processing. Protein N-termini retaining the initiator Met are designated “Native N-terminus”, whereas those processed by methionine aminopeptidase are termed “Met removed”. Periplasmic secreted proteins that have been processed by signal peptidases are grouped as “Signal peptide removed”, and likewise proteins with annotated propeptide cleavages are shown as “Propeptide removed”. All other N-termini correspond to internal cleavage-sites that have not been annotated, and thus are termed “Unascribed cleavage”. Protease-of-interest cleavage-sites were found by sieving this last group for cleavage-sites found only in the protease treated samples and cleaved after an Asp for human caspase-3, or a Glu for GluC.

	human caspase-3		Staphylococcal GluC	
Category	NR N-term	Spectra	NR N-term	Spectra
Native N-terminus	253	12,533	196	8,066
Met removed	209	9,991	148	7,448
Signal peptide removed	37	1,178	26	659
Propeptide removed	1	24	1	5
Unascribed cleavage	598	8,133	434	4,735
Total	**1,098**	**31,859**	**805**	**20,913**
Protease only	162	947	263	2,195
D or E in P1	57	490	100	767

Table 3.2 Caspase-3 cleavage-sites in *E. coli* proteins identified by N-terminomics. The names and SwissProt IDs of *E. coli* proteins cleaved by caspase-3 are shown with the cleavage-site shown from P4-P4', and the P1 amino acid number.

protein name	*SwissProt ID*	*cut site*	*obs P1*	*protein name*	*SwissProt ID*	*cut site*	*obs P1*
50S ribosomal protein L21	P0AG48	PFVD-GGVI	55	Elongation factor G	P0A6M8	CDPD-APII	401
30S ribosomal protein S1	P0AG67	DEVD-VALD	69	Adenylosuccinate lyase	P0AB89	DELD-HNWE	379
Alkyl hydroperoxide reductase subunit C	P0AE08	FVVD-PQGI	126	30S ribosomal protein S18	P0A7T7	DYKD-IATL	25
30S ribosomal protein S4	P0A7V8	LEVD-AGKM	174	Enolase	P0A6P9	TIAD-LAVG	382
UPF0381 protein yfcZ	P0AD33	CCMD-VGTI	16	Isocitrate lyase	P0A9G6	TDAD-AADL	236
Elongation factor G	P0A6M8	EVHD-GAAT	45	30S ribosomal protein S1	P0AG67	VLVD-AGLK	39
30S ribosomal protein S1	P0AG67	TAVD-AKGA	462	DNA-directed RNA polymerase subunit beta'	P0A8T7	CDTD-FGVC	891
Entericidin B	P0ADB7	DISD-GGNA	34	Type I restriction enzyme EcoKI M protein	P08957	DDLD-GDTQ	200
Uncharacterized protein yffQ	P76548	CFAD-VGDY	19	ATP-dependent Clp protease ATP-binding subunit clpX	P0A6H1	SVID-GQSK	404
Chaperone surA	P0ABZ6	SDVD-GLMQ	43	Elongation factor Tu	P0A6N1	GFLD-SYIP	197
30S ribosomal protein S2	P0A7V0	DGVD-FVIP	197	Seryl-tRNA synthetase	P0A8L1	EEAD-TSNY	228
Fructose-bisphosphate aldolase class 2	P0AB71	DGVD-NSHM	187	Dihydrolipoyl dehydrogenase	P0A9P0	DSTD-ALEL	167
30S ribosomal protein S7	P02359	LMVD-GKKS	33	50S ribosomal protein L9	P0A7R1	DIAD-AVTA	101
50S ribosomal protein L2	P60422	SGVD-AAIK	121	50S ribosomal protein L22	P61175	EHND-GADI	62
Chaperone protein htpG	P0A6Z3	DEVD-ESAK	498	Uncharacterized protein yeiA	P25889	SCYD-GGHQ	356
UPF0325 protein yaeH	P62768	VVAD-GVGQ	60	30S ribosomal protein S5	P0A7W1	ATID-GLEN	142
Protein translocase subunit secA	P10408	DEVD-SILI	212	Osmotically-inducible protein Y	P0AFH8	ETTD-GVVQ	163
Carbamoyl-phosphate synthase small chain	P0A6F1	DNPD-AALA	139	Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex	P0AFG6	HNLD-ASAI	128
Chaperone protein dnaK	P0A6Y8	DVVD-AEFE	628	Asparaginyl-tRNA synthetase	P0A8M0	DYTD-AVTI	315
Chaperone protein dnaK	P0A6Y8	DEVD-GEKT	211	Elongation factor Tu	P0A6N1	AHVD-CPGH	81
Elongation factor Tu	P0A6N1	AQMD-GAIL	100	Flavoprotein wrbA	P0A8G6	SKVD-GAEV	30
Tagatose-1,6-bisphosphate aldolase gatY	P37192	NEAD-ALYT	151	Uncharacterized protein yfeX	P76536	EEID-GDER	206
30S ribosomal protein S18	P0A7T7	QEID-YKDI	22	Phosphoenolpyruvate-protein phosphotransferase	P08839	ITLD-GHQV	265
Dihydrolipoyl dehydrogenase	P0A9P0	DCAD-GMTK	394	Phosphate acetyltransferase	P0A9M8	LMID-GPLQ	622
Sulfite reductase [NADPH] hemoprotein beta-component	P17846	DLND-GLTG	38	60 kDa chaperonin	P0A6F5	DAAD-LGAA	531
Elongation factor Tu	P0A6N1	DQID-NAPE	51	Protein recA	P0A7G6	SQPD-TGEQ	121
Elongation factor Tu	P0A6N1	GHVD-HGKT	22	Enolase	P0A6P9	DESD-WDGF	297
Aerobic respiration control protein arcA	P0A9Q1	STPD-TPEI	217	Bifunctional protein folC	P08192	DHTD-WLGP	175
Uncharacterized protein yfeX	P76536	DGVD-AGGS	165				

Table 3.3 GluC cleavage-sites in *E. coli* proteins identified by N-terminomics. The names and SwissProt IDs of *E. coli* proteins cleaved by GluC are shown with the cleavage-site shown from P4-P4', and the P1 amino acid number.

protein name	*SwissProt ID*	*cut site*	*obs P1*	*protein name*	*SwissProt ID*	*cut site*	*obs P1*
30S ribosomal protein S11	P0A7R9	KNLE-VMVK	83	Galactitol-specific phosphotransferase enzyme IIB component	P37188	EIKE-LCQN	25
30S ribosomal protein S18	P0A7T7	GVQE-IDYK	20	DNA gyrase subunit A	P0AES4	IEEE-LKSS	16
Osmotically-inducible protein Y	P0AFH8	DAKE-GSVK	126	DNA-binding protein H-NS	P0ACF8	ENGE-TKTW	105
Cell division protein ftsZ	P0A9A6	KRPE-ITLV	322	Trigger factor	P0A850	QGLE-AIEV	121
Osmotically-inducible protein Y	P0AFH8	TTNE-SAGQ	37	UPF0082 protein yebC	P0A8A0	IIRE-LVTA	31
GTP-dependent nucleic acid-binding protein engD	P0ABU2	AKAE-LAVL	161	Uncharacterized protein yeaG	P0ACY3	QRYE-AAKD	12
DNA-binding protein stpA	P0ACG1	TWLE-LMKA	63	DNA-binding protein HU-alpha	P0ACF0	SLKE-GDAV	38
Pyridoxine/pyridoxamine 5'-phosphate oxidase	P0AFI7	KFLE-LKQK	162	50S ribosomal protein L15	P02413	IQIE-FAKV	106
DNA-binding protein H-NS	P0ACF8	QYRE-MLIA	63	Alkyl hydroperoxide reductase subunit F	P35340	HQIE-TASG	306
Stationary-phase-induced ribosome-associated protein	P68191	IVTE-GDKS	26	Uncharacterized ABC transporter ATP-binding protein yjjK	P0A9W3	TLGE-TVKL	384
Uncharacterized protein yccJ	P0AB14	AIFE-VAGY	30	50S ribosomal protein L7/L12	P0A7K2	EKTE-FDVI	54
Glycyl-tRNA synthetase beta subunit	P00961	MVFE-FTDT	404	Elongation factor G	P0A6M8	MKRE-FNVE	476
Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex	P0AFG6	PAKE-SAPA	157	30S ribosomal protein S1	P0AG67	SLKE-IETR	15
Protein grpE	P09372	QHEE-IEAV	24	UTP--glucose-1-phosphate uridylyltransferase	P0AEP3	LLDE-VQSI	93
50S ribosomal protein L10	P0A7J3	AKFE-VKAA	107	Trigger factor	P0A850	PIVE-VTDA	131
Autonomous glycyl radical cofactor	P68066	VRVE-GGQH	68	Autonomous glycyl radical cofactor	P68066	AEDE-VVAV	42
30S ribosomal protein S6	P02358	RHYE-IVFM	5	Translation initiation factor IF-2	P0A705	ELEE-AVMS	375
50S ribosomal protein L11	P0A7J7	QLQE-IAQT	108	Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex	P06959	AKAE-GKSE	315
50S ribosomal protein L10	P0A7J3	IVAE-VSEV	14	30S ribosomal protein S5	P0A7W1	PASE-GTGI	101
Trigger factor	P0A850	PEVE-LQGL	116	Galactitol-specific phosphotransferase enzyme IIB component	P37188	AAEE-IKEL	22
10 kDa chaperonin	P0A6F9	KEVE-TKSA	18	30S ribosomal protein S3	P0A7V3	LTKE-LAKA	46
DNA-binding protein HU-alpha	P0ACF0	AITE-SLKE	34	DNA-directed RNA polymerase subunit beta	P0A8V2	EIKE-LLKL	1197
Trigger factor	P0A850	DSIE-TAVK	25	Translation initiation factor IF-2	P0A705	VETE-NGMI	436
DNA-binding protein H-NS	P0ACF8	REEE-SAAA	44	Uncharacterized oxidoreductase yajO	P77735	VSDE-VGKN	234
30S ribosomal protein S1	P0AG67	(N-term)MTE-SFAQ	3	Nucleoside diphosphate kinase	P0A763	ARFE-AAGF	28
30S ribosomal protein S3	P0A7V3	RKPE-LDAK	110	50S ribosomal protein L3	P60438	TKPE-AGHF	64
Tagatose-1,6-bisphosphate aldolase gatY	P37192	QVNE-ADAL	149	Cytidylate kinase	P0A6I0	AMAE-ALQW	27
30S ribosomal protein S5	P0A7W1	AVLE-VAGV	116	UPF0438 protein yifE	P0ADN2	PVTE-EEKL	55
GTP-binding protein lepA	P60785	MDLE-VVPV	125	Tryptophanase	P0A853	VVQE-GFPT	302
Translation initiation factor IF-2	P0A705	RENE-LEEA	372	Formate acetyltransferase 1	P09373	ASIE-GGQH	701
UPF0241 protein yihI	P0A8H6	PQAE-LELL	98	Elongation factor Tu	P0A6N1	KILE-LAGF	191
10 kDa chaperonin	P0A6F9	KRKE-VETK	16	Integration host factor subunit alpha	P0A6X7	TKAE-MSEY	7
2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	P62707	NADE-IAAK	235	30S ribosomal protein S1	P0AG67	AVIE-SENS	175
50S ribosomal protein L10	P0A7J3	EVSE-VAKG	17	Chaperone protein htpG	P0A6Z3	SAKE-AEKA	503
Outer membrane protein C	P06996	YDYE-GFGI	210	ATP-dependent Clp protease ATP-binding subunit clpX	P0A6H1	VDLE-FRDE	349
Pyruvate dehydrogenase [cytochrome]	P07003	VAME-MKAG	469	Aconitate hydratase 2	P36683	LTEE-GYYS	751
30S ribosomal protein S15	P0ADZ4	LSTE-ATAK	6	RNA polymerase sigma factor rpoS	P13445	SAEE-IAEQ	197
30S ribosomal protein S12	P0A7S3	NLQE-HSVI	76	Soluble pyridine nucleotide transhydrogenase	P27306	TIPE-ISSV	355
50S ribosomal protein L9	P0A7R1	RRAE-LEAK	53	Aconitate hydratase 2	P36683	ASAE-LAAV	809
UPF0325 protein yaeH	P62768	RHLE-SVVT	108	UTP--glucose-1-phosphate uridylyltransferase	P0AEP3	KGVE-LAPG	187
Biosynthetic arginine decarboxylase	P21170	SMQE-VAMS	24	ATP synthase subunit alpha	P0ABB0	KVTE-LLKQ	429
DNA-binding protein HU-beta	P0ACF4	TGKE-ITIA	68	DNA-binding protein H-NS	P0ACF8	DPNE-LLNS	74
Uncharacterized protein yibT	Q2M7R5	KLGE-NVPL	6	Phage shock protein E	P23857	THVE-NAGG	89
30S ribosomal protein S3	P0A7V3	NTKE-FADN	28	UPF0337 protein yjbJ	P68206	KVKE-QWGK	19
50S ribosomal protein L9	P0A7R1	AELE-AKLA	55	Phosphoenolpyruvate-protein phosphotransferase ptsP	P37177	AQKE-TAAI	237
50S ribosomal protein L4	P60723	DFNE-ALVH	25	Outer membrane protein A	P0A910	VEIE-VKGI	333
30S ribosomal protein S5	P0A7W1	QAGE-LQEK	10	50S ribosomal protein L9	P0A7R1	KLAE-VLAA	60
Protein sufB	P77522	NYKE-GFFT	23	Cell division protein zapB	P0AF36	LSQE-VQNA	42
50S ribosomal protein L18	P0C018	AIAE-QLKY	60	Uncharacterized ABC transporter ATP-binding protein yjjK	P0A9W3	VWEE-VSGG	407
DNA-binding protein HU-alpha	P0ACF0	TGKE-IKIA	68	Histidine biosynthesis bifunctional protein hisB	P06987	PHLE-YKAE	280

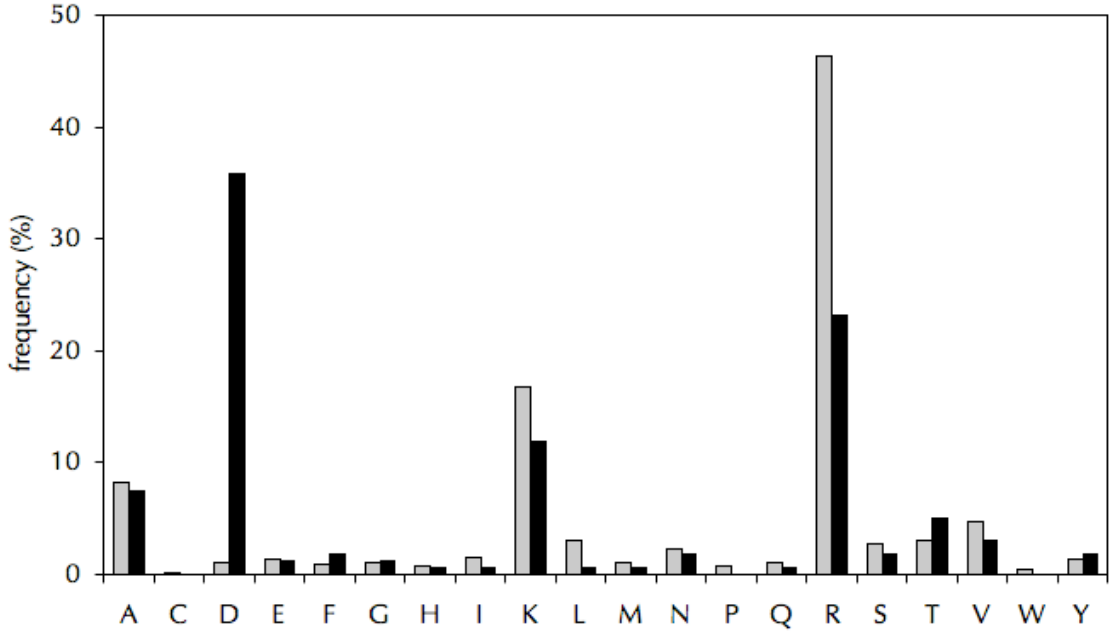


Figure 3.3 Caspase-3 samples have an increased frequency of D cleavage-sites. Protease only cleavage-sites (black bars) reveal the hallmark P1 specificity of human caspase-3. Cleavage-sites found in control samples (grey bars) suggest endogenous proteolysis at Ala, Lys, and Arg residues in the P1 position, and confirm the lack of endogenous Asp specific protease activity in *E. coli*.

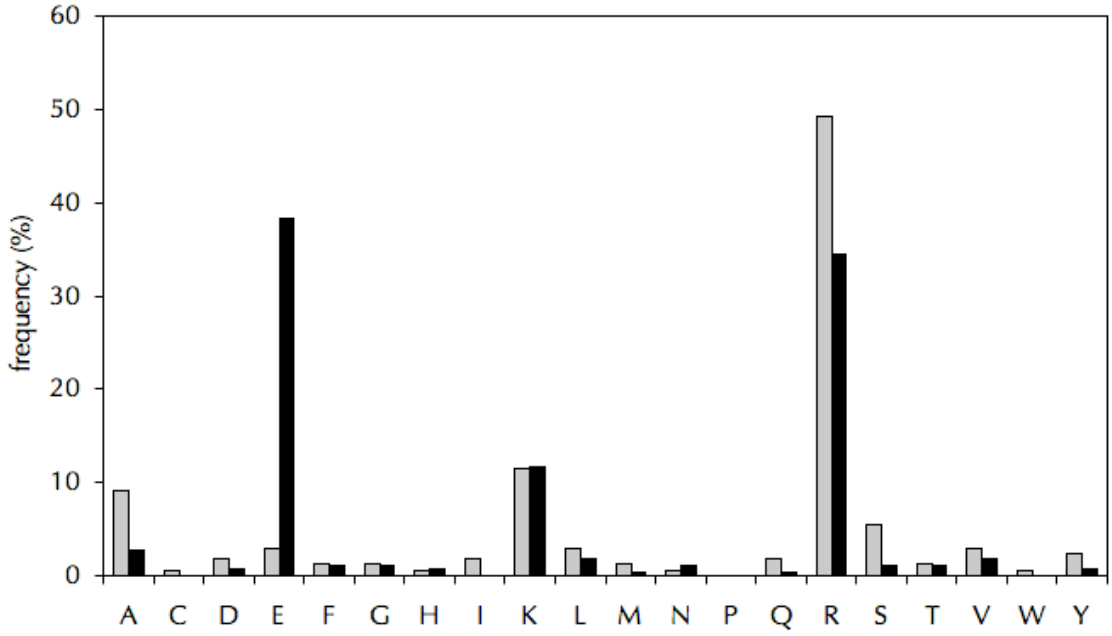


Figure 3.4 GluC samples have an increased frequency of E cleavage-sites. Protease only cleavage-sites (black bars) reveal the hallmark P1 specificity of Staphylococcal GluC. Cleavage-sites found in control samples (grey bars) suggest endogenous proteolysis at Ala, Lys, and Arg residues in the P1 position, and confirm the lack of endogenous Glu specific protease activity in *E. coli*.

of P1 amino acids was normalized to the total number of cleavage-sites in each sample to allow for the unequal numbers of cleavage-sites originating from control and protease treated samples. An anticipated consequence of the normalization is a decrease in the frequency of protease treated P1 amino acids other than the enhanced Asp or Glu residues. This accounts for the noticeable decrease in arginine and lysine from control to protease treated samples, with the implication that most of the constitutive proteolytic cleavages in *E. coli* are at arginine and lysine.

Sub-site amino acid preferences of caspase-3 and GluC. We investigated the specificity of both human caspase-3 and GluC using the WebLogo tool that identifies sequence conservation and amino acid frequency at positions flanking the cleavage-site (http://weblogo.berkeley.edu/logo.cgi) (Crooks, Hon et al. 2004). Search parameters were extended from position P10 to P10' in order to identify potential sub-site preferences not previously assessed by positional scanning peptide libraries. WebLogos were generated from the 57 caspase-3 cleavage-sites and 94 of the 100 GluC cleavage-sites identified by N-terminomics, whose sequences span P10 to P10' (**Fig. 3.5, 3.6**). Unfortunately, 6 of the GluC cleavage-sites were located within 10 residues of substrate N- or C-termini, and thus could not be included in the WebLogo analysis. Surprisingly, our WebLogo analysis of human caspase-3 revealed no preferences in position P3, despite the widely accepted notion that Glu is preferred. The WebLogo results also suggest a minor preference for small and uncharged amino acids in

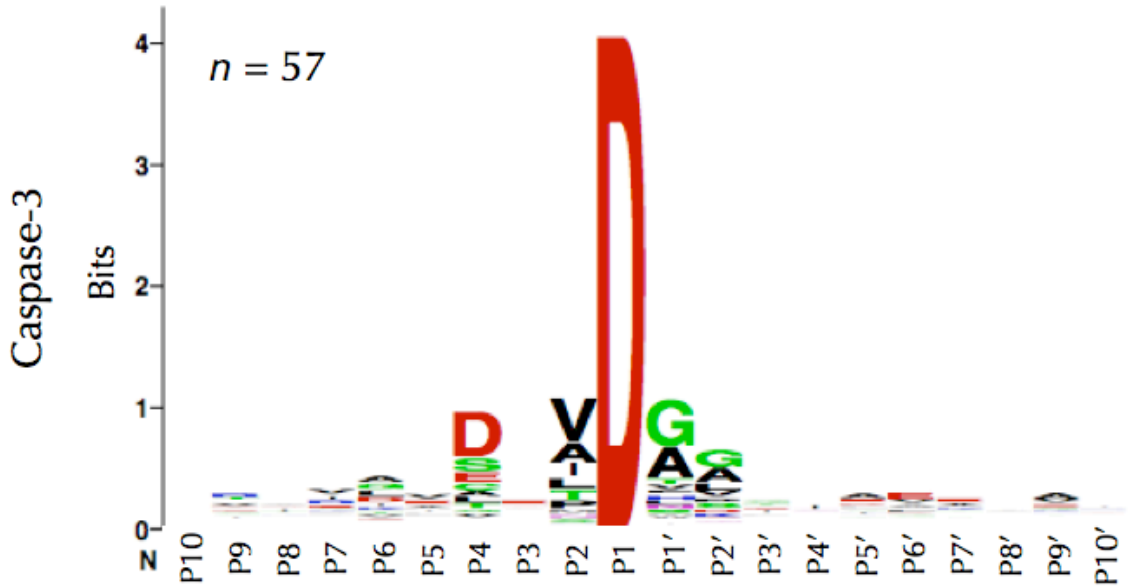


Figure 3.5 Extended specificity of caspase-3. WebLogo representations of protease cleavage-sites depict the amino acid conservation and frequency at each position. The classic human caspase-3 consensus sequence DEVD↓G based on peptide positional scanning libraries is recapitulated with two notable exceptions: there is no conservation in position P3 and a weak additional preference for small uncharged amino acids in position P2'.

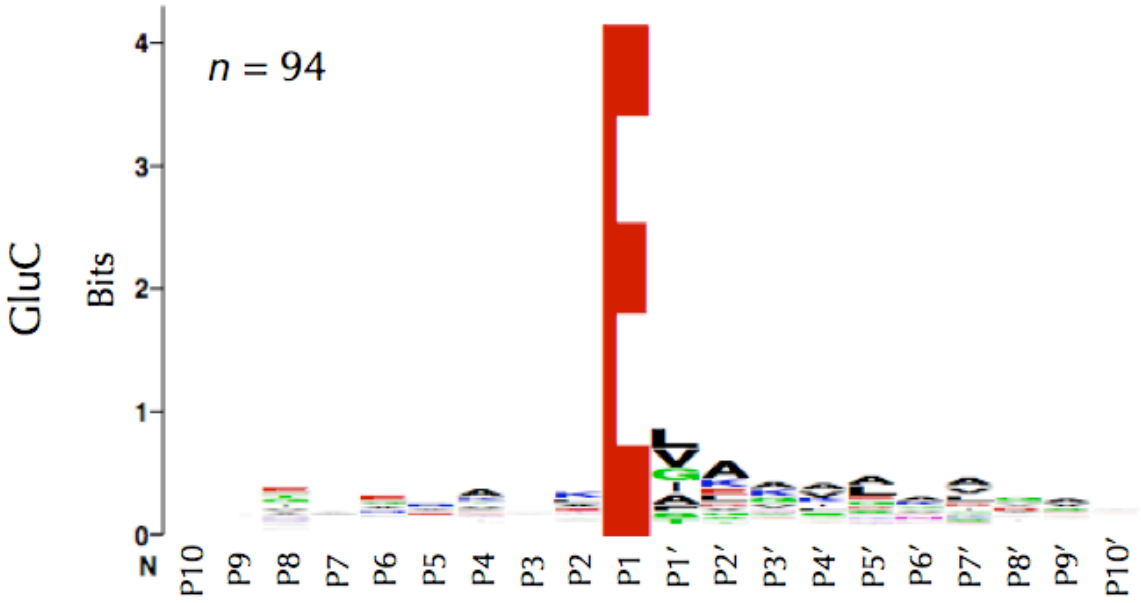


Figure 3.6 Extended specificity of GluC. WebLogo representations of protease cleavage-sites depict the amino acid conservation and frequency at each position. GluC is not known to possess any extended specificity; however, we see a weak preference for hydrophobic residues in the P1' position.

the P2' position, which is similar to the P1' preference. To control for potential amino acid bias surrounding the primary Asp and Glu residues that characterize the protease cleavage specificity, we analyzed 1087 Asp-containing and 1220 Glu-containing sequences from *E. coli* proteins by WebLogo, showing no inherent amino acid bias (**Fig. 3.7, 3.8)**. The lack of specificity for caspase-3 in position P3 could also be accounted for by a proteome-wide depletion of Glu residues in position P3 in relation to Asp sequences. Therefore, we performed a Two-Sample-Logo analysis of these control Asp and Glu sequences from *E. coli* with 1000 sequences containing an equal distribution of each amino acid at every position (http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi) (**Fig. 3.9, 3.10**) (Vacic, Iakoucheva et al. 2006). Indeed, Glu in P3 was not depleted in the *E. coli* proteome in relation to Asp in P1. However, on the whole these results support earlier specificity data for human caspase-3 and GluC based on substrates with no structure to them, although the contribution of the P3 sub-site in caspase-3 is inconsequential in the context of our structured substrate dataset. This means that the cleavage-site recognition in the immediate vicinity of the scissile bond (P4-P2' in the case of caspase-3) is largely dependent on the properties of the protease, not the substrate.

Structures preferred by human caspase-3 and GluC. Many of the *E. coli* proteins we identified as protease substrates had solved structures due to the efforts of the Joint Centers for Structural Genomics and individual researchers depositing protein crystal and NMR structures into the PDB.

Figure 3.7 WebLogo of control D containing sequences. WebLogo analysis of 1087 D containing sequences in 50 random *E. coli* proteins shows no amino acid preferences upstream or downstream of Asp. The extended specificities observed for caspase-3 are not due to the amino acid composition of the *E. coli* proteome.

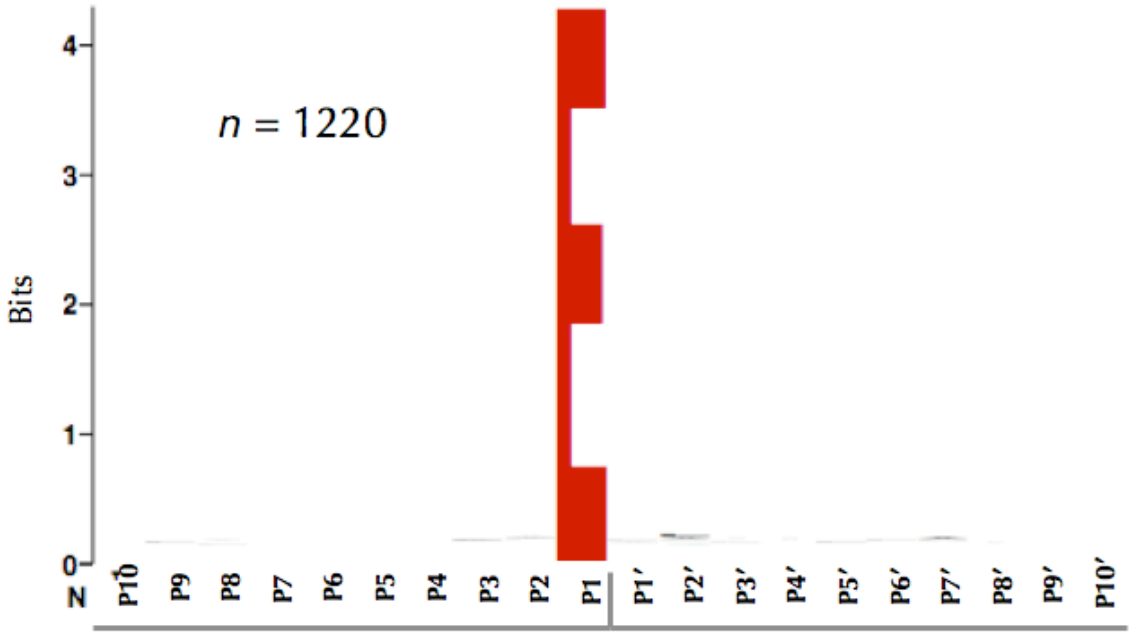


Figure 3.8 WebLogo of control E containing sequences. WebLogo analysis of 1220 E containing sequences in 50 random *E. coli* proteins shows no amino acid preferences upstream or downstream of Glu. The weak hydrophobic P1' specificity observed for GluC is independent of the amino acid composition of the *E. coli* proteome.

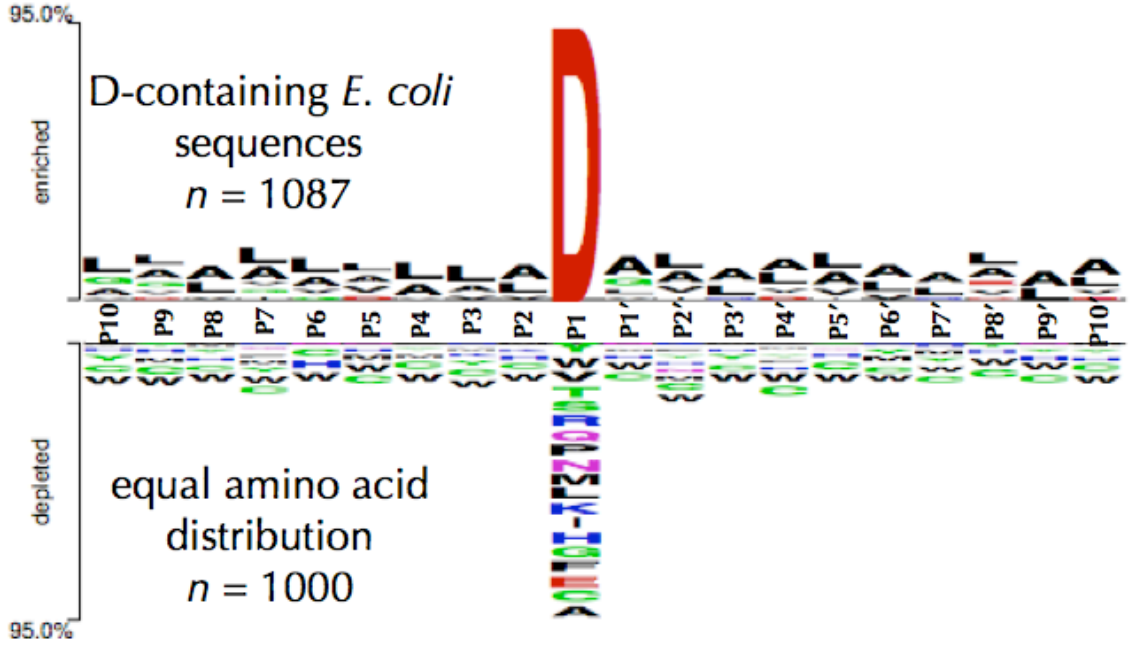


Figure 3.9 Two-Sample-Logo of amino acids in D containing sequences. We performed a Two-Sample-Logo analysis of the same 1087 control D sequences from *E. coli* by comparing them to an *in silico* generated dataset of 1000 sequences containing an equal amino acid distribution. This analysis revealed that the lack of specificity in position P3 for caspase-3 is not due to depletion in the frequency of Glu residues at that position in the *E. coli* proteome.

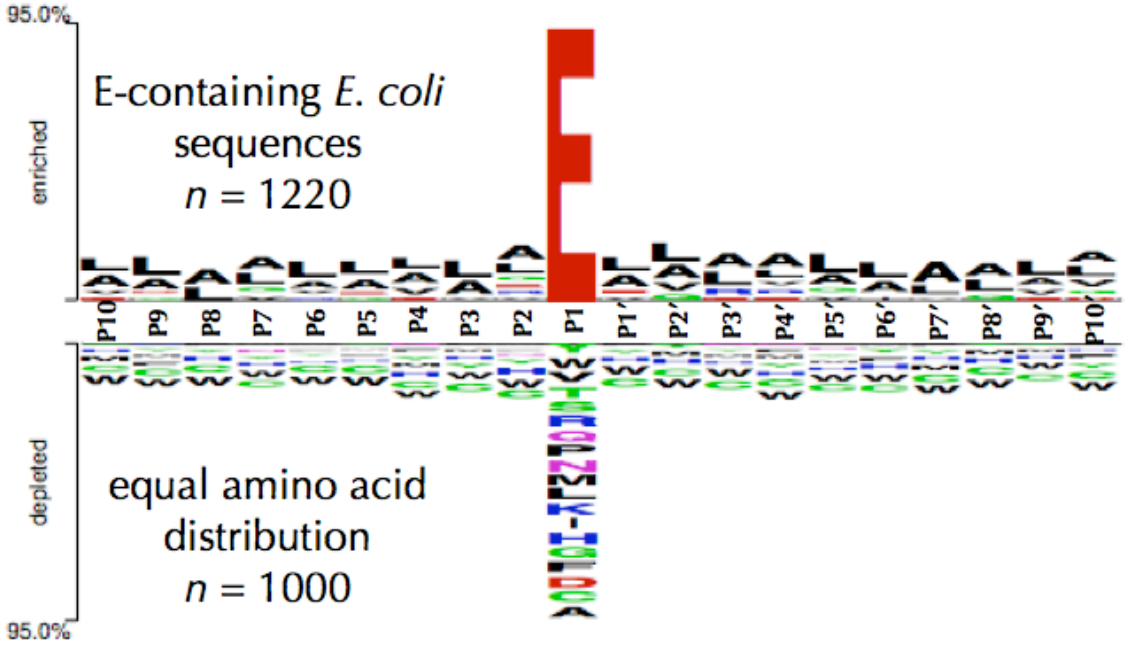


Figure 3.10 Two-Sample-Logo of amino acids in E containing sequences. We performed a Two-Sample-Logo analysis of the same 1220 control E sequences from *E. coli* by comparing them to an *in silico* generated dataset of 1000 sequences containing an equal amino acid distribution, revealing that the weak hydrophobic P1' specificity observed for GluC is independent of the amino acid composition of the *E. coli* proteome.

Mapping the cleavage-sites that we identified back to reported structures revealed the three-dimensional conformations that encompassed sites of proteolysis (**Fig. 3.11** thru **3.28**). Cleavage-sites were manually assessed and visualized using PyMOL software (DeLano 2002), and secondary structure assignments were retrieved from the Dictionary of Protein Secondary Structure (DSSP) (Kabsch and Sander 1983). Secondary structures were also predicted using the PSIPRED algorithm (http://128.16.10.201/psipred/psiform.html) (Jones 1999), allowing us to also assess cleavage-site conformations of structurally unresolved substrates. As expected, WebLogo analysis of cleavage-site secondary structures revealed frequent cleavage in extended loop structures or regions of no electron density; however, there were numerous cleavage-sites residing in α-helices for both proteases tested (**Fig. 3.11** thru **Fig. 3.14**) contradicting a central dogma of proteolysis. Although the majority of cleavage-sites are from crystal-derived structures, we also found cleavage-sites in solution structures solved by NMR lessening the possibility of helix formation induced by crystal growth. Thus the notion that cleavage-sites can reside in α-helices is strongly supported by the prevalent examples in both structural methods. Scattered reports of this phenomenon exist in the literature; however, the magnitude and relative prevalence compared to loop conformations has never been directly addressed in a structurally unbiased system like ours (Mahrus, Trinidad et al. 2008). As expected, extended β-strands were almost never tolerated in cleavage-sites from human caspase-3 or GluC. An analysis of control Asp and Glu site secondary structures (**Fig. 3.29, 3.30**) showed a tendency

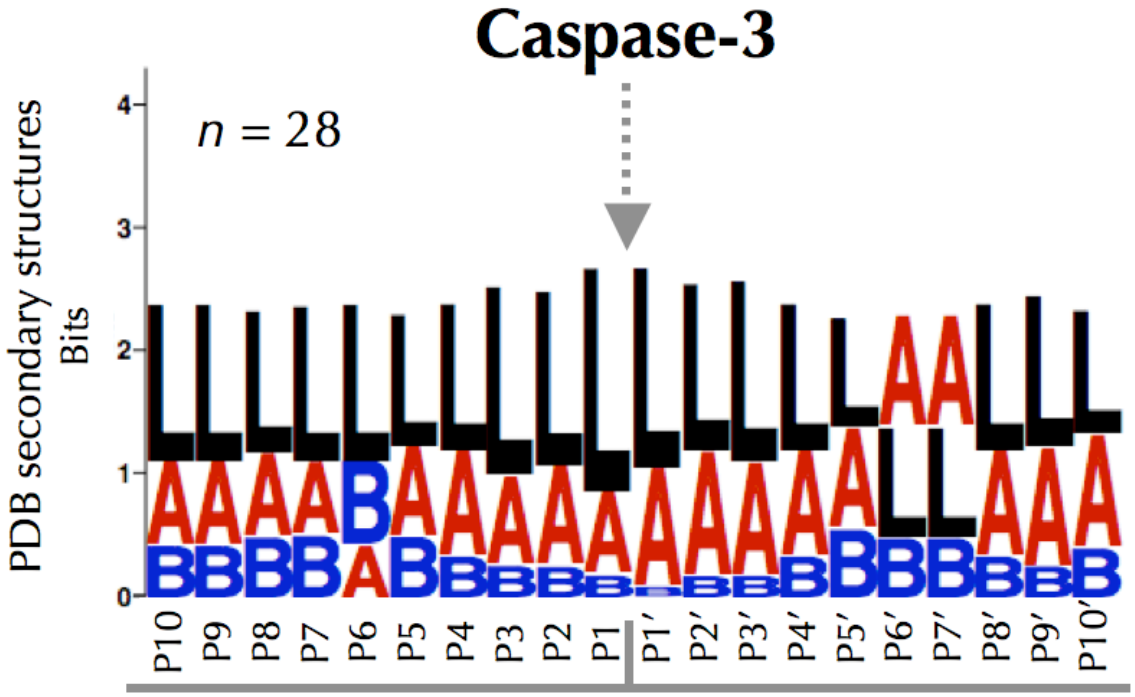


Figure 3.11 Structural preferences of human caspase-3 from substrates with solved structures. WebLogo representations of secondary structures from protease cleavage-sites with the scissile bond indicated by the grey arrow. Secondary structure assignments were determined from protein structures residing in the PDB as defined by DSSP. Human caspase-3 cleaved substrates in loops as well as α-helices, but almost never in β-strands. The intolerance of human caspase-3 for β-strands appears to be restricted around the scissile bond. Secondary structure assignments are as follows: L = loop, A = α-helix, B = β-strand.

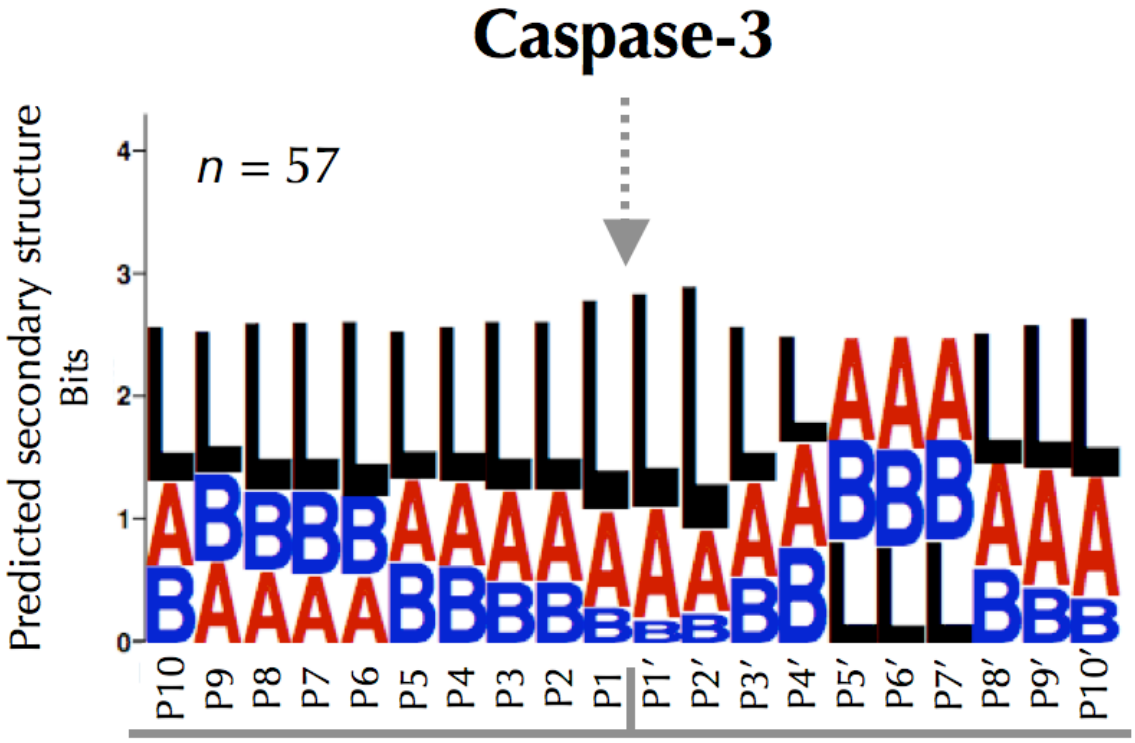


Figure 3.12 Structural preferences of human caspase-3 from predicted structures. WebLogo representations of secondary structures from protease cleavage-sites with the scissile bond indicated by the grey arrow. Secondary structure assignments were predicted by the PSIPRED algorithm. Human caspase-3 cleaved substrates in loops as well as α-helices, but almost never in β-strands. The intolerance of human caspase-3 for β-strands appears to be restricted around the scissile bond. Secondary structure assignments are as follows: L = loop, A = α-helix, B = β-strand.

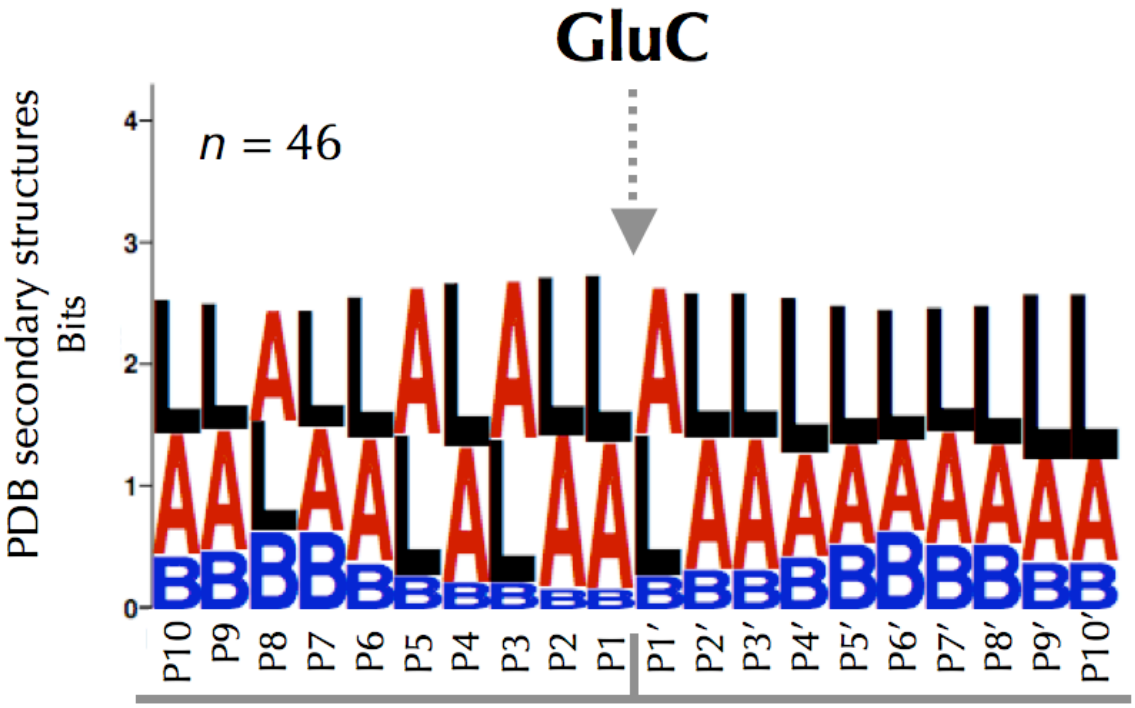


Figure 3.13 Structural preferences of Staphylococcal GluC from substrates with solved structures. WebLogo representations of secondary structures from protease cleavage-sites with the scissile bond indicated by the grey arrow. Secondary structure assignments were determined from protein structures residing in the PDB as defined by DSSP. GluC cleaved substrates in loops as well as α-helices, but almost never in β-strands. The intolerance of GluC for β-strands appears to be restricted around the scissile bond, but is shifted toward the substrates N-terminus. Secondary structure assignments are as follows: L = loop, A = α-helix, B = β-strand.

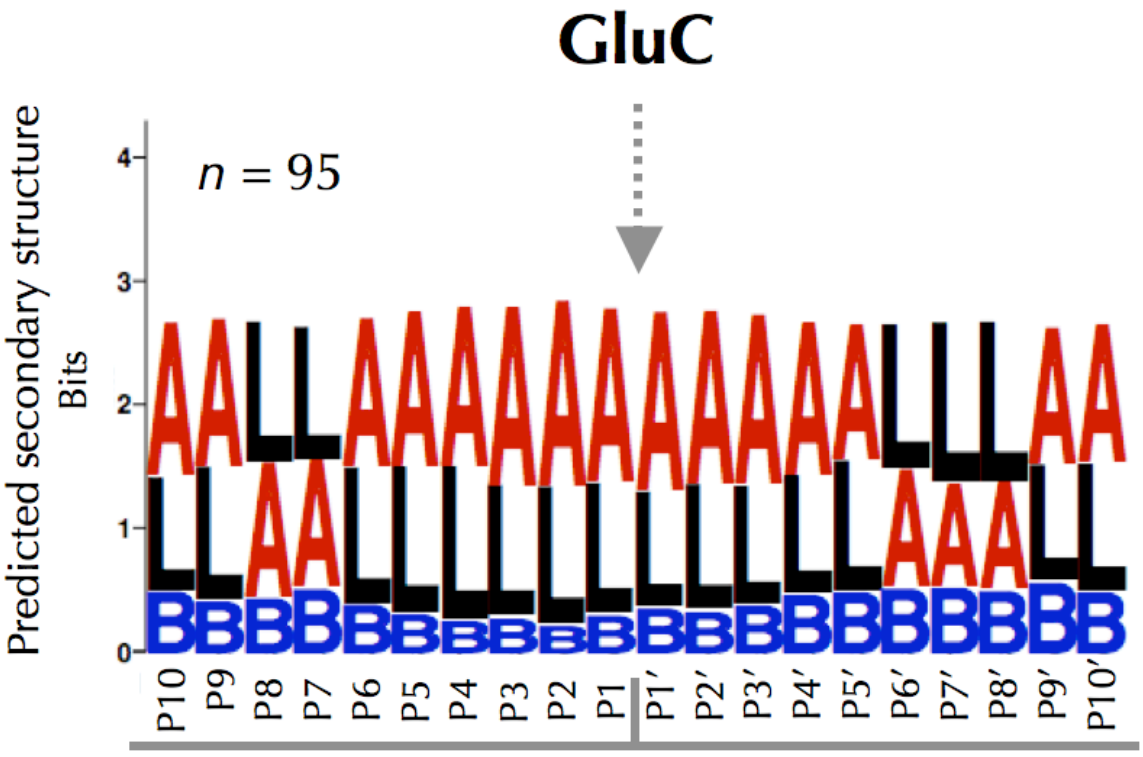


Figure 3.14 Structural preferences of Staphylococcal GluC from predicted structures. WebLogo representations of secondary structures from protease cleavage-sites with the scissile bond indicated by the grey arrow. Secondary structure assignments were predicted by the PSIPRED algorithm. GluC cleaved substrates in loops as well as α-helices, but almost never in β-strands. The intolerance of GluC for β-strands appears to be restricted around the scissile bond, but is shifted toward the substrates N-terminus. Secondary structure assignments are as follows: L = loop, A = α-helix, B = β-strand.

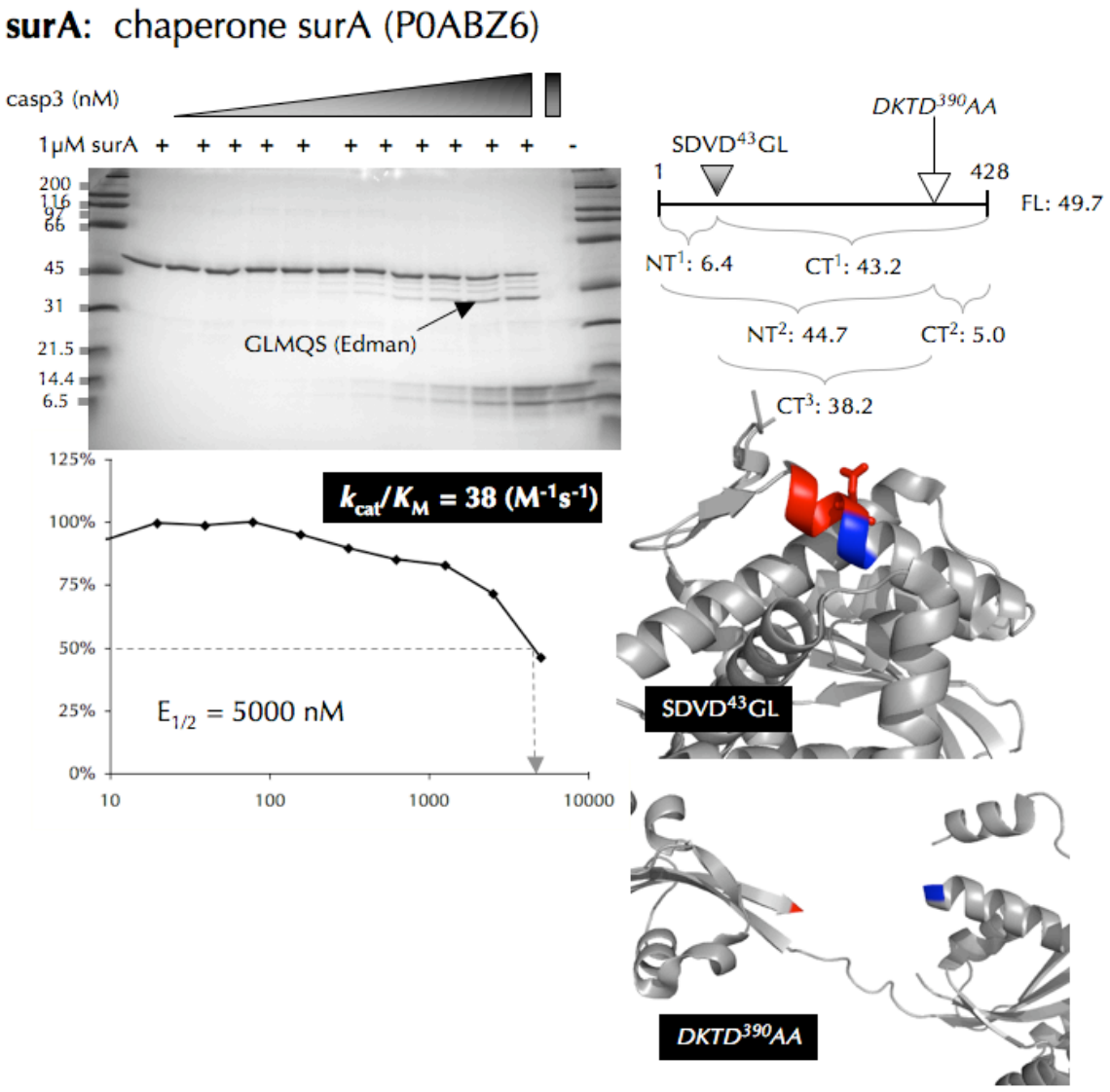


Figure 3.15 Biochemical validation and kinetic analysis of surA. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

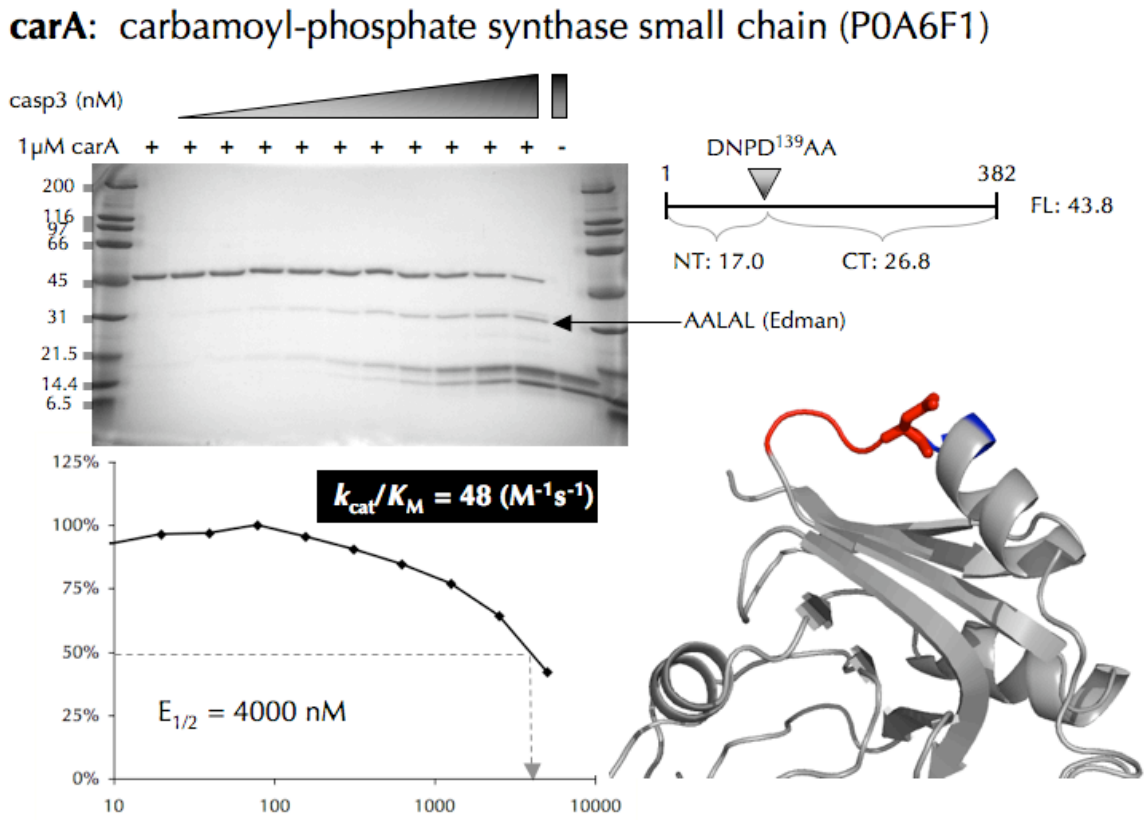


Figure 3.16 Biochemical validation and kinetic analysis of carA. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

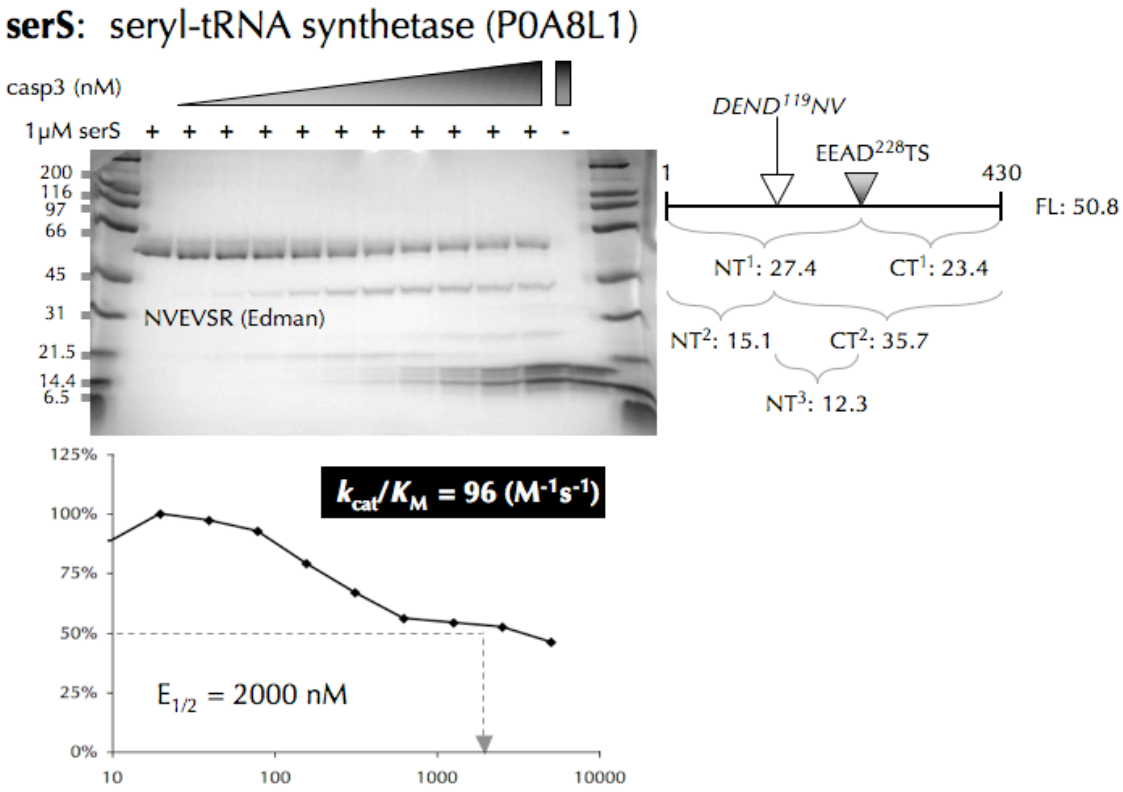


Figure 3.17 Biochemical validation and kinetic analysis of serS. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

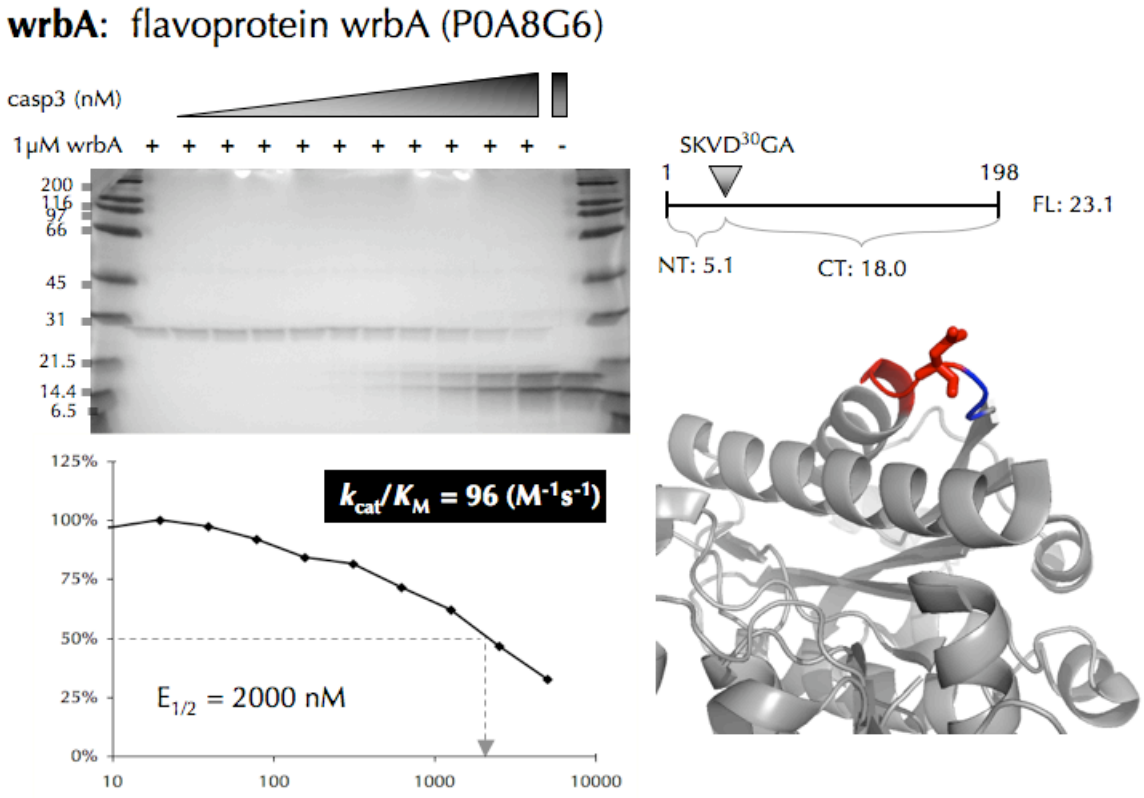


Figure 3.18 Biochemical validation and kinetic analysis of wrbA. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

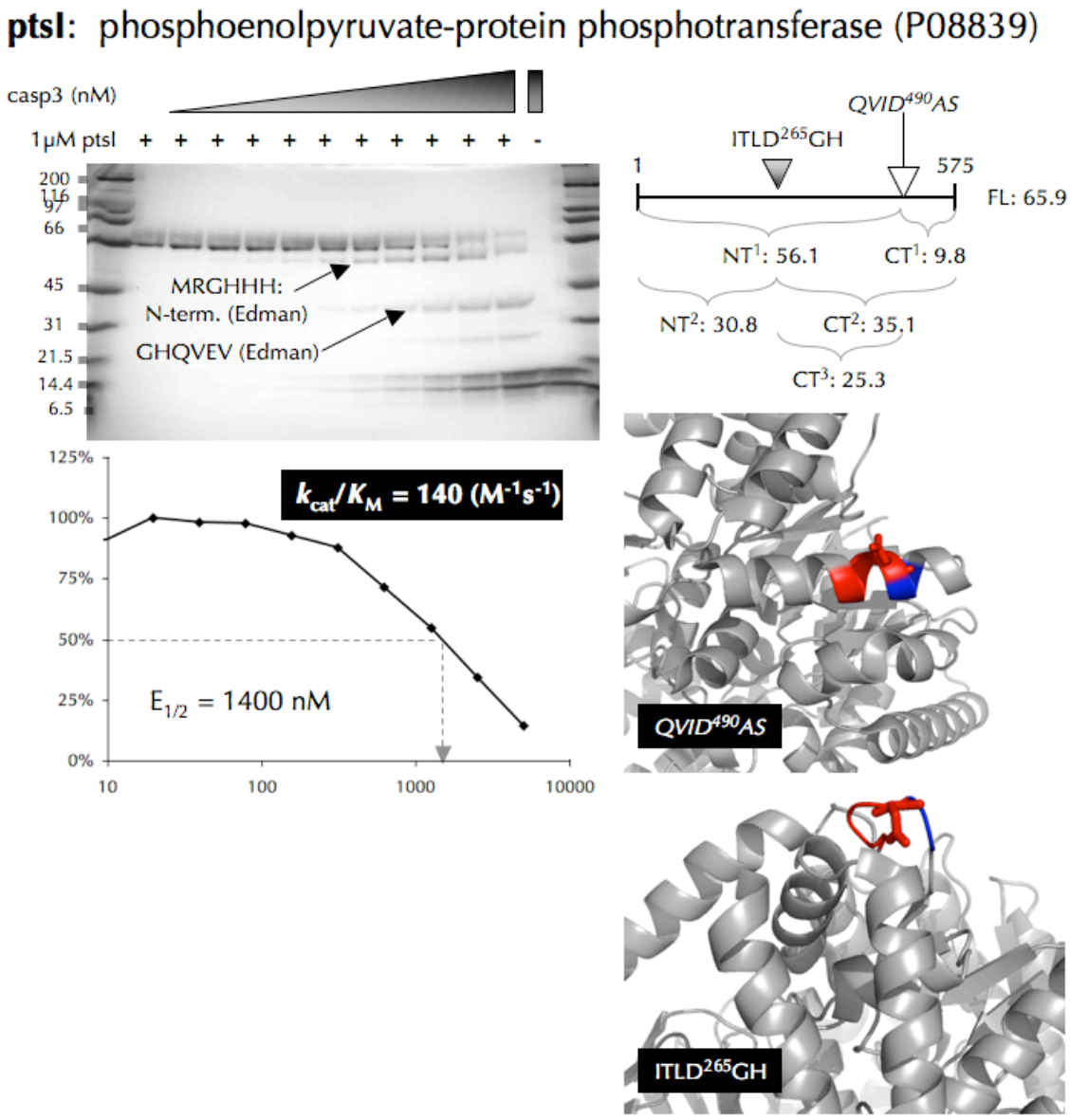


Figure 3.19 Biochemical validation and kinetic analysis of ptsI. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

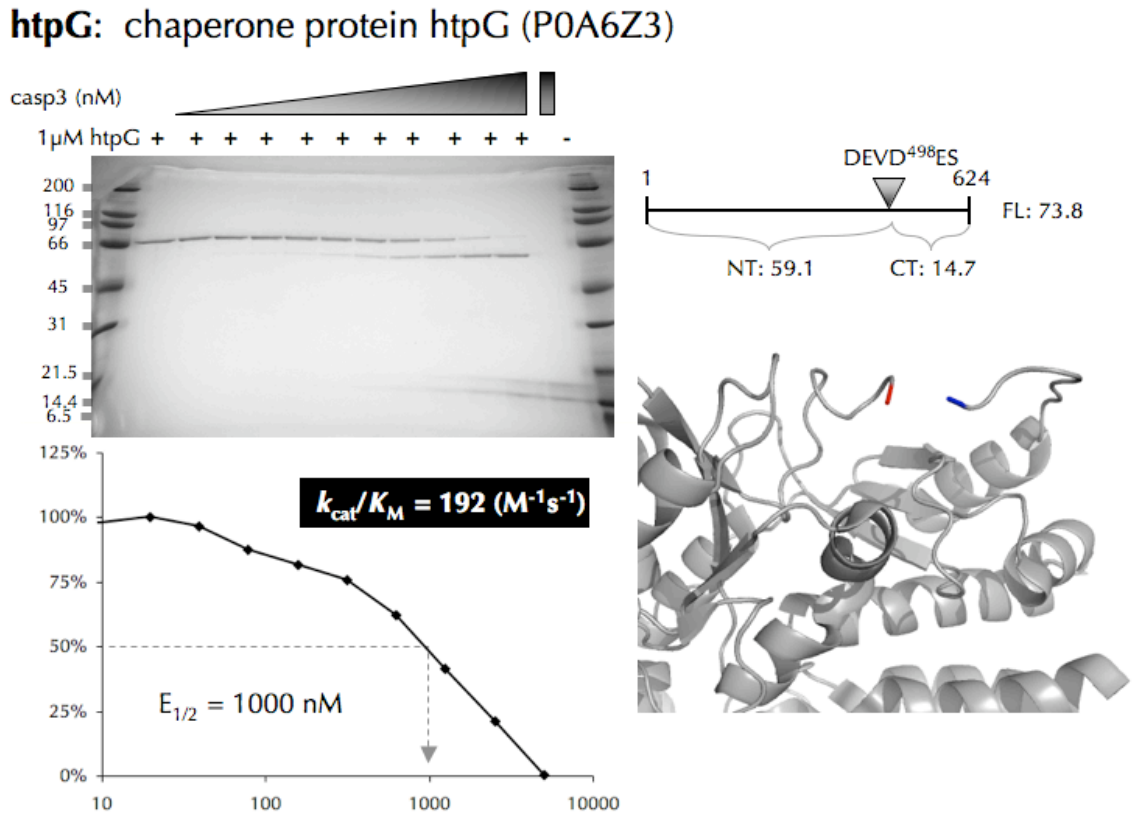


Figure 3.20 Biochemical validation and kinetic analysis of htpG. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

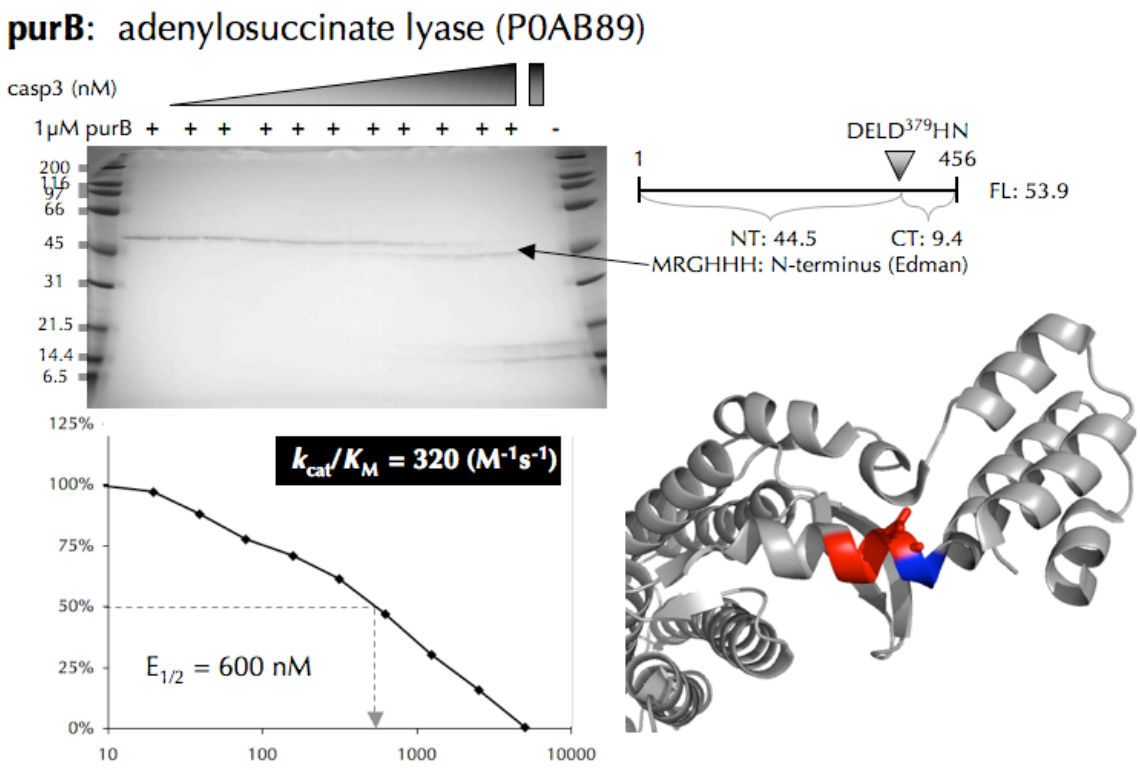


Figure 3.21 Biochemical validation and kinetic analysis of purB. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

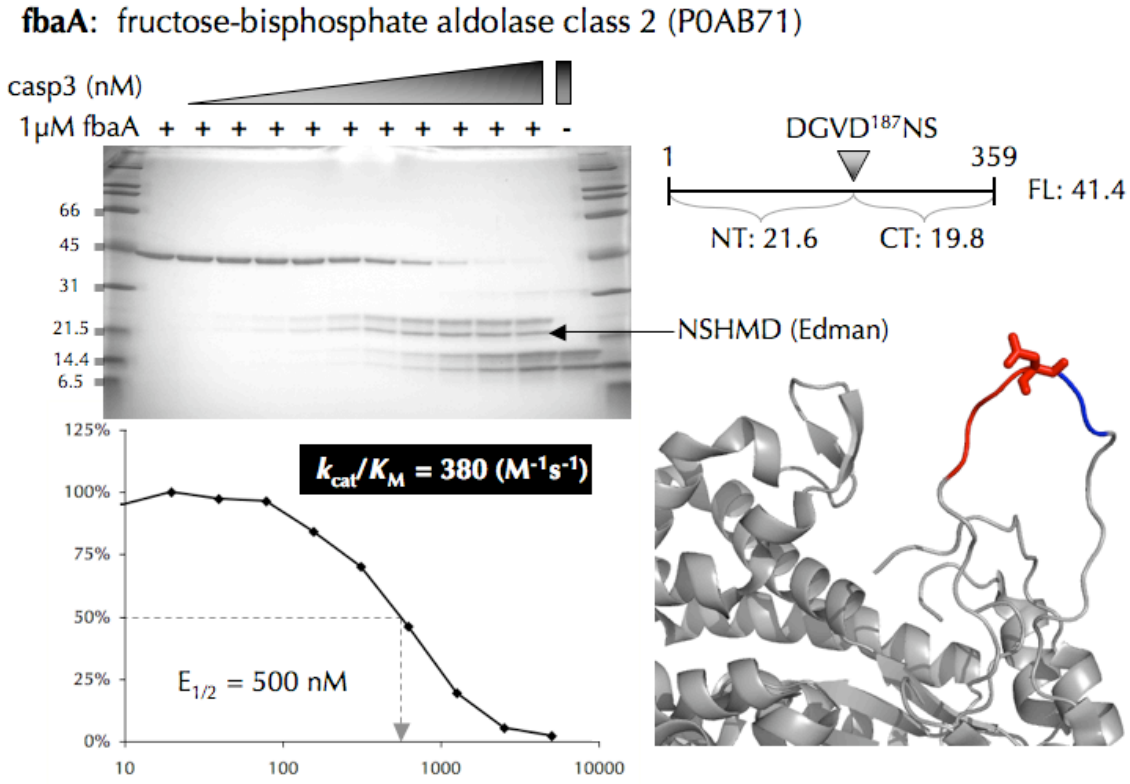


Figure 3.22 Biochemical validation and kinetic analysis of fbaA. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

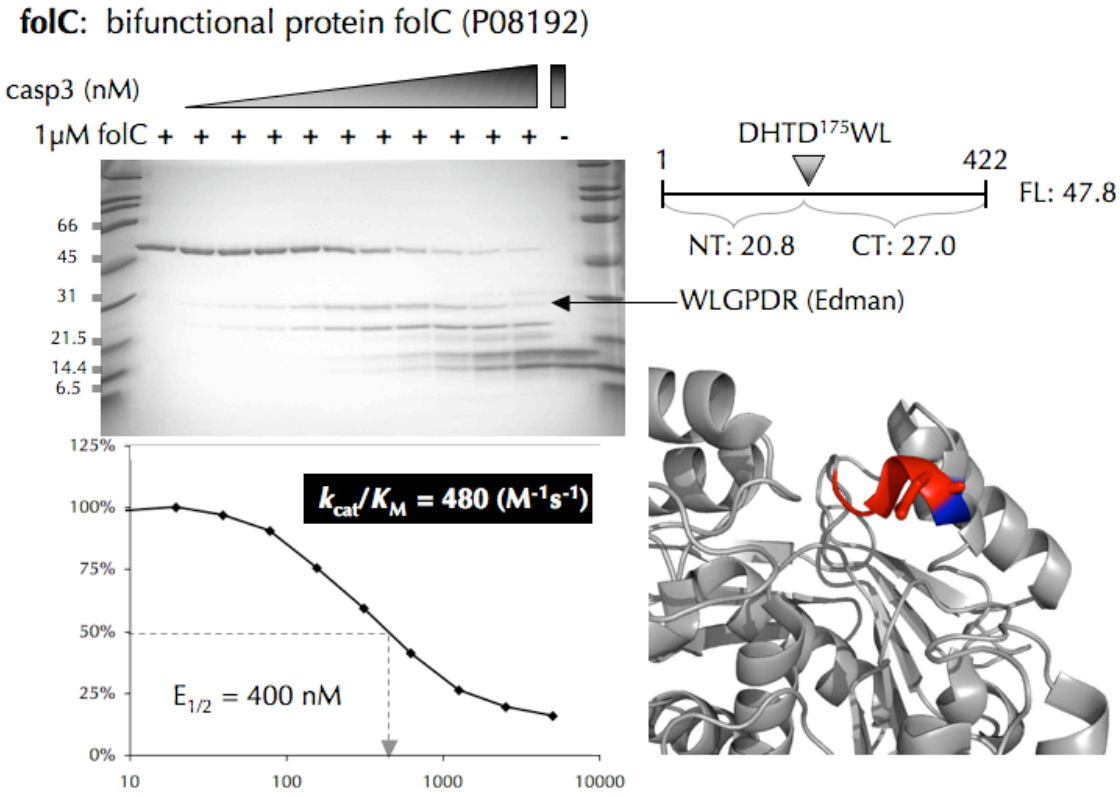


Figure 3.23 Biochemical validation and kinetic analysis of folC. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

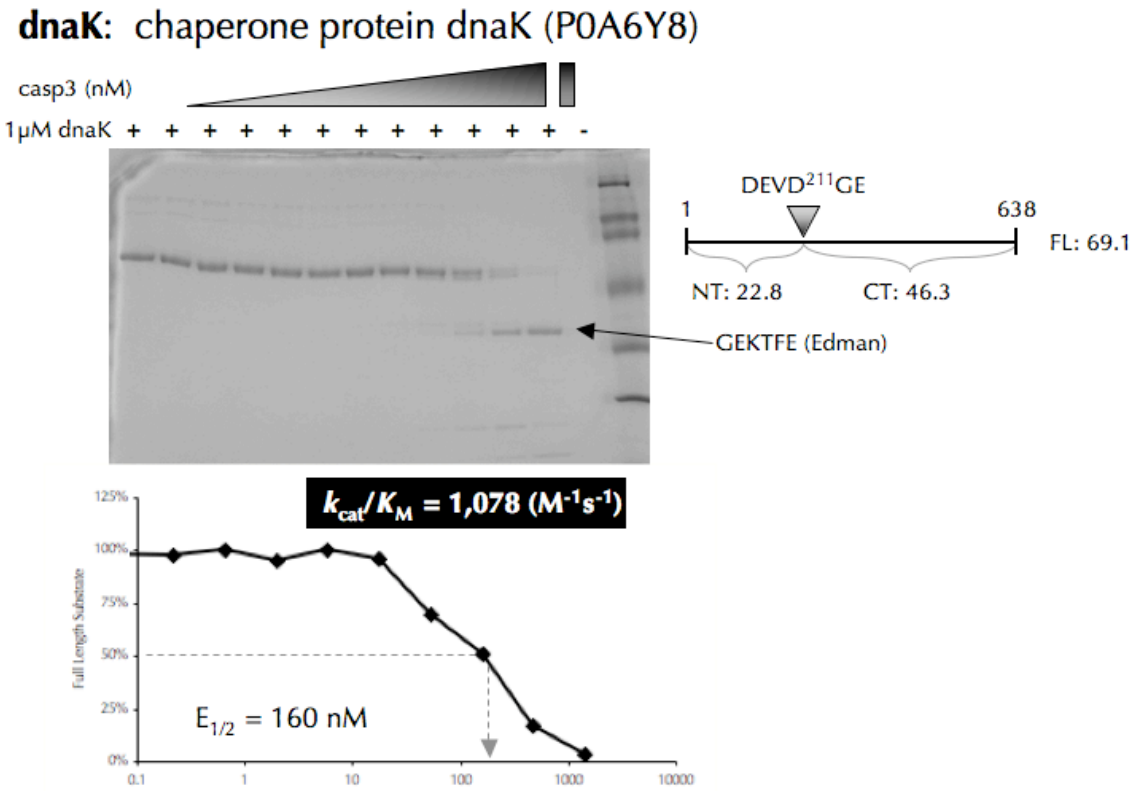


Figure 3.24 Biochemical validation and kinetic analysis of dnaK. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

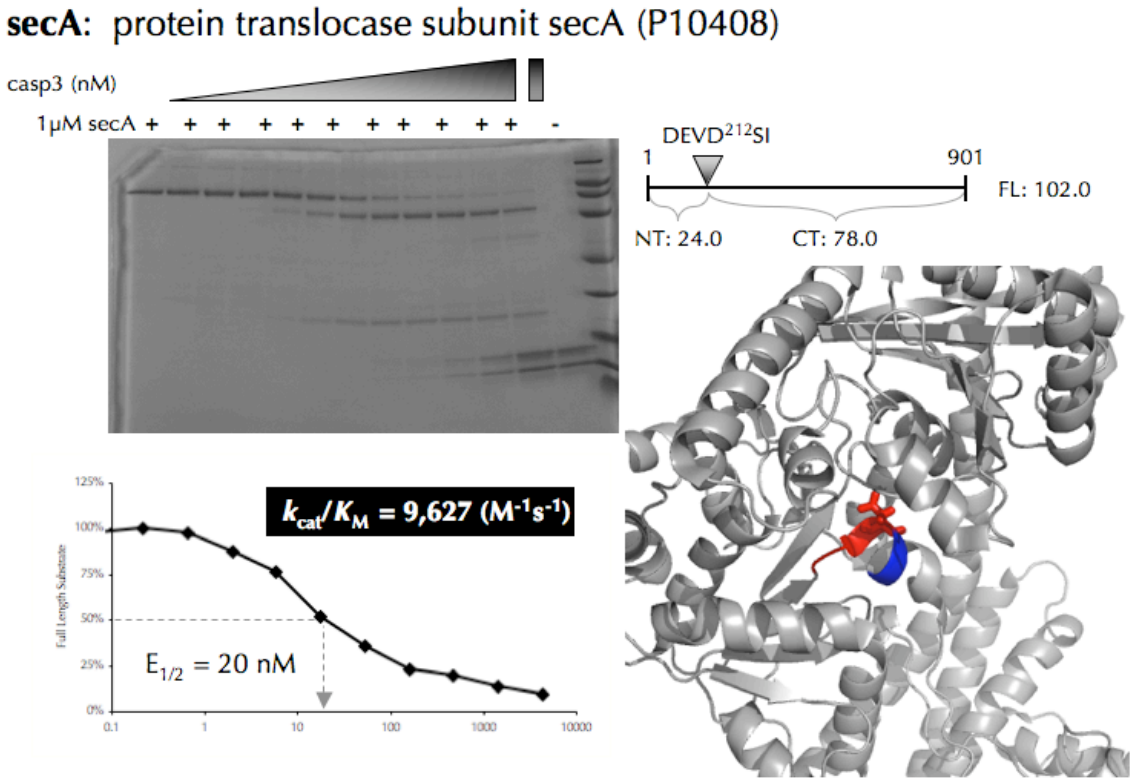


Figure 3.25 Biochemical validation and kinetic analysis of secA. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

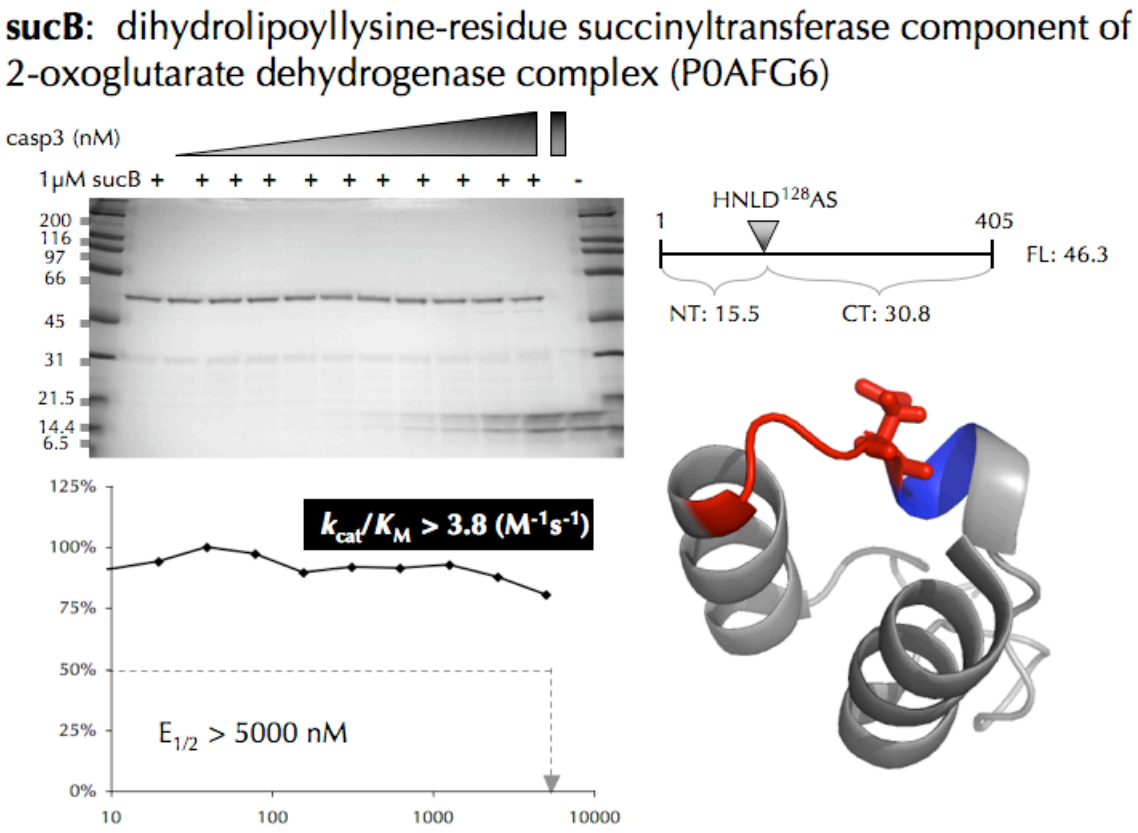


Figure 3.26 Biochemical validation and kinetic analysis of sucB. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

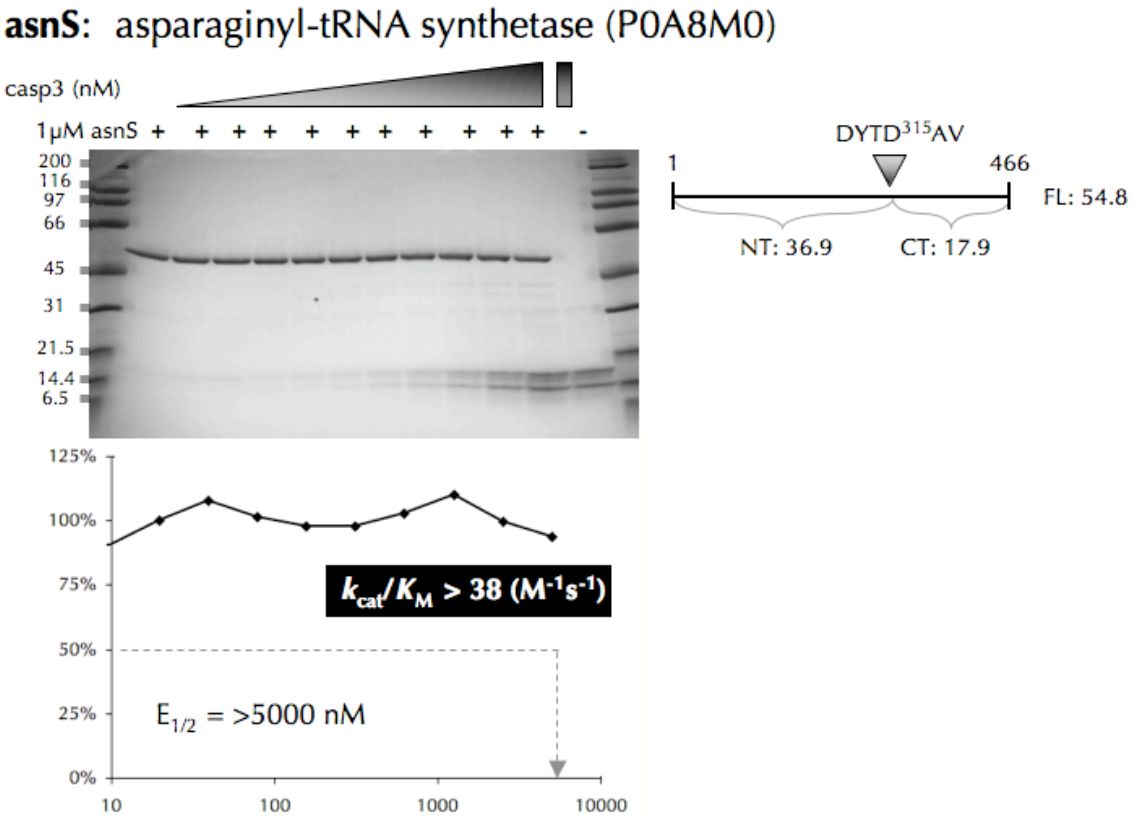


Figure 3.27 Biochemical validation and kinetic analysis of asnS. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

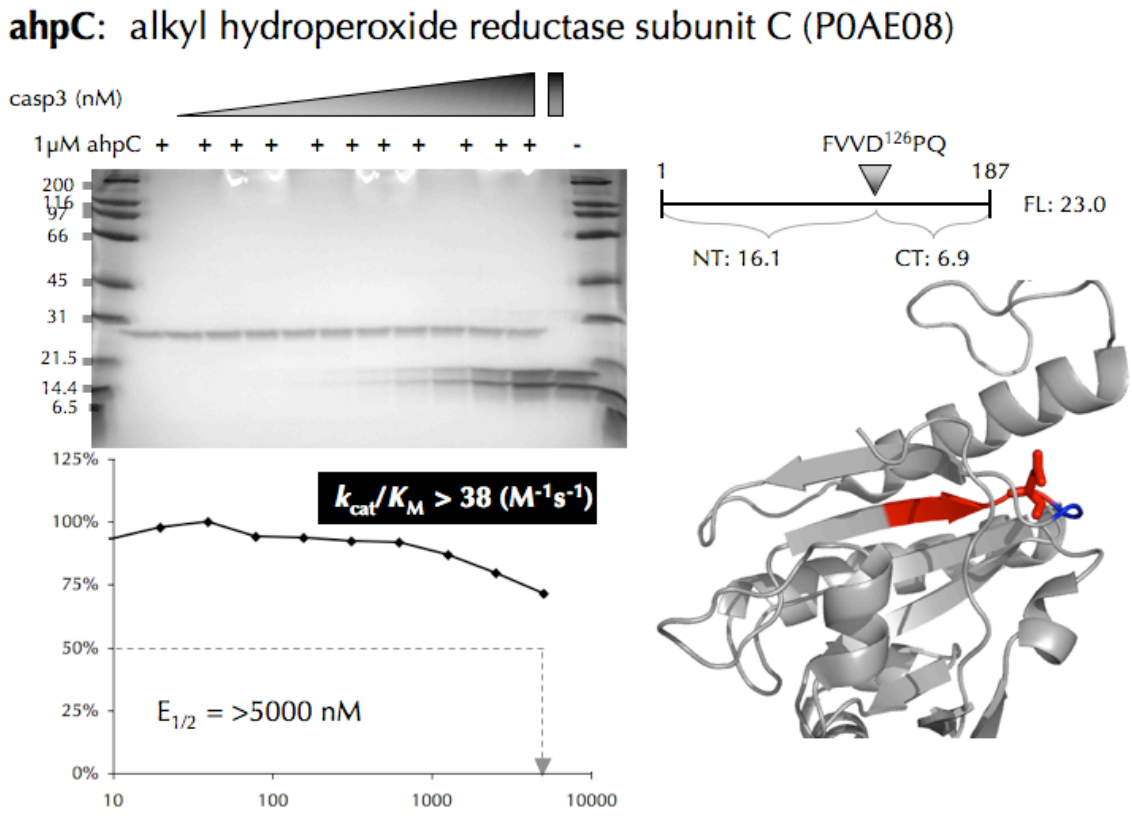


Figure 3.28 Biochemical validation and kinetic analysis of ahpC. Representative *E. coli* proteins that were identified as substrates of caspase-3 were recombinantly expressed, purified, and subjected to *in vitro* cleavage. The $E_{1/2}$ values were measured based on the disappearance of the full-length substrate using densitometry. These values were used to calculate k_{cat}/K_M for each substrate. Substrates with solved structures were retrieved from the Protein Data Bank and visualized with PyMOL. Structures are shown with cleavage-sites colored red from P4 to P1, and blue from P1' to P2', with the P1 residue visualized in stick format.

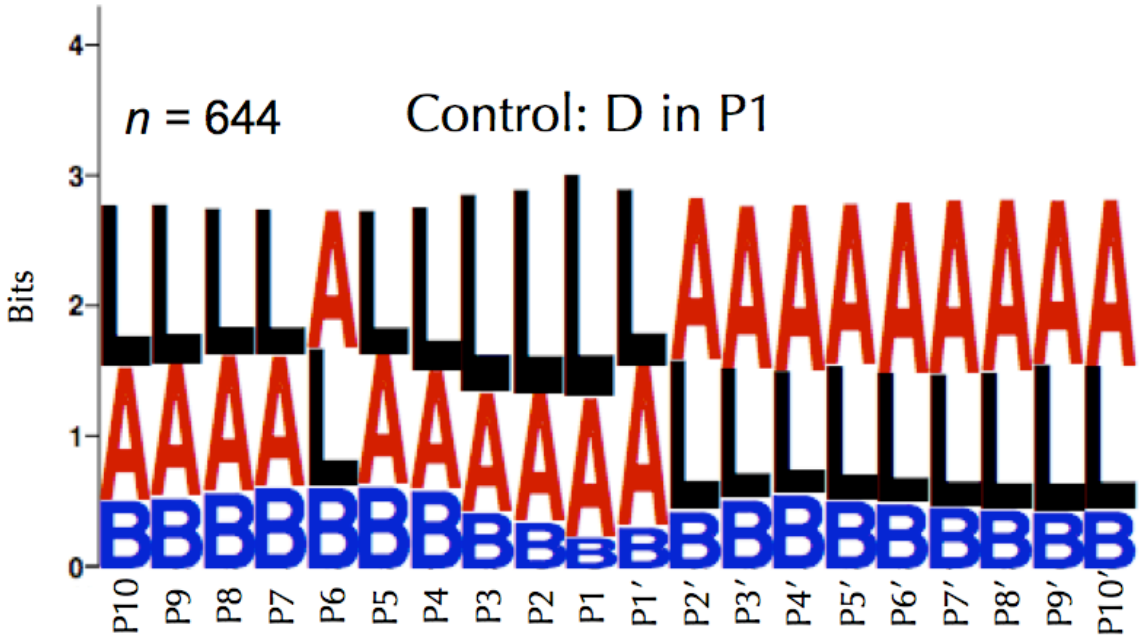


Figure 3.29 Secondary structure WebLogo analysis of Asp fixed in the P1 position from random *E. coli* proteins with solved structures. As a control for the secondary structure analysis of human caspase-3 sites, we randomly selected *E. coli* proteins with PDB files, and then arbitrarily chose amino acid sequences within those proteins, which contained an Asp at position P1. The corresponding secondary structures were retrieved from the DSSP database for all sites in continuous amino acid stretches. The resultant 644 control Asp sites represent secondary structural tendencies for sites with Asp in position P1. Secondary structure assignments are as follows: L = loop, A = α-helix, B = β-strand.

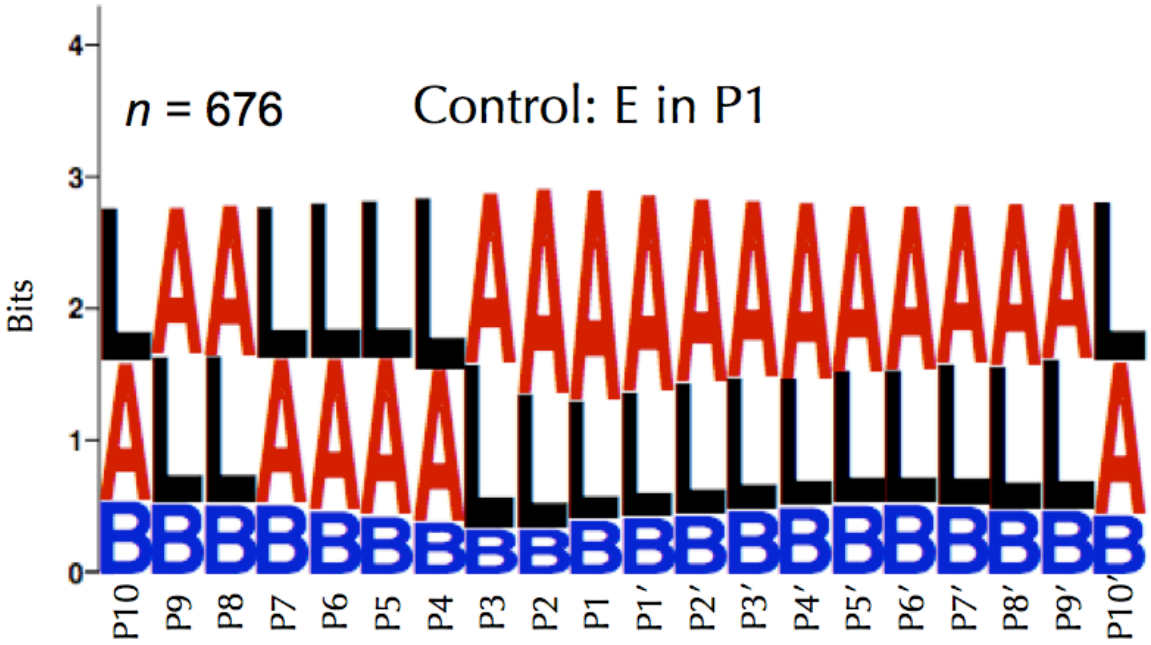


Figure 3.30 Secondary structure WebLogo analysis of Glu fixed in the P1 position from random *E. coli* proteins with solved structures. As a control for the secondary structure analysis of GluC sites, we randomly selected *E. coli* proteins with PDB files, and then arbitrarily chose amino acid sequences within those proteins, which contained a Glu at position P1. The corresponding secondary structures were retrieved from the DSSP database for all sites in continuous amino acid stretches. The resultant 676 control Glu sites represent secondary structural tendencies for sites with Glu in position P1. Secondary structure assignments are as follows: L = loop, A = α-helix, B = β-strand.

toward loop and helical structures; however, this does not conflict with our finding that both caspase-3 and GluC cleaved *E. coli* substrates in α-helices as well as loops.

Biochemical verification of cleavage-sites identified by N-terminomics. To directly validate the cleavage-sites revealed by N-terminomics and kinetically characterize them, we selected several representative substrates cleaved at low, intermediate, and high human caspase-3 concentration for biochemical characterization. Substrates were selected that contained only 1 cleavage-site and had previously been shown to express well from a dataset of all cloned *E. coli* genes (Kitagawa, Ara et al. 2005). We were able to rapidly evaluate these 14 proteins due to the availability of the Genobase ASKA collection of *E. coli* open reading frames cloned into N-terminal 6 histidine-tagged expression vectors. A substantial advantage of using this tagged protein collection is that the proteins are expressed in a homologous system and are purified under similar conditions used to produce the original *E. coli* lysate, enhancing the likelihood that the expressed proteins are conformationally similar to those in the original lysate. Potential substrates of human caspase-3 were expressed, purified, and

evaluated in an *in vitro* cleavage assay. Rates of cleavage (k_{cat}/K_M) were measured by treating 1 μM substrate with a dilution series of protease for 1 hour at 37ºC, followed by SDS-PAGE detection of cleavage products. The enzyme concentrations corresponding to 50% cleavage of the full-length substrates were

determined by densitometry. K_M values for these substrates were determined by competition with a fluorogenic substrate (**Table. 3.4)** and were all above 10 μM, so substrates in our assay are at least 10 fold below the K_M. This allowed us to determine their respective k_{cat}/K_M values from the following half-life equation:

$$\frac{k_{cat}}{K_M} = \frac{\ln 2}{tE_{1/2}}$$

$E_{1/2}$ is the concentration of protease that produces 50% substrate cleavage in time = *t*. The half-life equation used to determine k_{cat}/K_M values is based on a single cleavage-site; therefore substrates with two or more sites are overestimates of each cleavage-site k_{cat}/K_M value. We also verified that the cleavage-products observed in our cleavage assay matched the expected sizes based on the cleavage-sites identified in the N-terminomics screen. In addition, we sequenced many of the C-terminal cleavage products by Edman degradation to confirm the exact cleavage-sites. Examples of the validation and kinetic analysis are shown in **Figure 3.15** thru **3.28**. Several of the substrates assayed revealed multiple cleavage-sites, despite our selection of substrates that were identified from a single N-terminomics hit. These additional cleavage-sites were identified through Edman degradation and for the most part would produce tryptic peptides too short to be unambiguously identified by MS. We tested 14 recombinant *E. coli* proteins in our cleavage assay, and only 3 (21%) could not be confirmed as human caspase-3 substrates either because the cleavage-

Table 3.4 Kinetic parameters of caspase-3 substrates. K_M values were measured for caspase-3 substrates by competition with the fluorogenic peptide substrate Ac-DEVD-AFC. k_{cat}/K_M values were determined by cleavage-assay as described in the text, and k_{cat} values were estimated by

	caspase-3 substrate	cleavage-site P4-P1'	K_M (μM)	k_{cat}/K_M ($M^{-1}s^{-1}$)	k_{cat} (s^{-1})
	Ac-DEVD-AFC	DEVD↓Φ	21.1	551,342	11.63
Human	caspase-9 (C285A)	DQLD↓A	29.5	17,503	0.52
Human	caspase-7 (C285A)	DSVD↓A	42.1	12,836	0.54
Human	ICAD (D244E)	DETD↓S	13	64,180	0.83
Human	ICAD (D117E)	DAVD↓T	34.9	38,508	1.34
E. coli	carA	DNPD↓A	87.3	68.0	0.01
E. coli	carA (DEVD)	DEVD↓G	44.8	172.0	0.01
E. coli	dnaK	DEVD↓G	19	1,077.0	0.02
E. coli	fbaA	DGVD↓N	39	321.0	0.01
E. coli	folC	DHTD↓W	247	550.0	0.14
E. coli	purB	DELD↓H	151	320.9	0.05
E. coli	htpG	DEVD↓E	139	192.5	0.03

products co-migrated with the human caspase-3 large and small subunits on the gel, or because no cleavage was observed at the concentrations tested. Lack of biochemical confirmation of these substrate cleavage-sites in the purified system does not necessarily mean that these substrates were not cleaved in *E. coli* lysate. Possible reasons for this include altered substrate conformation in complexes with other proteins in the *E. coli* lysate, or partial unfolding due to the action of cellular chaperones. Our kinetic validation of protease cleavage-sites identified through N-terminomics is the first of its kind and serves as a general proof-of-principle that high quality MS/MS data from single peptides accurately identifies protein N-termini and sites of proteolytic cleavage.

Substrate engineering. We next designed a panel of mutants to dissect how substrate features affect the rate of proteolysis. The *E. coli* protein carA was chosen as a template for engineering. The protein expressed well and the cleavage products were clearly distinct from the full-length precursor allowing for accurate quantitation in our cleavage assay. The wild type carA is cleaved at the sequence DNPD↓A in a short loop linking a β-strand to an α-helix. Mutants were designed to optimize the cleavage-site amino acid sequence to DEVDG (Thornberry, Rano et al. 1997; Stennicke, Renatus et al. 2000), to extend the cleavage-site loop making it more flexible (GSGSGDNPD↓AGSGSG), and both optimize the amino acid sequence and extend the loop (GSGSGDEVD↓GGSGSG). These carA mutants were evaluated in our cleavage assay to determine their respective k_{cat}/K_M values (**Fig. 3.31, 3.32**). The

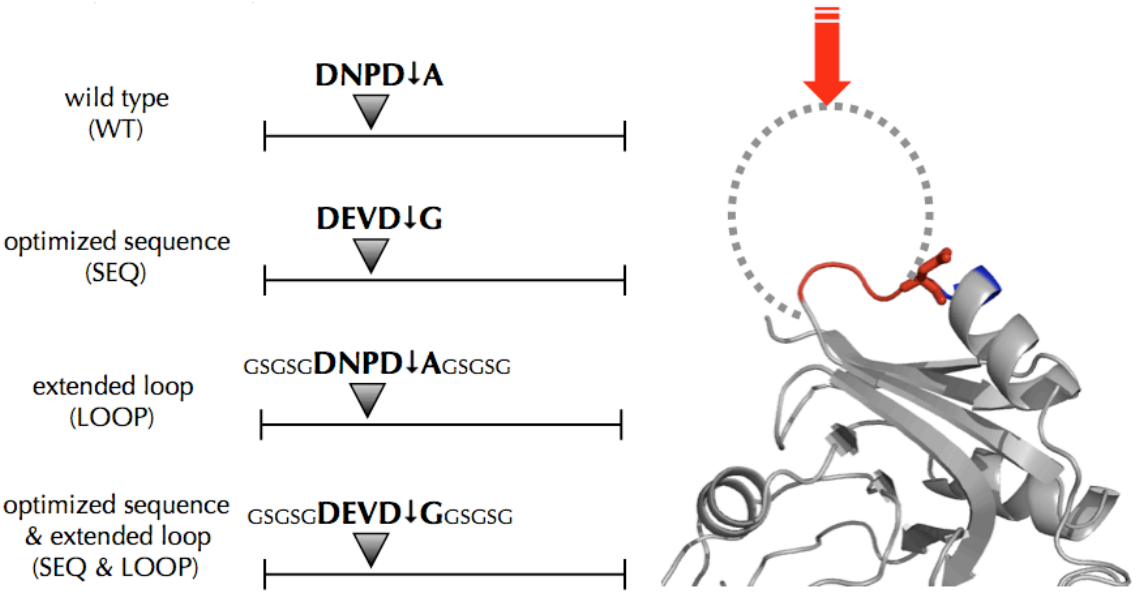


Figure 3.31 Engineered carA cleavage-site mutants. The *E. coli* caspase-3 substrate was engineered to dissect the contribution of sequence and structure on cleavage efficiency. The various constructs are shown with an optimized sequence, an extended loop, and a combination of both.

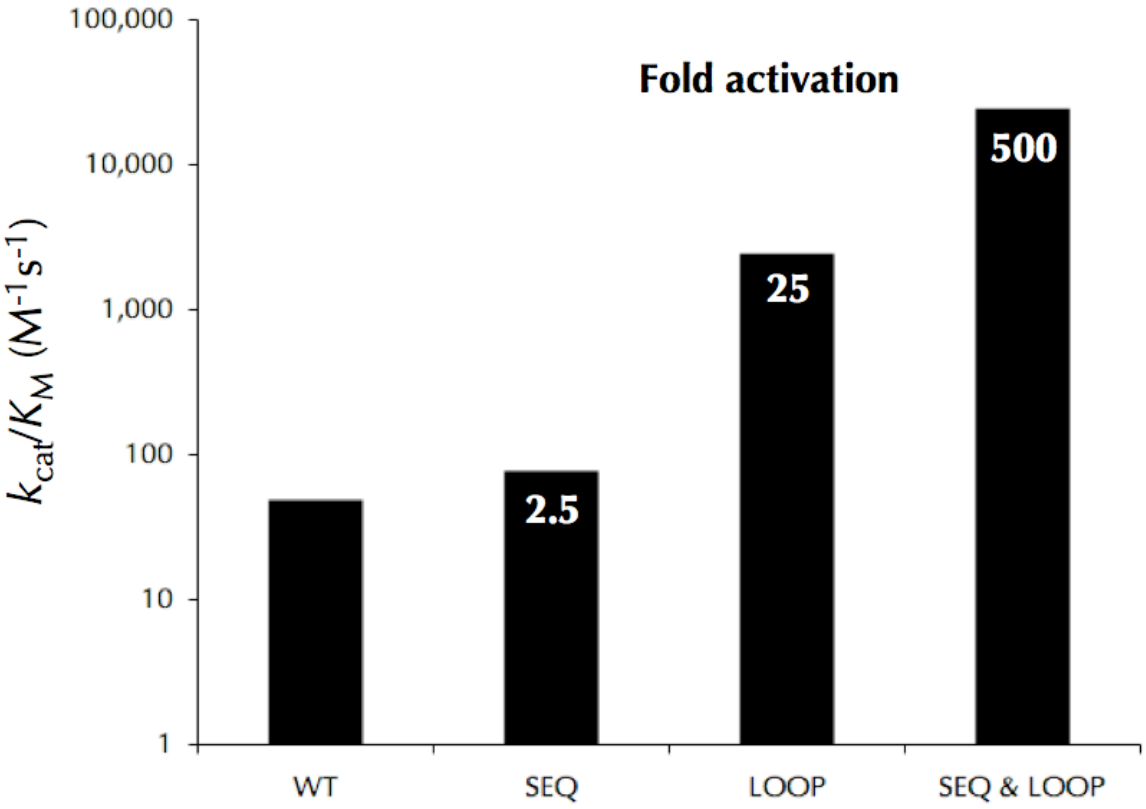


Figure 3.32 Engineered carA mutants are cleaved more efficiently than wild type. The *E. coli* caspase-3 substrate was engineered to dissect the contribution of sequence and structure on cleavage efficiency. Relative rates of cleavage were measured for these mutants revealing that an extended loop conformation improves the k_{cat}/K_M more dramatically than an optimized sequence. However, an optimized sequence in parallel with an extended loop synergizes to elicit efficient cleavage.

optimized sequence was cleaved 2.5 times better than wild type, and the extended loop mutant was cleaved over 25 times better than wild type. Importantly, combining the optimized sequence with the extended loop dramatically increased the k_{cat}/K_M to over 500 fold above wild type making it comparable to some natural caspase substrates, and defines a threshold that *bona fide* substrates should satisfy or exceed (see below). This suggests that an extended loop is more influential than an optimal sequence in sensitizing a substrate to proteolytic cleavage, and flexible extended sites are cleaved dramatically better with an optimized amino acid sequence. These results strongly support the idea that substrates co-evolve with proteases to present both an extended loop and an optimal cleavage-site sequence.

Kinetic comparison of *E. coli* substrates with natural human caspase-3 substrates. A goal of this study was to compare the specificity of caspase-3 on an unbiased natively folded substrate proteome (*E. coli*) with a naturally evolved one (human) containing caspase signaling targets. The extensive characterization of apoptotic caspase substrates in the literature is a valuable resource from which we chose several well-characterized substrates for our study (Fischer, Janicke et al. 2003; Timmer and Salvesen 2007). Kinetic values were determined experimentally in our cleavage assay, or collected from reported literature values (Casciola-Rosen, Nicholson et al. 1996; Stennicke, Jurgensmeier et al. 1998) (**Fig. 3.33**). Natural caspase-3 substrates measured in our assay had significantly higher k_{cat}/K_M values than non-natural *E. coli*

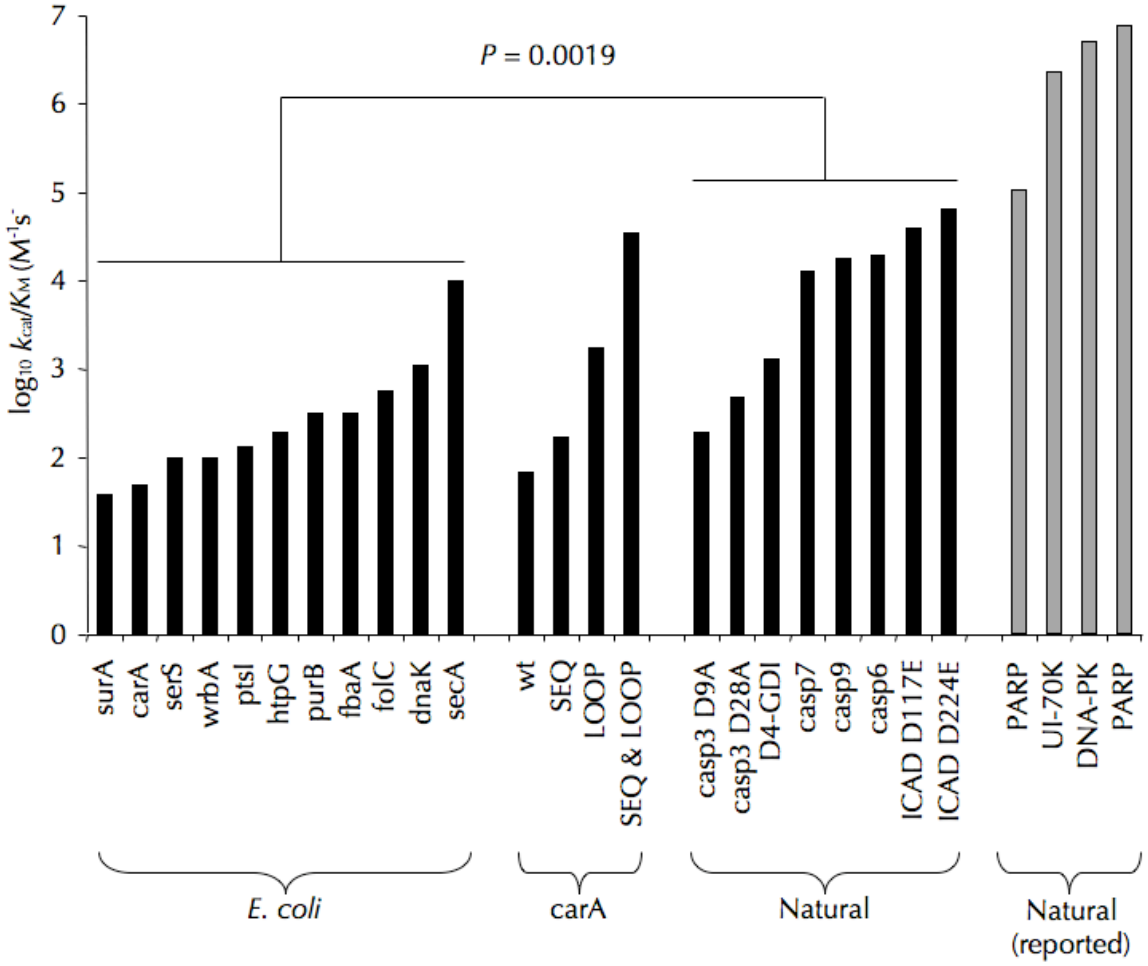


Figure 3.33 Natural human caspase-3 substrates are kinetically superior to *E. coli* substrates. Human caspase-3 cleaves most *E. coli* substrates with k_{cat}/K_M between 50 and 2,000 $M^{-1}s^{-1}$, while several biologically relevant human caspase-3 substrates were cleaved with values greater than 10,000 $M^{-1}s^{-1}$. The propensity of natural substrates to be kinetically superior to E. coli substrates was shown to be statistically significant (p=0.0019). Engineering the *E. coli* substrate carA to contain an optimized cleavage-site sequence in an extended loop improved the k_{cat}/K_M value to over 30,000 $M^{-1}s^{-1}$. Some natural substrates have k_{cat}/K_M values greater than 30,000 $M^{-1}s^{-1}$, implying additional mechanisms, such as exosite interactions, to enhance catalysis.

substrates (p = 0.0019). Only secA and the optimized carA mutant were able to approach values comparable to natural substrates. We cannot rationalize how some natural substrates are cleaved more efficiently than the optimized carA mutant due to the amino acid sequence or local structure of their cleavage-sites. Despite minor variations in experimental conditions between our study and others in the literature, their reported kinetics show extremely efficient cleavage of natural substrates, which reiterates our findings and suggests that these natural caspase-3 substrates may have evolved exosites to increase their rates of cleavage. Importantly, it is not certain that limited proteolysis of substrates that results in gain-of-function needs to be rapid, but there is a good chance that loss-of-function cleavages should be rapid enough to remove biological activity, as proposed earlier (Timmer and Salvesen 2007). However, all of the activating cleavages of natural substrates in our study do maintain elevated cleavage kinetics.

DISCUSSION

We have successfully probed the specificity and structural preferences of human caspase-3 and Staphylococcal GluC in the context of a folded protein substrate library, overcoming the limitation of other methods that unlink amino acid composition from protein structure. New focused proteomic technologies have enabled us to address this question on a proteome-wide level. Although the number of caspase-3 and GluC cleavages-sites in *E. coli* proteins that we report is not extraordinary, these cleavage-sites are the most relevant for our structure-

function study of protease substrates. In preliminary experiments we found that incubation with copious amounts of protease produces hundreds of cleavage sites, including multiple cleavage-sites per substrate. Multiple cleavages in a protein are likely to destroy its native organization, thereby precluding any structural interpretation. Accordingly, our experimental conditions produced mainly single cleavage-sites in *E. coli* proteins, thus making our structural interpretation of these cleavage-sites valid.

Despite conditions of limited proteolysis, there remains a massive background of exposed N-terminal amines originating from the *E. coli* lysate, and not a result of exogenous proteolytic cleavage. It is a daunting task to systematically evaluate and identify protease-of-interest cleavage-sites from these hundreds and thousands of N-terminal peptides generated by N-terminal proteomic analysis. We used three criteria to circumvent this problem: removing annotated co-translational proteolytic events, as well as cleavage-sites also found in control samples, and only including cleavage-sites that corresponded to the strict P1 specificity of the proteases tested. This intrinsic specificity filter is not applicable to all proteases, especially those with weak specificity requirements, where incomplete sampling will account for some of the observed differences found in the protease treated sample and not in control samples. Conversely, additional complications could arise from the proteolytic activation of other proteases when using an endogenous substrate proteome. Thus care must be exercised when designing N-terminomic experiments and interpreting the results.

These techniques are not without caveats relating to sensitivity of detection. However, the signature of proteolysis obtained from N-terminomics gives sufficient data for clear-cut analysis of specificity parameters. N-terminomics cannot reveal what quantity of precursor substrate is cleaved to generate the proteolytic fragments identified, even in conjunction with techniques employing heavy and light isotope methodologies. Consequently, we coupled traditional biochemical techniques to quantitate proteolysis and confirm proteomic results from representative substrates.

As expected, a clear relationship was observed between the number of cleavage-site peptides and the caspase-3 concentration used to generate them (**Fig. 3.34**). However, the k_{cat}/K_M values did not appear to correlate well with the concentration of caspase-3 needed to identify cleavage-sites by N-terminal proteomics (**Fig. 3.35**). This result is likely explained by two dominant factors: (1) the well-documented phenomenon that peptides do not all ionize equally well, creating bias for cleavage-site peptides that do ionize well, and (2) proteins are not maintained at equal levels in *E. coli*, and thus the effective substrate concentration varies dramatically for different proteins. We addressed the second concern by correlating the reported abundance levels of the substrates we identified for each concentration of human caspase-3 tested (**Fig. 3.36**)

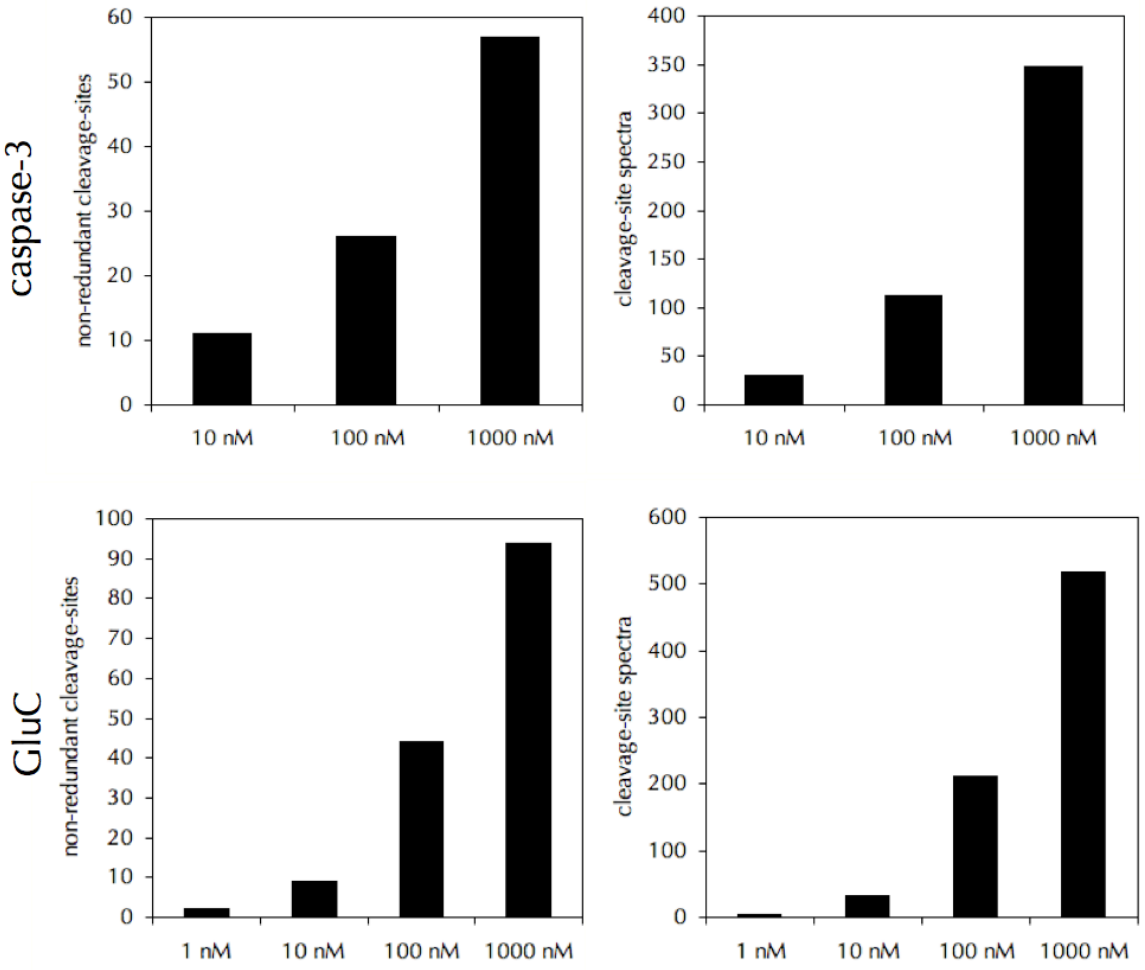


Figure 3.34 Cleavage-site peptide identification is related to protease concentration. The number of cleavage-sites and MS/MS spectra from cleavage-sites increases as the concentration of protease applied increases.

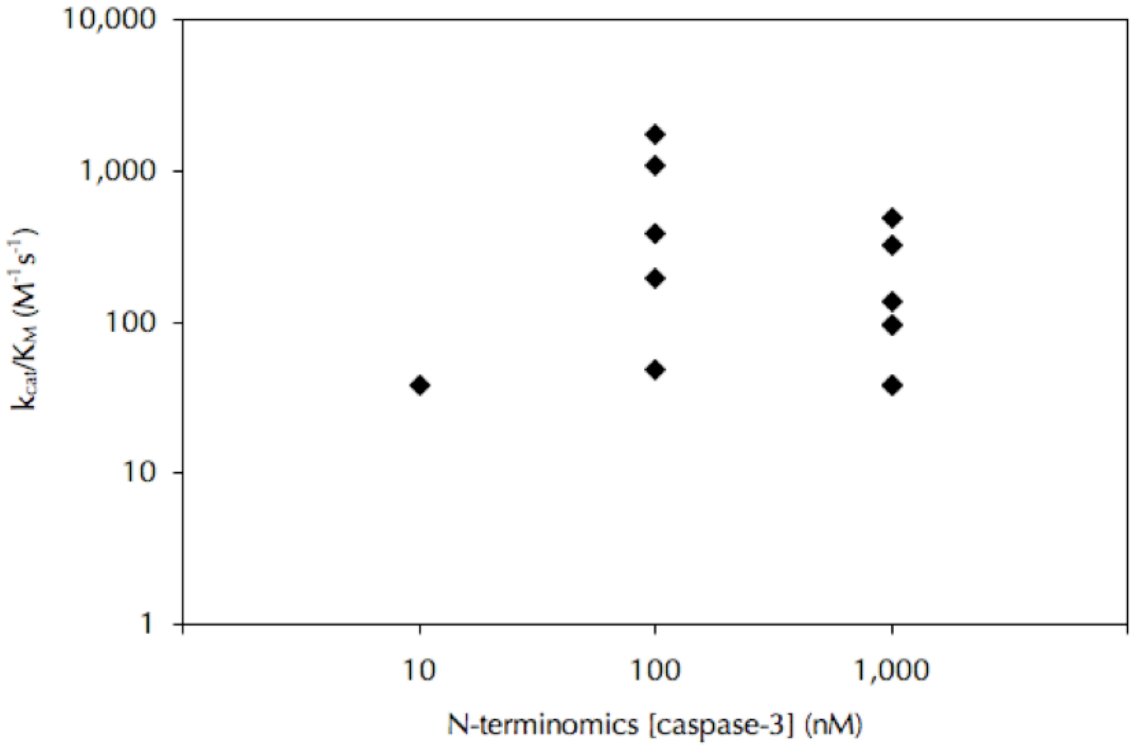


Figure 3.35 Cleavage efficiency is not the dominant factor relating to N-terminomic peptide identification. Catalytic rates were measured for many *E. coli* substrates, and grouped according to the concentration of caspase-3 used for N-terminomics. There was no clear relationship between k_{cat}/K_M values and N-terminomics protease concentration used.

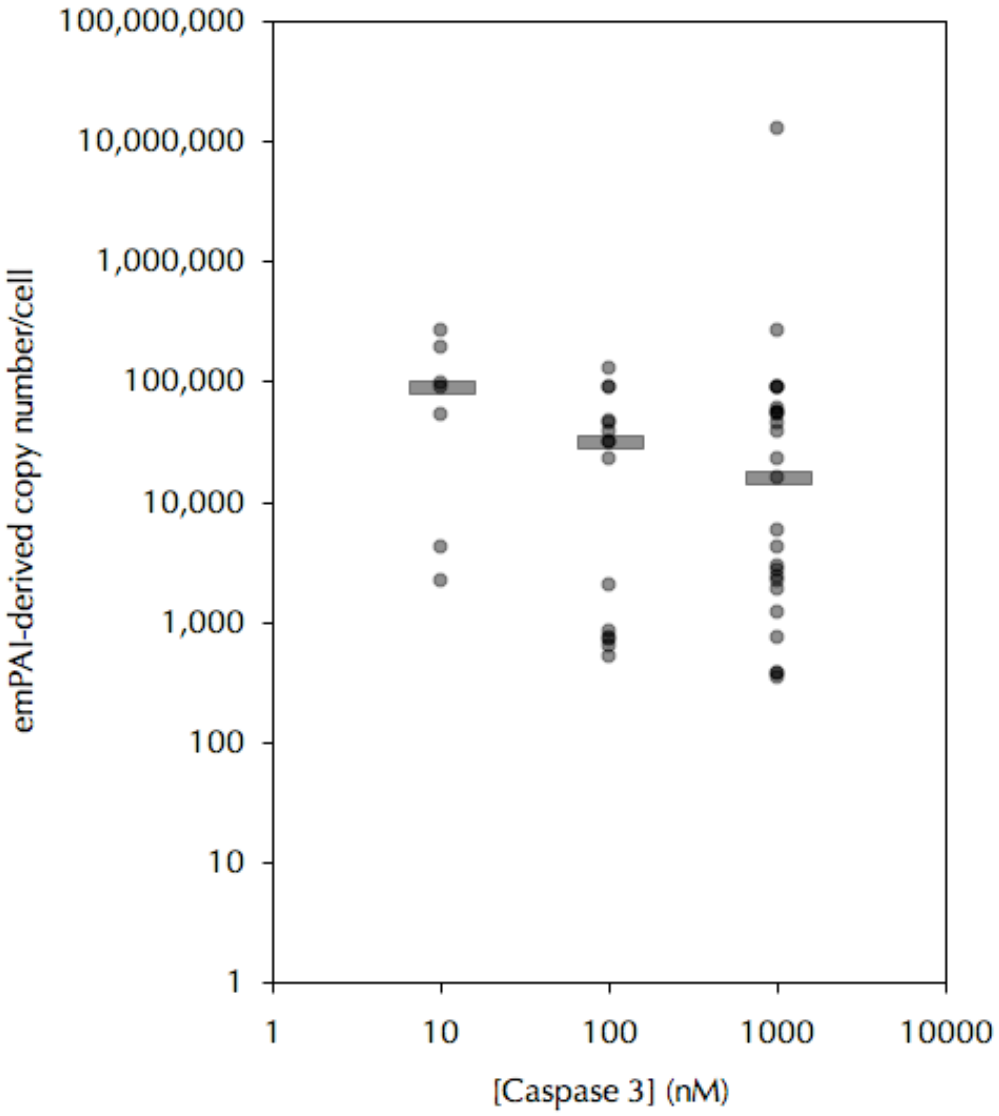


Figure 3.36 Substrate abundance levels of *E. coli* proteins found to be caspase-3 substrates. Caspase-3 substrates in *E. coli* found at the lowest concentration of protease with N-terminomics were slightly more abundant than substrates found at higher caspase-3 concentration. Likewise, substrates found at intermediate caspase-3 concentration were more abundant than substrates found only at the highest protease concentration. However, the dynamic range of caspase-3 substrate abundance found by N-terminomics covered nearly 5 orders of magnitude, making the differences seen between N-terminomics less substantial in comparison.

(Ishihama, Schmidt et al. 2008). Understandably, abundant substrates tended to be identified at lower caspase-3 concentration, but the caspase-3 N-terminomics revealed cleavages in proteins varying in concentration by nearly 5 orders of magnitude. This suggests that protein abundance is of modest importance in the acquisition of N-terminomics data, and attests to the sensitivity of our N-terminomics methodology, which we attribute to the affinity enrichment of cleavage-site derived peptides, and the concomitant massive sample simplification where irrelevant peptides are discarded. While preceding proteomic studies have produced lists of numerous substrate cleavage-sites (Gevaert, Goethals et al. 2003; McDonald, Robertson et al. 2005; Van Damme, Martens et al. 2005; Dean and Overall 2007; Enoksson, Li et al. 2007; Timmer, Enoksson et al. 2007; Dix, Simon et al. 2008; Impens, Van Damme et al. 2008; Mahrus, Trinidad et al. 2008; Schilling and Overall 2008), fundamental aspects of kinetics, recognition of structured elements, and other factors influencing proteomic detection have remained unanswered until now. Our results suggest that substrate abundance, cleavage-site k_{cat}/K_M, and the capacity of the resultant peptide to ionize all contribute to the identification of cleavage-sites by N-terminal proteomics. This study is the first to biochemically confirm and kinetically evaluate the significance of proteome-wide proteolysis, allowing us to address the sequence/structure relationship that underlies limited proteolysis by signaling proteases such as caspase-3.

A critical feature of our study was the choice of *E. coli* lysate as a substrate library. It contains approximately 4500 proteins representing a variety of structural conformations. Moreover, *E. coli* does not possesses any known Asp or Glu specific proteases, and therefore has no evolutionary pressure to alter the distribution of these amino acids throughout its proteome. The absence of endogenous caspase- or GluC-like proteases also substantially diminished the possibility of exosite interactions that would skew our results. The folding and stability of *E. coli* proteins should also be compatible with the proteases tested because it normally grows at human body temperature, and therefore its proteins should be evolutionarily selected to express folded proteins that are stable at 37ºC – the temperature we use in our assays. Likewise, the proteases tested are from humans or the commensal skin bacteria *Staphylococcus aureus*, which produce the protease GluC.

The starting hypothesis stated that we would find cleavage-sites predominantly located distinct from regions of ordered secondary structure (helices and sheets). Our data of cleavages on folded proteins with reported structures falsify this hypothesis. Cleavage is almost as frequently observed in α-helices as in regions without secondary structure. Are the helices unfolding, or can proteases cleave the helices directly? The parsimonious explanation is that helices are cut without unfolding; however, attempts to dock helical substrates in the active site of caspase-3 failed due to its deep and narrow active site cleft (unpublished observation). Indeed, previous structures of caspase-3 covalently bound to a tetrapeptide inhibitor show an extended conformation in the active

site. Conversely, GluC presents a broad and exposed active site that helical structures may be able to access; suggesting that unfolding of cleavage-sites may not be required for proteolysis. To reconcile the observed α-helical caspase-3 cleavage-sites, we propose flexible reordering of either the substrate or the active site, and that only the exposed P1 Asp is required to accommodate the S1 sub-site, while other sub-sites need not be occupied. Interestingly, analysis of cleavage-sites located in solved protein structures from the CutDB database (http://cutdb.burnham.org/) also showed a high frequency of cleavage-sites in α-helices and loops, with significantly less cleavage in β-strands (personal communication) (Igarashi, Eroshkin et al. 2007).

An alternative explanation for helical substrate cleavage is a local helical unfolding concomitant with protease binding, and we find many examples of short helices positioned within larger loops that may be in a dynamic equilibrium fluctuating between helix and loop in solution. However, we find no other indications, such as enhanced local temperature factors in the structures, which could account for flexibility of the cleaved helices. Unfolding of substrates is more difficult to envision for cleavage-sites in long stretches of α-helix that may not be as structurally dynamic, and yet these are cleaved also. We found two examples of cleavage-sites in long α-helixes resolved by NMR (**Fig. 3.37, 3.38**), which excludes the possibility that α-helixes are formed during protein crystallization. Likely a combination of dynamic α-helices sampling loop conformations, and rigid helices with a protruding P1 amino acid that accommodates the protease

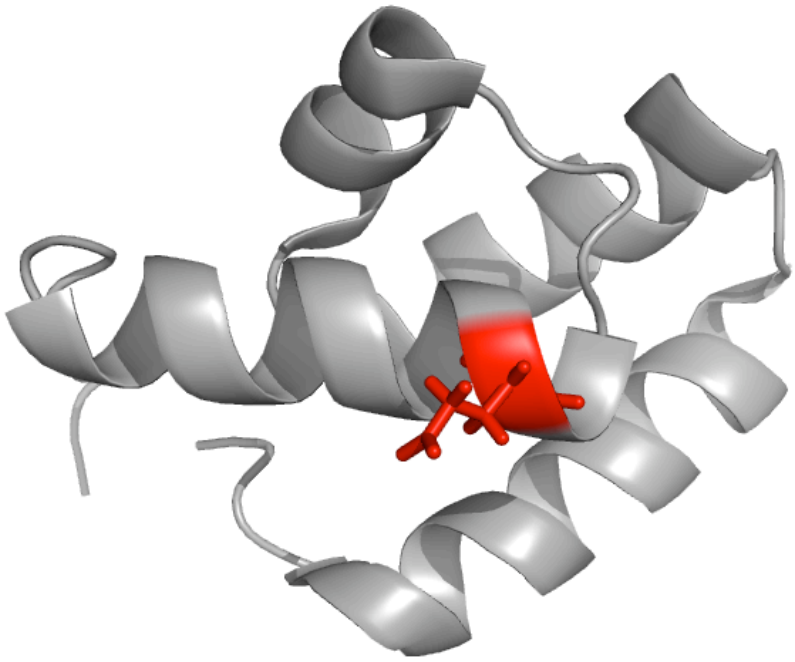


Figure 3.37 GluC cleaved protein yjbJ in an α-helix resolved by NMR. Protein yjbJ (P68206) is cleaved at E19. The P1 residue is shown in red in stick format.

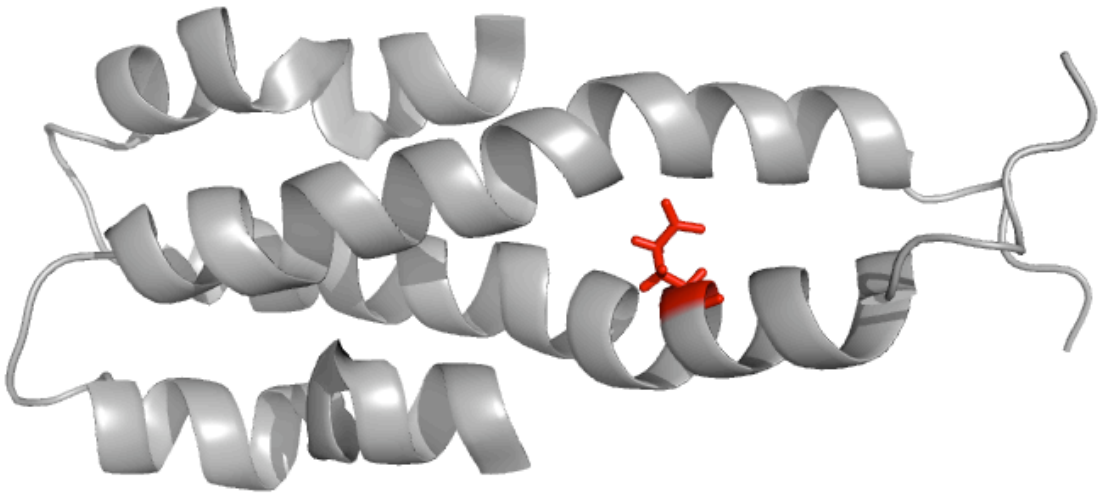


Figure 3.38 GluC cleaved DNA-binding protein H-NS in an α-helix resolved by NMR. DNA-binding protein H-NS (P0ACF8) is cleaved at E44. The P1 residue is shown in red in stick format.

specificity, together account for the observed helical cleavage-sites in *E. coli* proteins.

We analyzed *E. coli* cleavage-sites for common biophysical features using the DSSP database of standardized secondary structure assignments (ftp://ftp.ebi.ac.uk/pub/databases/dssp/). The solvent accessibility and hydrogen bonding energy of each amino acid for every cleavage-site was collected, and the average values and standard deviations were plotted from P10 to P10' (**Fig. 3.39** thru **3.42**). Interestingly, no obvious biophysical features were associated with cleavage-sites.

We propose that the folded *E. coli* proteome serves as a baseline of unbiased caspase-3 cleavage specificity and kinetics. How does a protease elevate its activity above this baseline (k_{cat}/K_M values in the range 10^2-10^3 $M^{-1}s^{-1}$) several orders of magnitude to tackle natural substrates (k_{cat}/K_M values in the range 10^4-10^6 $M^{-1}s^{-1}$)? We demonstrate that one way to improve substrate hydrolysis is to extend the cleavage-site loop away from the surface of the protein and incorporate an optimal sequence. Both of these features separately enhance the cleavage efficiency by caspase-3, together synergizing to account for a 500-fold enhancement (k_{cat}/K_M). All of the catalytic rates for *E. coli* proteins were less than this, probably because of suboptimal position and composition of their cleavage-sites, or steric hindrance from other regions in protein structures.

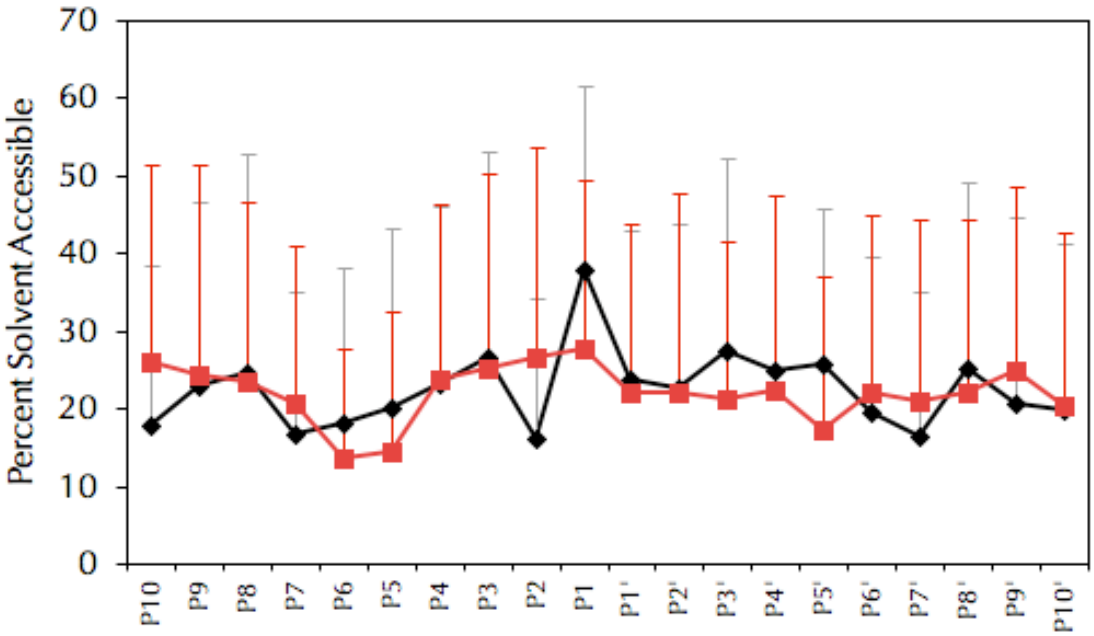


Figure 3.39 Solvent accessibility of caspase-3 cleavage-sites. The solvent accessibility of *E. coli* proteins with solved structures was determined for caspase-3 cleavage-sites (black) or control sites in *E. coli* proteins with Asp in P1 (red). Human caspase-3 cleaved substrates containing P1 residues that were solvent accessible (≥30%). The hydrophobic trend of human caspase-3 in the P2 position seems to be reflected in the decreased solvent accessibility at that position.

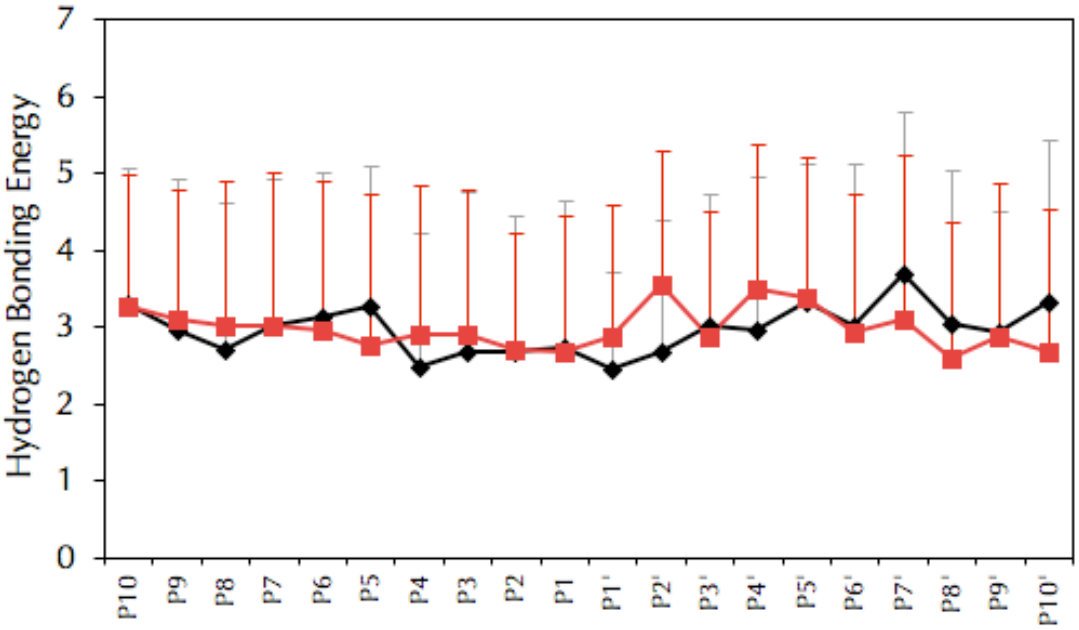


Figure 3.40 Hydrogen-bonding of caspase-3 cleavage-sites. No obvious decrease in hydrogen-bonding was observed for human caspase-3 cleavage-sites (black) compared to control Asp sequences in *E. coli* proteins (red).

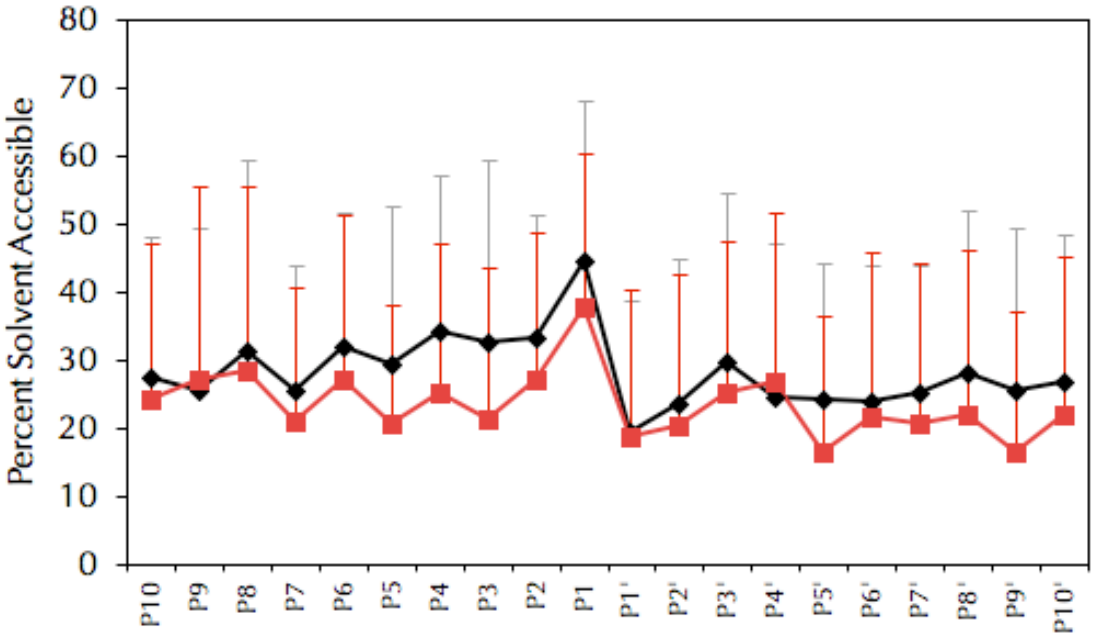


Figure 3.41 Solvent accessibility of GluC cleavage-sites. The solvent accessibility of *E. coli* proteins with solved structures was determined for GluC cleavage-sites (black) or control sites in *E. coli* proteins with Glu in P1 (red). Human caspase-3 cleaved substrates containing P1 residues that were solvent accessible (≥30%). The hydrophobic trend of GluC in the P1' position seems to be reflected in the decreased solvent accessibility at that position.

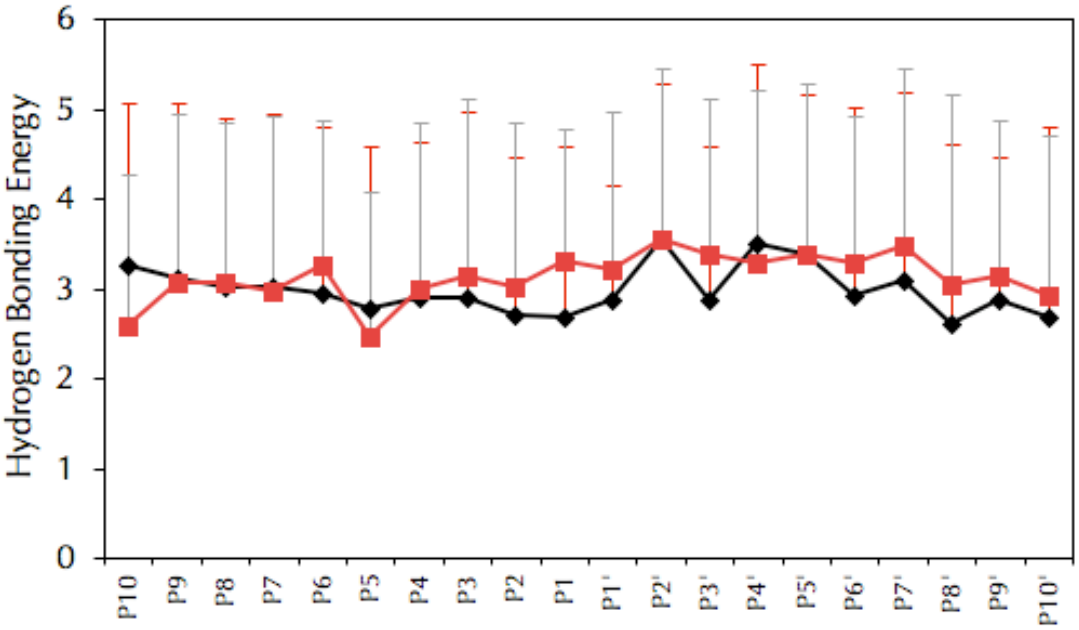


Figure 3.42 Hydrogen-bonding of GluC cleavage-sites. No obvious decrease in hydrogen-bonding was observed for GluC cleavage-sites (black) compared to control Glu sequences in *E. coli* proteins (red).

The catalytic enhancement seen in natural substrates is likely the result of co-evolution of human caspase-3 and its substrates, producing cleavage-sites on extended unstructured loops with optimized amino acids in P4, P2, P1, and P1'. Although we have not directly ruled out the possibility that α-helices can be cleaved as efficiently as extended loops, this possibility seems unlikely in light of the carA engineering experiments, which show that a flexible loop is the critical feature of kinetically superior sites. Therefore, even though caspase-3 can cleave α-helices inefficiently, we do not anticipate natural signaling substrates of caspase-3 to be cleaved in α-helices. One further way catalysis is elevated is through the utilization of exosites, which are surfaces distinct from the active site, providing a weak secondary interaction that, together with the catalytic cleft interaction, provides enhanced and very specific substrate catalysis. This has been convincingly demonstrated for the protease thrombin and its substrate fibrinogen (Stubbs and Bode 1995), as well as matrix metalloproteases and their cognate substrates (Overall 2002). This mechanism may account for the extremely efficient cleavage kinetics reported for several natural caspase-3 substrates.

Our results have immediate implications in several areas of protease research. First, we present data defining a kinetic threshold that many physiologically relevant substrates of signaling proteases, such as caspase-3, do fulfill. We propose that substrates be challenged with this kinetic litmus test to strengthen claims of biological relevance. Likewise, other proteases could be investigated in a similar manner to define protease-specific kinetic thresholds for

natural substrate validation. Second, our specificity and structural analysis of caspase-3 cleavage-sites can be incorporated into substrate prediction algorithms by weighting positions P4, P2, and P1' with Asp fixed in P1, and limiting predictions to sites in known or predicted loop structures. This dual specificity-structure filter could dramatically diminish the exorbitant number of false positive predictions, making biochemical evaluation possible where it would otherwise be unrealistic. Indeed, the methods and results of this study serve as a broadly applicable template to characterize the amino acid specificity and structural permissiveness of diverse proteases.

MATERIALS & METHODS

***E. coli* lysate preparation.** *E. coli* K12 strain MG1655 was cultured in 3 liters of 2x yeast tryptone medium at 37ºC in a shaking incubator. Bacterial cultures were collected at OD_{600} = 1.0, and pelleted by centrifugation. Cell pellets were resuspended in 1x assay buffer (20 mM PIPES pH 7.2, 100 mM NaCl, 10% sucrose, 0.1% CHAPS, and 1mM EDTA), and ruptured by sonication on ice for 3 minutes at 50% intensity and 50% duty cycle. Insoluble material was removed by centrifugation, and clarified with a 0.45 µm filter yielding a final concentration of 20 mg/mL total protein. The resultant lysate was aliquoted and frozen at -20ºC. No precipitant was observed upon thawing of the frozen lysate, indicating that proteins in the lysate were folded and soluble.

Expression and purification of human caspase-3. Human caspase-3 was expressed as previously described (Denault and Salvesen 2008). Human caspase-3 was previously cloned into pET23b with a C-terminal 6 histidine tag. *E. coli* strain BL21-DE3 was transformed with this plasmid and selected for using ampicillin at 50 µg/mL. Transformant colonies were used to inoculate 3 liters of 2x yeast tryptone broth, which were induced to express caspase-3 at an OD_{600} = 1.0 with 0.2 mM Isopropyl β-D-1-thiogalactopyranoside, and maintained at 20ºC for 18 hours. Bacteria were pelleted by centrifugation at 4,000 RCF for 10 minutes at 4ºC. Cells were resuspended in buffer A (50 mM HEPES pH 8.0, 100 mM NaCl), sonicated on ice for 6 minutes, centrifuged at 10,000 RCF at 4ºC for 30 minutes in an SS34 rotor, and clarified with a 0.45 µm filter. Histidine tagged protein was bound to Ni-NTA agarose (Invitrogen) in batch format at 4ºC, washed extensively with buffer B (50 mM HEPES pH 8.0, 500 mM NaCl), and eluted with a gradient of buffer A and buffer C (50 mM HEPES pH 8.0, 100 mM NaCl, and 200 mM imidazole). Fractions were analyzed by SDS-PAGE for abundance and purity of the P20 and P10 bands. Appropriate fractions were pooled, and dialyzed overnight in 50 mM HEPES pH 7.2, 100 mM NaCl, and the protein concentration was measured by absorbance (A_{280}). The active concentration of caspase-3 was determined by titration with the caspase inhibitor p35. Residual caspase-3 activity was measured with the fluorogenic peptide substrate Ac-DEVD-AFC in the SpectraMax Gemini EM plate reader (Molecular Devices). The residual caspase-3 rates were plotted against the concentration of inhibitor, and a linear

fit was used to calculate the concentration of inhibitor where the slope intersects the x-axis; representing the concentration of active caspase-3.

Cleavage of *E. coli* lysate proteins by purified proteases. Aliquots of frozen *E. coli* lysate were thawed at room temperature and spun down at 13,000 RCF for 5 minutes at 4ºC prior to protease treatment to remove any potential aggregated or unfolded proteins that had precipitated in the freeze/thaw process. No visible protein pellet was observed, indicating *E. coli* lysate proteins maintained their native conformations and remained soluble. Purified caspase-3 (described above), and GluC (Roche) were prepared at 5x or 10x final concentration in 1x assay buffer (20 mM PIPES pH 7.2, 100 mM NaCl, 10% sucrose, 0.1% CHAPS, 1mM EDTA, and 5 mM fresh DTT) and pre-activated at 37ºC for 15 minutes. Active protease was then added to *E. coli* lysate pre-incubated at 37ºC, mixed and the cleavage reaction was maintained at 37ºC for 1 hour. Dry guanidine HCl was added to 6 M final concentration, DTT was added to 10 mM, and the samples were boiled for 10 minutes to terminate the cleavage reaction by denaturation.

N-terminomic sample preparation. Samples were prepared as described previously (Timmer, Enoksson et al. 2007). Denatured samples boiled in 6 M guanidine HCl, were cooled to room temperature, whereby cystine sulfhydryls were alkylated with 30 mM iodoacetamide at room temperature in the dark for 30 minutes. Dry *o*-methylisourea was added to 0.5 M final concentration,

the pH was adjusted to 10.5 with NaOH. Samples were kept at 4ºC overnight to allow lysine guanidination without N-terminal amine modification. PD-10 column size exclusion chromatography was then used to buffer exchange samples into 8 M urea, 50 mM HEPES pH 7.8, and 100 mM NaCl, and remove amine containing small molecules as well as any residual DTT. Free N-terminal amines were reacted with 5 mM sulfo NHS-SS-biotin (Pierce) for 1 hour at 37ºC, and quenched by the addition of 50 mM ammonium bicarbonate (AmmBic). Samples were again buffer exchanged using PD-10 columns into 8M M urea, 50 mM HEPES pH 7.8, 100 mM NaCl to remove unconjugated sulfo NHS-SS-biotin. 10 mM AmmBic buffer was used to dilute samples down to 2 M urea, which allowed for efficient trypsin digestion at 37ºC overnight. Boiling samples for 10 minutes inactivated the remaining trypsin, and any flocculant was removed by centrifucation at 13,000 RCF for 5 minutes at room temperature. The soluble samples were neutated with pre-washed high capacity neutravidin agarose resin (Thermo) for 30 minutes to allow biotinylated peptides to be captured. Disposable 2 mL polystyrene columns were used to capture the resin, and allowed for extensive washing on the column with 2 M urea, then with 10 mM AmmBic. The resin was transferred to 2 mL tubes, where immobilized biotinylated peptides were eluted with the addition of 5 mM tris(2-carboxyethyl)phosphine at 37ºC for 30 minutes. Samples were then loaded onto disposable micro bio-spin columns (Bio-Rad) allowing liberated peptides to be collected. In preparation for LC-MS/MS, the samples were loaded onto C_{18} Sep-Pak Vac 6cc cartridges (Waters), washed with 0.1% TFA, and eluted with 50% MeCN, 0.1% TFA. Peptides were

dried in a vacuum centrifuge, and solubilized in 50 µL of 0.1% TFA. Samples were then analyzed by LC-MS/MS to identify N-terminal peptides.

Sample analysis by nano LC–MS. The automated NanoLC-LTQ system consists of an Eksigent Nano-2D LC, an autosampler, a switch valve, a C_{18} trap column (Agilent), a capillary separation column (100 µm i.d. ×15 cm length, Michrom Magic C18AQ), and a LTQ ion-trap mass spectrometer (Thermo Electron). The separation column is mounted into the Michrom ADVANCE spray source. First, trypsin-digested peptides (5 µl) were loaded by autosampler on to the trap column in 100% solvent A (2% acetonitrile and 0.1% formic acid) using a flow rate of 10 µl/minute for 4 minutes. After sample loading and washing, the valve was switched, and the gradient was delivered to the trap and separation column at 500 nl/minute. Peptides were separated with a 100-120 minute linear gradient of 10–60% solvent B (80% acetonitrile and 0.1% formic acid), and was then eluted directly into the LTQ spectrometer. The fully automated NanoLC-LTQ was operated via an Instrument Method of Xcalibur. MS/MS spectra were collected automatically during the LC-MS runs. Each scan was set to acquire a full MS scan followed by four MS/MS scans of the four most intense ions from the preceding MS scan. Each sample was analyzed 3 or more times.

Database searching. After data acquisition, MS/MS spectra were combined and loaded on a Sorcerer 2 system (SageN) and searched against a concatenated forward and reverse semi-tryptic *E. coli* database (Swiss-Prot)

using Sorcerer SEQUEST. A molecular mass of 88 Da was added to the differential search of all N-termini to account for NHS-SS-Biotin modification. A molecular mass of 57 Da was added to all cysteine residues to account for carboxyamidomethylation, and a differential search was performed for methionine+16 Da (oxidation) and lysine+42 Da (guanidination). SEQUEST search results were filtered for peptides identified from two or more spectra with a minimum probability score of 0.8, and a cross correlation value (Xcorr) of at least 2.0. The peptide false discovery rate was less than 2%.

***E. coli* protein expression and purification**. *E. coli* proteins found to be substrates of caspase-3 chosen for biochemical analysis were retrieved from the Genobase collection of open reading frames cloned into an N-terminal 6 histidine tagged inducible expression vector and maintained in frozen stock cultures (Kitagawa, Ara et al. 2005). Clones were streaked out for isolation on chloramphenicol plates. Single colonies were used for protein expression in 500 mL of 2x yeast tryptone broth, inducing with 0.2 mM IPTG at OD_{600} = 0.8, and maintained at 30ºC for 4 hours. Bacteria were pelleted by centrifugation, and purified as described above for caspase-3. Proteins were eluted with buffer C (50 mM HEPES pH 8.0, 100 mM NaCl, and 200 mM imidazole) supplemented with 10% sucrose. Protein concentration was determined by A_{280} and purity was verified by SDS-PAGE. Aliquots of each protein were made and stored at -80ºC.

Cloning and engineering of carA. The caspase-3 *E. coli* substrate, carA wild type was sub-cloned into pET-15b using NdeI and BamHI restriction sites. Full length carA was amplified using the NdeI containing forward oligonucleotide primer CACACACATATGATTAAGTCAGCGCTATTG, and the BamHI containing reverse primer CACACAGGATCCTTACTTAGCGGTTTTACG. Engineering the carA cleavage-site was achieved by overlapping PCR of the N- and C-terminal sections with mutations/insertions contained in the overlapping region. The optimized cleavage-site mutant was constructed from an N-terminal fragment resulting from the NdeI containing forward primer CACACACATATGATTAAGTCAGCGCTATTG and the reverse primer TCCATCTACCTCATCGCCCGCGATAATGCAGCC, and the C-terminal product of the forward primer GATGAGGTAGATGGAGCGGCGCTGGCGTTAGAA, and the BamHI containing reverse primer CACACAGGATCCTTACTTAGCGGTTTTACG. The extended loop mutant was constructed from an N-terminal fragment resulting from the NdeI containing forward primer CACACACATATGATTAAGTCAGCGCTATTG and the reverse primer CGCATCCGGGTTATCACCACTTCCACTACCGCCCGCGATAATGCAGCC, and the C-terminal product of the forward primer GATAACCCGGATGCGGGTAGCGGTAGTGGAGCGGCGCTGGCGTTAGAA, and the BamHI containing reverse primer CACACAGGATCCTTACTTAGCGGTTTTACG. The optimized cleavage-site & extended loop mutant was constructed from an N-terminal fragment resulting

from the NdeI containing forward primer CACACACATATGATTAAGTCAGCGCTATTG and the reverse primer TCCATCTACCTCATCACCACTTCCACTACCGCCCGCGATAATGCAGCC, and the C-terminal product of the forward primer GATGAGGTAGATGGAGGTAGCGGTAGTGGAGCGGCGCTGGCGTTAGAA, and the BamHI containing reverse primer CACACAGGATCCTTACTTAGCGGTTTTACG. The caspase-9 linker mutant was constructed from an N-terminal fragment resulting from the NdeI containing forward primer CACACACATATGATTAAGTCAGCGCTATTG and the reverse primer GTCCAGCTGGTCAAAGGTCCTGAGACCGCCCGCGATAATGCAGCC, and the C-terminal product of the forward primer GACCAGCTGGACGCCATTAGCAGCGCGGCGCTGGCGTTAGAA, and the BamHI containing reverse primer CACACAGGATCCTTACTTAGCGGTTTTACG. Addition of the p35 spout exosite to the N-terminus of carA was executed by amplifying the full length wild type carA with the NdeI containing forward primer CACACACATATGTTTACTACAGAATCGAGCTGGGGCAAATCCGAAAAGTATA ATTGGAAAATTAAGTCAGCGCTATTG, and the XhoI containing reverse primer CACACACTCGAGCTTAGCGGTTTTACGGTACTG, which was cloned into the corresponding sites in pET-23b (C-terminal 6 histidine tag). These constructs were expressed and purified as described above for the other *E. coli* proteins from the Genobase collection.

Statistical analysis of substrate kinetics. Statistical analysis of the cleavage kinetics of caspase-3 substrates was performed using the Student's *t*-test using an unpaired and two-tailed analysis with equal variance.

Determination of K_M for caspase-3 substrates. The K_M for caspase-3 cleavage of various protein substrates was calculated by using the modified Michaelis-Menten equation when two competitive substrates were present simultaneously in the enzyme reaction (Morrison 1982). The hydrolysis of the fluorescent synthetic tetrapeptide Ac-DEVD-AFC was monitored by the change in fluorescence in the presence of recombinant proteins of *E. coli* or human origin that were previously established to be caspase-3 substrates.

The initial velocity (v_{DEVD}) for hydrolysis of Ac-DEVD-AFC (*DEVD*) in the presence of a competitive non-fluorescent substrate of caspase-3 (*Prot*) was fit using the equation:

$$v_{DEVD} = \frac{\frac{V_{DEVD}[DEVD]}{K_{M_{DEVD}}}}{1 + \frac{[DEVD]}{K_{M_{DEVD}}} + \frac{[\mathrm{Pr}ot]}{K_{M_{\mathrm{Pr}ot}}}} \qquad \text{(Equation 1)}$$

In Equation 1, [*DEVD*] and [*Prot*] are the concentrations for Ac-DEVD-AFC and the protein substrate, respectively, while K_M represents the Michaelis-Menten constant for the indicated substrate.

The buffer of recombinant proteins was exchanged to standard caspase assay buffer (10 mM Pipes pH 7.2, 100 mM NaCl, 5% sucrose, 0.1% CHAPS, 10 mM DTT) and the proteins were concentrated to 50-700 μM stock solutions, determined by absorbance at 280 nm. For determination of K_M, Ac-DEVD-AFC was serially diluted in caspase buffer (2-300 μM final concentration), mixed with the protein of interest at constant concentration and incubated in 96-well plates at 37°C for 15 min. Caspase-3, pre-incubated in assay buffer at 37°C for 15 min, was then added with a multi-channel pipette to the substrate mix at 1 nM final concentration, and kinetics of fluorescence generated by AFC was recorded immediately using the SpectraMax Gemini EM plate reader (Molecular Devices). The experiment was performed using at least three different final concentrations of recombinant protein in the mix, typically ranging between 5-40 μM for mammalian substrates and 25-200 μM for *E. coli* substrates. The control contained no alternative substrate and was done under the same experimental conditions. Resulting initial velocity was plotted against the Ac-DEVD-AFC concentration and the data was fit using Equation 1. The fit with Equation 1 generated $K_{M,DEVD}$ for DEVD-AFC cleavage when no alternative substrate was present ([*Prot*]=0] or the $K_{M,Prot}$ for protein substrate cleavage when ([*Prot*] was kept constant.

ACKNOWLEDGEMENTS

The work in this chapter was supported by NIH Roadmap Initiative National Biotechnology Resource Center grant RR20843 for the Center on Proteolytic Pathways, CA69381 from the NCI, and by Training Grant 5T32CA77109-9 from the NCI. This chapter was reproduced with permission from co-authors, modified from the manuscript in review for publication: Timmer JC, Zhu W, Pop C, Snipas SJ, Eroshkin AM, Riedl SJ, Salvesen GS. Structural and kinetic determinants of protease substrates. *Nature Structure and Molecular Biology.* Accepted.

REFERENCES

Casciola-Rosen, L., D. W. Nicholson, et al. (1996). "Apopain/CPP32 cleaves proteins that are essential for cellular repair: a fundamental principle of apoptotic death." J Exp Med **183**(5): 1957-64.

Coombs, G. S., R. C. Bergstrom, et al. (1998). "Directing sequence-specific proteolysis to new targets. The influence of loop size and target sequence on selective proteolysis by tissue-type plasminogen activator and urokinase-type plasminogen activator." J Biol Chem **273**(8): 4323-8.

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-90.

Dean, R. A. and C. M. Overall (2007). "Proteomics discovery of metalloproteinase substrates in the cellular context by iTRAQ labeling reveals a diverse MMP-2 substrate degradome." Mol Cell Proteomics **6**(4): 611-23.

DeLano, W. L. (2002). "The PyMOL Molecular Graphics System." on the World Wide Web http://www.pymol.org.

Denault, J. B. and G. S. Salvesen (2008). "Apoptotic caspase activation and activity." Methods Mol Biol **414**: 191-220.

Deng, S. J., D. M. Bickett, et al. (2000). "Substrate specificity of human collagenase 3 assessed using a phage-displayed peptide library." J Biol Chem **275**(40): 31422-7.

Ding, L., G. S. Coombs, et al. (1995). "Origins of the specificity of tissue-type plasminogen activator." Proc Natl Acad Sci USA **92**: 7627-7631.

Dix, M. M., G. M. Simon, et al. (2008). "Global mapping of the topography and magnitude of proteolytic events in apoptosis." Cell **134**(4): 679-91.

Elias, J. E., W. Haas, et al. (2005). "Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations." Nat Methods **2**(9): 667-75.

Enoksson, M., J. Li, et al. (2007). "Identification of proteolytic cleavage sites by quantitative proteomics." J Proteome Res **6**(7): 2850-8.

Fischer, U., R. U. Janicke, et al. (2003). "Many cuts to ruin: a comprehensive update of caspase substrates." Cell Death Differ **10**(1): 76-100.

Gettins, P. G. (2002). "Serpin structure, mechanism, and function." Chem Rev **102**(12): 4751-804.

Gevaert, K., M. Goethals, et al. (2003). "Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides." Nat Biotechnol **21**(5): 566-9.

Gevaert, K., F. Impens, et al. (2007). "Applications of diagonal chromatography for proteome-wide characterization of protein modifications and activity-based analyses." Febs J **274**(24): 6277-89.

Harris, J. L., B. J. Backes, et al. (2000). "Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries." Proc Natl Acad Sci U S A **97**(14): 7754-9.

Hubbard, S. J., S. F. Campbell, et al. (1991). "Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors." J Mol Biol **220**: 507-530.

Igarashi, Y., A. Eroshkin, et al. (2007). "CutDB: a proteolytic event database." Nucleic Acids Res **35**(Database issue): D546-9.

Impens, F., P. Van Damme, et al. (2008). "Mechanistic insight into taxol-induced cell death." Oncogene **27**(33): 4580-91.

Ishihama, Y., T. Schmidt, et al. (2008). "Protein abundance profiling of the Escherichia coli cytosol." BMC Genomics **9**: 102.

Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." J Mol Biol **292**(2): 195-202.

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**(12): 2577-637.

Kelly, C. A., M. Laskowski, Jr., et al. (2005). "The role of scaffolding in standard mechanism serine proteinase inhibitors." Protein Pept Lett **12**(5): 465-71.

Kitagawa, M., T. Ara, et al. (2005). "Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research." DNA Res **12**(5): 291-9.

Mahrus, S., J. C. Trinidad, et al. (2008). "Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini." Cell **134**(5): 866-76.

McDonald, L., D. H. Robertson, et al. (2005). "Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides." Nat Methods **2**(12): 955-7.

Morrison, J. F. (1982). "The slow-binding and slow, tight-binding inhibition of enzyme-catalysed reactions." Trends Biochem. Sci. **3**: 102-105.

Nazif, T. and M. Bogyo (2001). "Global analysis of proteasomal substrate specificity using positional-scanning libraries of covalent inhibitors." Proc Natl Acad Sci U S A **98**(6): 2967-72.

Overall, C. M. (2002). "Molecular determinants of metalloproteinase substrate specificity: matrix metalloproteinase substrate binding domains, modules, and exosites." Mol Biotechnol **22**(1): 51-86.

Puente, X. S., L. M. Sanchez, et al. (2003). "Human and mouse proteases: a comparative genomic approach." Nat Rev Genet **4**(7): 544-58.

Salvesen, G. S. and J. M. Abrams (2004). "Caspase activation - stepping on the gas or releasing the brakes? Lessons from humans and flies." Oncogene **23**(16): 2774-84.

Schilling, O. and C. M. Overall (2008). "Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites." Nat Biotechnol **26**(6): 685-94.

Smith, M., L. Shi, et al. (1995). "Rapid identification of highly active and selective substrates for stromelysin and matrilysin using bacteriophage peptide display libraries." J Biol Chem **270**: 6440-6449.

Stennicke, H. R., J. M. Jurgensmeier, et al. (1998). "Pro-caspase-3 is a major physiologic target of caspase-8." J Biol Chem **273**(42): 27084-27090.

Stennicke, H. R., M. Renatus, et al. (2000). "Internally quenched fluorescent peptide substrates disclose the subsite preferences of human caspases 1, 3, 6, 7 and 8." Biochem J **350**(Pt 2): 563-568.

Stubbs, M. T. and W. Bode (1995). "The clot thickens: clues provided by thrombin structure." Trends Biochem Sci **20**(1): 23-8.

Thornberry, N. A., T. A. Rano, et al. (1997). "A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis." J Biol Chem **272**(29): 17907-11.

Timmer, J. C., M. Enoksson, et al. (2007). "Profiling constitutive proteolytic events in vivo." Biochem J **407**: 41-48.

Timmer, J. C. and G. S. Salvesen (2007). "Caspase substrates." Cell Death Differ **14**(1): 66-72.

Turk, B. E., L. L. Huang, et al. (2001). "Determination of protease cleavage site motifs using mixture-based oriented peptide libraries." Nat Biotechnol **19**(7): 661-7.

Vacic, V., L. M. Iakoucheva, et al. (2006). "Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments." Bioinformatics **22**(12): 1536-7.

Van Damme, P., L. Martens, et al. (2005). "Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis." Nat Methods **2**(10): 771-7.

Chapter IV

Conclusions and Perspectives

Improving substrate discovery. A significant challenge in the protease field has long been to identify the physiologically relevant substrates of proteases *in vivo*. We have developed N-terminomics, a proteomics-based tool, which now enables us to identify protease substrates from pertinent cellular or tissue samples. Despite the broad utility of N-terminomics, it is limited by factors such as sensitivity, incomplete sampling, and peptide ionization, which all proteomic methods are subject to. However, protease prediction algorithms can readily survey entire genomes *in silico* for amino acid sequences matching a protease's specificity, and can compliment the limitations of N-terminomics to improve substrate discovery and cleavage-site identification.

Prioritizing cleavage-sites by structure. Yet N-terminomics and prediction algorithms both produce immense lists of cleavage-sites that must be verified by a secondary method to quantitatively and temporally verify the initial hit. Indeed, my *in vitro* cleavage assay confirmed many of the *E. coli* proteins identified by the N-terminomics analysis as substrates of human caspase-3. Paradoxically, the validation of the cleavage-sites from these proteomic or prediction lists is vastly more challenging that the initial screen. The difficulty of substrate validation in my study of caspase-3 and GluC was substantially reduced because these proteases had very clear cleavage-site specificities, and because I had access to the Genobase collection of *E. coli* open reading frames cloned into expression vectors. Yet many proteases do not have strict

specificities or have not had their specificity determined, making it more challenging to decipher the specific substrates of a given protease. It is almost certain that most if not all N-terminomic hits will require cloning into affinity tagged expression vectors or epitope tagged vectors for biochemical or cell-based validation if specific antibodies are not available. Substrate validation is the rate-limiting step in the process; therefore, N-terminomic hits and predictions must be prioritized by some measure so as to make validation experimentally realistic. I propose that an understanding of the cleavage-site structures tolerated by proteases can aid in the prioritization of cleavage-sites identified by proteomic methods, or predicted by computational algorithms.

Caspase-3 cleaves loops *in vivo*. My investigation of the structural permissiveness of two model proteases has shown that both α-helices and loops are susceptible to cleavage, while β-strands are not. I also demonstrated that for the apoptotic protease caspase-3, a flexible extended loop structure is dramatically more influential in sensitizing a cleavage-site to proteolysis than is an optimal amino acid sequence. In finding that the natural apoptotic substrates of caspase-3 were cleaved very efficiently, we concluded that they were likely only cleaved in loop structures. This structural assessment of caspase-3 substrates indicates that proteomic efforts or prediction approaches aimed at identifying caspase-3 specific substrates should prioritize validation efforts on cleavage-sites in flexible loops before α-helical sites, and exclude sites in β-strands.

Can α-helices be cleaved efficiently? Although my study of caspase-3 indicated that this protease likely does not cleave signaling substrates at α-helical sites during apoptosis, other proteases may not be so stringent. Indeed, the other protease we tested was Staphylococcal GluC, a secreted bacterial protease with little known about its physiological substrates. The protease's crystal structure shows a shallow and broad active site, more so than the narrow and deep active site of caspase-3. This suggests that substrates with α-helical sites may dock directly in the active site of GluC and get cleaved efficiently. However, without a kinetic baseline determined empirically from natural substrates, it is difficult to assess whether GluC is likely to cleave α-helical substrates *in vivo*.

Dynamic α-helices. In my analysis of cleavage-site structures, sites were categorized into three distinct groups: loops, α-helices, or β-strands. Yet, this minimalist interpretation does not account well for the various conformations that I identified in the crystal and NMR structures of *E. coli* substrates. This approach also ignores the dynamic nature of protein structures; α-helices and loops often fluctuate between the two states. These unstable regions of proteins have been implicated as frequent sites of protein-protein interaction. Therefore, it is not unlikely that proteases also interact with and cleave substrates at these sites. Herein lies the difficulty: how does one tease apart the influence of structure on

cleavage if the structure is not fixed? Is the protease cleaving the α-helical conformation as well as the loop, or only the loop?

The first step towards answering these questions is to show unequivocally that at least some α-helices are cleaved in helical states. Peptide substrates are poor candidates for this inquiry, because even though "helical" peptides can be designed and synthesized, they are not fixed structured, but tend to be helical, or more helical than an unstructured peptide. Therefore, rigid folded protein substrates with stabilizing interactions to maintain their α-helical cleavage-sites should be the preferred test subjects. In this way, the rigidity of the cleavage-site helix can be measured by circular dichroism or hydrogen-deuterium exchange in the presence of a catalytically inactive protease harboring an active site mutation. If these rigid α-helices substrates are cleaved as helices as the data so far suggests, then a few critical questions arise: are helical substrates involved in proteolytic signaling, and if so are helical substrates cleaved with kinetics comparable to that of signaling substrates with cleavage-sites in flexible loops? Helical substrates also beg the question: do some proteases prefer α-helical cleavage-sites than loops? I speculate that the architecture of each protease's active site will determine its structural permissiveness, such that broad and open active sites like Staphylococcal GluC can accommodate both α-helices or loops, while more recessed active sites will strongly prefer loops over α-helices like caspase-3 does. However, it will be interesting to find an example of a protease that prefers α-helices to loops. For a protease to bind α-helices well while

excluding loops, I envision the active site could be dynamically occluded by flexible loops that productively orient upon helical substrate binding, while substrates with loops would fail to properly orient the active site into its competent conformation.

Protease activity in space and time. With these advances in protease substrate identification on a proteome-wide scale now available, the next challenge is to apply these techniques to answer specific questions regarding the spatiotemporal aspects of proteolysis, and which proteases are cleaving the observed substrates. These are particularly difficult questions to answer due to the complexity of the biological systems that proteases participate in. Cellular settings generally contain hundreds of diverse proteases with varying mechanisms of activation and regulation. In addition, these systems nearly always contain multiple proteases with overlapping substrate specificities and substrate repertoires. Thus, just identifying a cleaved substrate generally is not sufficient evidence to implicate a given protease as the causative agent without further evidence for its activation.

Efficient experimentation. Successful approaches to these questions will likely employ time course studies in conjunction with protease inhibitors. In this way a single control at time zero suffices to compare later time points to, as well as inhibitor treated samples. This minimization in the number of samples reduces

the investment of resources to reasonable levels, while providing the necessary controls to discriminate changes between samples.

Understanding the functional consequences of cleavage. While these experiments aimed at identifying specific protease substrates cleaved *in vivo* are essential, they are just the first steps to understanding the molecular mechanisms of how proteases perform their essential functions in biological and pathological processes. Further studies are required to reveal how cleavage of a substrate at a particular site functionally changes that protein in such a way as to realize a given outcome. Yet, the advances in substrate determination and cleavage-site identification presented here will hopefully empower other scientists in the protease field to demonstrate the molecular mechanisms that proteases utilize to conduct their diverse and crucial functions.