GETTING INTIMATE WITH TRYPSIN, THE LEADING PROTEASE IN PROTEOMICS

Elien Vandermarliere,[1,2] Michael Mueller,[3] and Lennart Martens[1,2]*
[1]*Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium*
[2]*Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium*
[3]*EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK*

Received 20 November 2012; revised 15 February 2013; accepted 15 February 2013

Published online 15 June 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/mas.21376

Nowadays, mass spectrometry-based proteomics is carried out primarily in a bottom-up fashion, with peptides obtained after proteolytic digest of a whole proteome lysate as the primary analytes instead of the proteins themselves. This experimental setup crucially relies on a protease to digest an abundant and complex protein mixture into a far more complex peptide mixture. Full knowledge of the working mechanism and specificity of the used proteases is therefore crucial, both for the digestion step itself as well as for the downstream identification and quantification of the (fragmentation) mass spectra acquired for the peptides in the mixture. Targeted protein analysis through selected reaction monitoring, a relative newcomer in the specific field of mass spectrometry-based proteomics, even requires a priori understanding of protease behavior for the proteins of interest. Because of the rapidly increasing popularity of proteomics as an analytical tool in the life sciences, there is now a renewed demand for detailed knowledge on trypsin, the workhorse protease in proteomics. This review addresses this need and provides an overview on the structure and working mechanism of trypsin, followed by a critical analysis of its cleavage behavior, typically simply accepted to occur exclusively yet consistently after Arg and Lys, unless they are followed by a Pro. In this context, shortcomings in our ability to understand and predict the behavior of trypsin will be highlighted, along with the downstream implications. Furthermore, an analysis is carried out on the inherent shortcomings of trypsin with regard to whole proteome analysis, and alternative approaches will be presented that can alleviate these issues. Finally, some reflections on the future of trypsin as the workhorse protease in mass spectrometry-based proteomics will be provided. © 2013 Wiley Periodicals, Inc. Mass Spec Rev 32:453–465, 2013

Keywords: *trypsin; protein cleavage; peptides; missed cleavage; proteomics*

I. INTRODUCTION

Trypsin is currently the most-frequently used protease in mass spectrometry-based proteomics experiments. It owes this position to its rather stringent cleavage specificity: it cleaves carboxy-terminal (C-terminal) of Arg and Lys. In human proteins, this cleavage pattern results in an average of 61 peptides per protein as calculated by DBToolkit (Martens, Vandekerckhove, & Gevaert, 2005). The peptides are composed of 1–2,374 residues, with an average of 9 and a standard deviation of 15, a length that is suitable for analysis by the mass spectrometer. When one miscleavage is allowed, these numbers change to 120 peptides per protein. In this situation, the peptides are composed of 1–2,433 residues, but with an average of 14 and a standard deviation of 20. Furthermore, the presence of a C-terminal basic residue enhances (positive) ionization, again beneficial to mass spectrometrical analysis. A final, related benefit is provided by the balance in basicity between the free amine of the amino-terminus (*N*-terminus) of the peptide, and the C-terminal Arg or Lys. This balance typically leads to good fragmentation, as understood through the mobile-proton hypothesis (Tabb et al., 2004). Yet, it has been repeatedly observed empirically that cleavage does not occur after all Arg or Lys residues. Several studies have therefore evaluated the cleavage specificity of trypsin, to result in a set of rules of which the best known is the failure to cleave at an Arg or Lys followed by a Pro. Some of these rules are implemented in commonly used peptide-to-spectrum matching algorithms to generate an accurate set of theoretical peptides. These theoretical peptides are fragmented in silico by the algorithm to yield theoretical fragmentation spectra. Identification is achieved through matching theoretical and experimental fragmentation spectra (Vaudel, Sickmann, & Martens, 2012). An accurate understanding of trypsin cleavage has even more value in situations where the quantification of peptides is attempted. Indeed, if a particular peptide can be split between a correctly cleaved variant that falls within the analytical range of a mass spectrometer, and a miscleaved variant that is too long to analyze, then the quantification of that peptide will rely exclusively on the observations of the correctly cleaved peptide, and underestimate the actual occurrence of the precursor protein as a result. In the case of targeted proteomics approaches, where specific peptides are a priori selected as representative of a given protein of interest, it would also be highly beneficial if the cleavage of these peptides could be assessed before the choice is made. A thorough understanding of trypsin, along with an appreciation of the remaining gaps in our knowledge of this key protease, is crucial to any researcher interested in the use of proteomics

Michael Mueller's present address is Clinical Genome Informatics Facility, Faculty of Medicine, Imperial College London, London, W12 0NN, UK
Contract grant sponsor: IWT O&O "Kinase Switch" Project; Contract grant sponsor: Ghent University; Contract grant sponsor: European Union 7th Framework Program; Contract grant number: 262067, 260558; Contract grant sponsor: EMBL International PhD program.
**Correspondence to*: Lennart Martens, Department of Medical Protein Research, Universiteit Gent—VIB. A. Baertsoenkaai 3 B-9000 Ghent, Belgium. E-mail: lennart.martens@vib-ugent.be

as an analytical tool. In the following sections, an overview of the working mechanism and specificity of trypsin is therefore presented, along with a brief history of the evolution of this enzyme into a heavily engineered workhorse. Furthermore, the accumulated knowledge on the finer details of its cleavage preferences will be described, with critical notes on the contradictions that can be found in these findings. An analysis is also presented of the shortcomings of trypsin for specific tasks, along with alternatives that can overcome these limitations. Finally, an outlook on the future of trypsin in the field of mass spectrometry-based proteomics and beyond will be given.

II. THE OCCURRENCE OF TRYPSIN IN NATURE

Trypsin is one of the most-important digestive proteases in most vertebrates. It is produced in the acinar cells of the pancreas and is released via the pancreatic duct into the gastrointestinal tract. After activation by enterokinase or autocatalysis, it cleaves—together with chymotrypsin and elastase—dietary proteins into peptides. The resultant peptides are further digested by a variety of exopeptidases and the small peptic fragments are subsequently absorbed by the intestine (Whitcomb & Lowe, 2007).

In 1876, Keuhne discovered trypsin in the pancreatic fluid of the bovine intestine during his studies on the intermediate digestion of proteins in the gastrointestinal tract. Shortly after, he found that trypsin was initially inactive but was subsequently converted to its active form (Keuhne, 1877). This latter observation formed the basis of the concept that zymogen activation proceeds by proteolytic processing. Almost a century after this discovery, in 1964, trypsin was one of the first proteins that had its full amino acid sequence determined. At the same time, the sequence of chymotrypsin became available and comparison of both sequences demonstrated their extensive similarity (Walsh & Neurath, 1964). Only later did it become clear that both protein sequences contained the same experimental error; in trypsin, Asp 177, which is located in the primary substrate-binding pocket, was identified as Asn, whereas in chymotrypsin, Asp 102, which is part of the charge-relay system, was also reported as Asn. This erroneous identification was a common sequencing error at the time because there was no good method available yet to discriminate between a free acid and its amide (Neurath, 1994).

Although the main function of trypsin is the degradation of digestive proteins, it can also activate the insulin receptor by limited cleavage of its extracellular domain to result in an insulin-like response in the absence of insulin (Clark et al., 1991). Apart from its direct role as a protease, trypsin can also serve as an indirect signaling molecule that regulates the cell by cleaving and hence triggering proteinase-activated receptors (PARs). PARs are a family of G protein-coupled receptors, which are activated by different proteases: PAR-1 and PAR-3 are activated by thrombin (Vu et al., 1991), whereas PAR-2 is activated by trypsin or mast cell tryptase (Nystedt et al., 1994). The protease cleaves the extracellular *N*-terminus of the PAR. The released peptide subsequently activates the cleaved receptor by binding to it. PAR-2, which is activated by trypsin via cleavage of the trypsin cleavage site SKGR SLIGR, is located at the apical membrane of enterocytes (Déry et al., 1998). Apart from its involvement in inflammation, little is currently known about the physiological and pathophysiological role of PAR-2 (Steinhoff et al., 2000).

A. Trypsin Amongst the Serine Proteases

Trypsin belongs to the group of serine proteases that emerged as one of the most-abundant and functionally diverse group of proteins during evolution. Serine proteases are found in all kingdoms of life and have been identified in viral genomes (Page & Di Cera, 2008). Serine proteases owe their name to the nucleophilic Ser, which is part of the active site. They are grouped into thirteen clans based on their catalytic mechanism, and are subsequently divided into 40 families on the basis of common ancestry (Rawlings & Barrett, 1994; Hedstrom, 2002). A clan is identified by two letters: the first represents the catalytic residue, which all of the families included in the clan have in common. For trypsin, this first letter is P and not S as one would intuitively think because the clan to which trypsin belongs contains families with different catalytic residues. The families are grouped by catalytic type, which is represented by another two-letter denomination: the first letter is S (Ser) for trypsin. This grouping with the respective group members can be found in the MEROPS database (Rawlings, Barrett, & Bateman, 2012).

Trypsin belongs to the clan PA proteases and is the representative member. Members of this clan, such as chymotrypsin, have the trypsin fold and participate in many processes such as digestion, coagulation, and immunity. Within this clan, which contains twelve families, trypsin is found in the S1 family; the S refers to the catalytic Ser. This family is further composed of two subfamilies: S1A and S1B. The S1B proteases are all intracellular enzymes, whereas S1A proteases are extracellular enzymes. Trypsin, which is secreted into the gastrointestinal tract, is a member of the S1A subfamily (Di Cera, 2009).

B. The Structure of Trypsin

The trypsin fold is characterized by a two beta-barrel structure; by two six-stranded beta-barrels where the beta-strand topology obeys the classic Greek-key architecture. The catalytic residues are located at the interface between these barrels. His 57 and Asp 102 originate from the *N*-terminal beta-barrel, whereas Ser 195 originates from the C-terminal barrel. Figure 1 gives a cartoon representation of the trypsin fold. The structure is stabilized by six completely conserved disulfide bridges to withstand the extracellular reducing environment of the gastrointestinal tract (Halfon & Craik, 1998). Trypsin shares this fold with other pancreatic proteinases such as chymotrypsin (Bode, Fehlhammer, & Huber, 1976).

Trypsin and its zymogen trypsinogen were both amongst the first proteins that had their structure determined (Bode & Schwager, 1975; Bode, Fehlhammer, & Huber, 1976; Fehlhammer, Bode, & Huber, 1977). This knowledge of structure allowed an early comparison between both enzymes. The transition from trypsinogen to trypsin, catalyzed by either enterokinase (Kunitz, 1939) or trypsin itself (Kay & Kassell, 1971), involves some chain rearrangements that do not influence the overall fold. During this activation step, the bond between Lys 15 and Ile 16 is cleaved. The new *N*-terminus, which was originally unstructured and floating in the solvent,

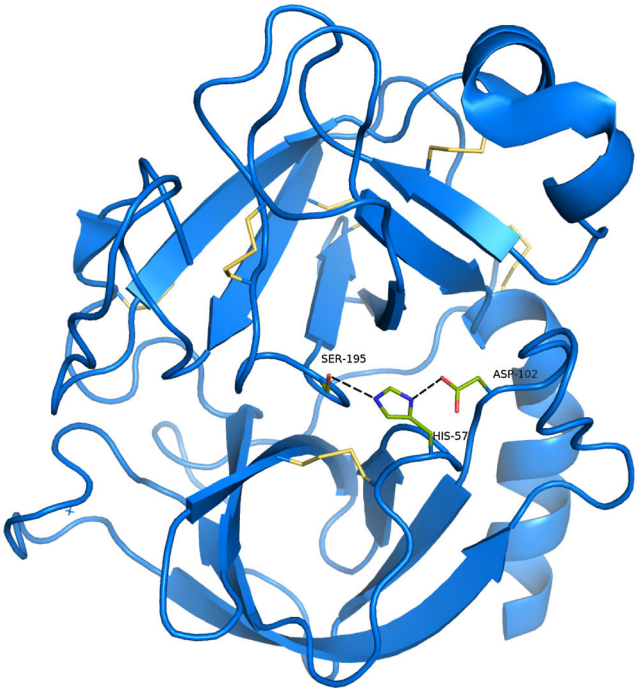

FIGURE 1. Cartoon representation of trypsin. This cartoon nicely illustrates the two beta-barrel structure of trypsin. The three residues of the charge-relay system are shown in green. The black dashes represent the connections between the catalytic residues. The conserved disulfide bridges are shown in yellow. PDB-entry 2PTN of bovine trypsin was used to create this figure (Walter et al., 1982).

subsequently forms a salt bridge with Asp 194, which is located next to the catalytic Ser (Oppenheimer, Labouesse, & Hess, 1966). At the same time, the residues that surround the active site change from their flexible state in trypsinogen, to an ordered state in trypsin (Fehlhammer, Bode, & Huber, 1977). This disorder-to-order transition results in a competent active site and hence active trypsin.

Trypsin, chymotrypsin, and elastase bear the same fold but have different specificity: the S1 pocket of trypsin is specific for Arg and Lys, whereas the S1 pocket of chymotrypsin is specific for the large aromatic residues Phe, Tyr, and Trp. This difference originates from the residues that line the bottom of this pocket: trypsin has an Asp at the bottom to allow the formation of a salt bridge with Arg or Lys; chymotrypsin has a Ser at the bottom of its S1 pocket that changes the specificity towards aromatic residues. Elastase is specific for small aliphatic residues in its S1 pocket due to the presence of Val 216 and Thr 226, which line the wall of the S1 pocket to prevent large residues from entering. In trypsin and chymotrypsin, both these residues are Gly, as illustrated in Figure 2 (Shotton & Watson, 1970).

C. The Reaction Mechanism of Trypsin

The reaction mechanism of serine proteases catalyzes the hydrolysis of a peptide bond in two steps as illustrated in Figure 3. In the first step, the nucleophilic Ser can attack the substrate scissile bond by donating a proton to the catalytic His, where the corresponding nitrogen has increased electronegativity due to the hydrogen bridge formed by the catalytic Asp. This attack results in formation of a covalent acyl-enzyme intermediate and release of the *N*-terminal fragment of the cleaved peptide bond. In the second step, a water molecule is activated by the His and attacks the acyl-enzyme intermediate. This attack is followed by release of the C-terminal fragment of the peptide bond. This latter release is achieved by the return of the proton from the histidine to the catalytic serine to reset the enzyme for another cycle (Hedstrom, 2002; Voet & Voet, 2004).

This mechanism of catalysis is often referred to as the charge-relay system (Blow, Birktoft, & Hartley, 1969), and has three residues at its base, which form the catalytic triad: Ser 195, His 57, and Asp 102. This conventional numbering scheme follows the residue numbering of chymotrypsin. Within this catalytic triad, the Ser functions as a nucleophile to initiate the reaction, whereas the His has a dual function. In the first step (acylation step), the catalytic His first acts as a general base to accept a hydrogen atom from the hydroxyl group of the catalytic Ser, which promotes the formation of the tetrahedral intermediate, and then the catalytic His acts as an acid by donating the acquired proton group to the amine that leaves when the tetrahedral intermediate collapses. Similar to the first step, in the second step (deacylation step), the catalytic His acts as a base to

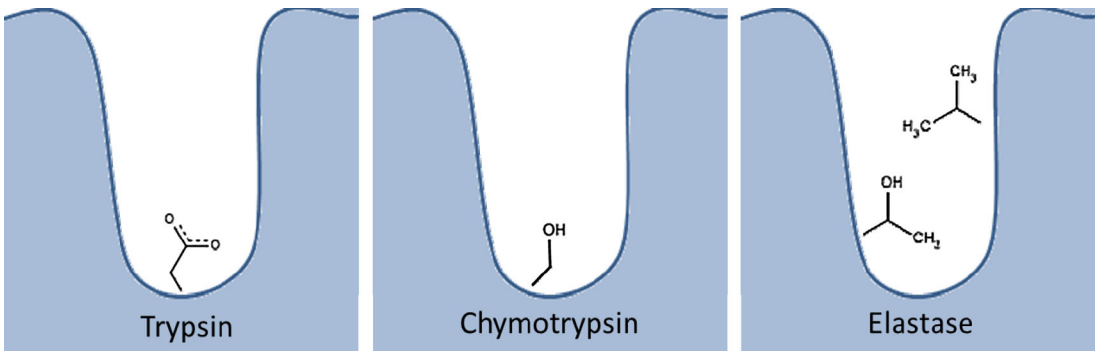

FIGURE 2. Schematic representation of the specificity of the S1 pocket of trypsin, chymotrypsin and elastase. The residues lining the S1 pocket determine the specificity of the different S1A proteases. Trypsin has an Asp at the bottom of this pocket that forms a salt bridge with Arg or Lys while chymotrypsin has a Ser at this position accounting for the preference of aromatic residues. Elastase on the other hand, has a Thr and Val lining the wall of the S1 pocket, hence, only small aliphatic residues are able to bind the S1 pocket.

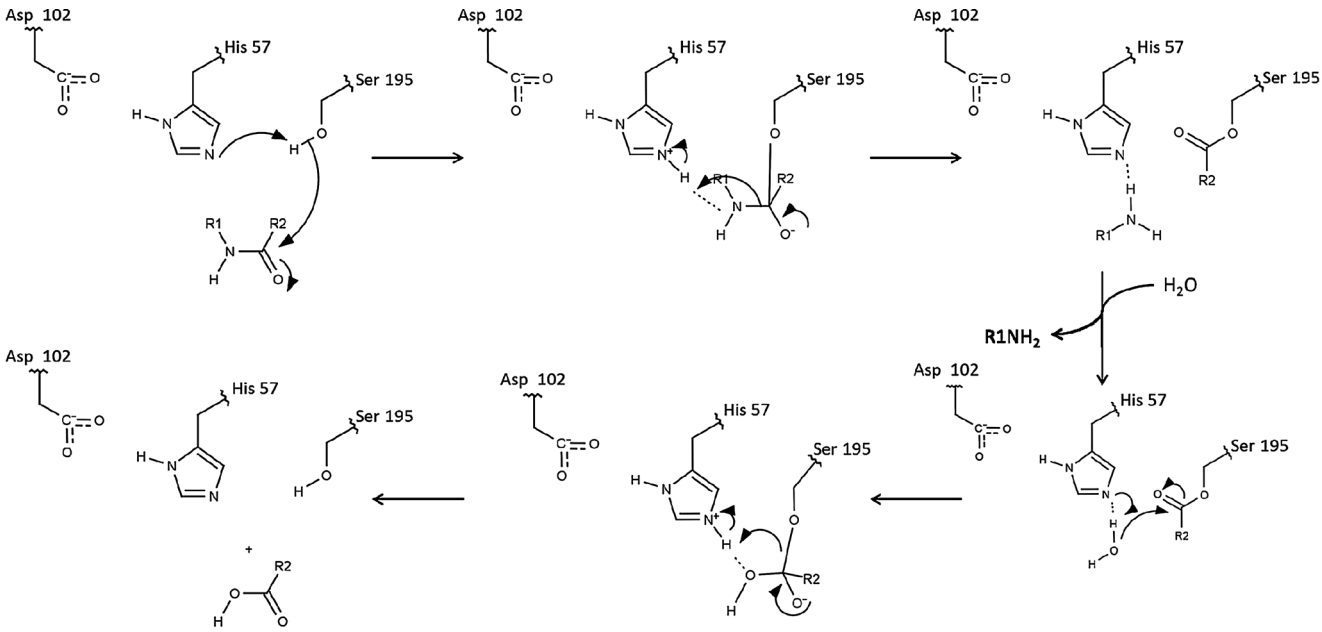

FIGURE 3. The reaction mechanism of trypsin. The arrows illustrate the flow of electrons at each of the different steps in the reaction mechanism of trypsin. In the first step, an acyl-enzyme intermediate is formed and the *N*-terminal part of the cleaved peptide bond is released. Next, an H_2O molecule activated by His attacks the acyl-enzyme intermediate which results in the release of the C-terminal part of the cleaved peptide bond.

activate the water molecule, and then acts as an acid and donates the proton to the Ser hydroxyl oxygen that leaves. The Asp is necessary to correctly orient the His to form a hydrogen bond with it. This orientation facilitates the abstraction of the proton from Ser and generates a potent nucleophile (Sprang et al., 1987). This charge-relay system exists in all other related serine proteases such as chymotrypsin (Freer et al., 1970) and elastase (Shotton & Watson, 1970), but also in bacterial proteinases such as subtilisin (Kraut, 1971). Other serine proteases have a variation on the classic catalytic triad where the Ser/His/Asp is replaced by Ser/His/Glu, Ser/His/His, or Ser/Glu/Asp (Ekici, Paetzel, & Dalbey, 2008). Still other serine proteases use a simpler catalytic mechanism composed of a dyad system where a Lys or His assists the catalytic Ser (Paetzel & Strynadka, 1999). Even though trypsin is arguably the most-studied enzyme, its catalytic mechanism is a generally accepted mechanism, which is still actively debated (Hedstrom, 2002; Polgár, 2005).

Experiments in which wild type trypsin hydrolysis of beta-casein is carried out at a pH of 7.0, show that the apparent rate constant for the hydrolysis of Arg-X (3.9×10^4 c^{-1}) is larger than the apparent rate constant for the hydrolysis of Lys-X (0.5×10^4 c^{-1}). This difference in apparent rate constant indicates that trypsin has a certain preference for Arg (Vorob'ev et al., 2000). Analysis of the pK_a value at the different steps of the catalytic mechanism revealed a pK_a value of 6.5, which corresponds to the deprotonation of the His. A pK_a value of 9.8 (Arg-X) and 10.5 (Lys-X) corresponds to the ionization process that can be attributed to the protonation of the amino group of Ile 26 which interacts with Asp 194. This stabilizes the active conformation of trypsin (Vorob'ev et al., 2000).

The nomenclature of the substrate-binding subsites (also referred to as specificity pockets, proposed by Schechter & Berger, 1967) is commonly used as the convention nomenclature. In most proteolytic enzymes, the active site is surrounded by several subsites that each accommodates one amino acid residue of the peptide substrate. The subsites occupied with the residues located *N*-terminally to the cleavage site on the substrate are referred to as S1, S2, etc. Likewise, the subsites on the C-terminal end of the cleavage side are referred to as S1', S2', etc. The cleaved peptide bond is the one between the residues in subsites S1 and S1'. The positions of the residues in the substrate peptide are referred to as P2, P1, P1', P2', etc. in analogy to the subsites they occupy. The numbering in a given peptide thus depends on the bond that is split (Schechter & Berger, 1967). Figure 4 provides a schematic illustration of this numbering convention.

III. THE USE OF TRYPSIN IN PROTEOMICS

Proteases are used extensively in protein chemistry, proteomics, and biopharmaceutical manufacture. Amongst their tasks are protein identification by peptide mass fingerprinting, protein

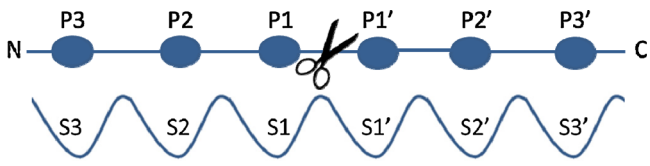

FIGURE 4. Schematic representation of the specificity pocket numbering convention in proteases. Conventionally, the nomenclature of Schechter and Berger is used to address the specificity pockets. Residues and specificity pockets located at the *N*-terminal side of the cleaved bond are referred to as P1, P2, … and S1, S2, … respectively. Likewise, residues and specificity pockets located at the C-terminal side of the cleaved bond are designated P1', P2', … and S1', S2', … respectively. The cleaved peptide bond is situated between the P1 and P1' positions in the peptide.

fragmentation, protein domain separation, and protein truncation. Because the focus of this review is the position of trypsin in proteomics, an overview of a typical shotgun experiment follows with emphasis on the steps where trypsin is involved.

The first step of a shotgun experiment involves isolation of the proteome of interest from a cell, tissue, or organelle. The protein mixture is denaturated and site-specifically cleaved into peptides with the aid of a protease, a role typically carried out with trypsin because of its favorable characteristics for mass spectrometry. The resulting, highly complex peptide mixture is separated with reversed-phase liquid chromatography (LC), which relies on the solubility of the peptides in an organic solvent compared to their affinity for a hydrophobic stationary phase (Berg, Tymoczko, & Stryer, 2002). After elution from the LC column, the peptides enter the mass spectrometer, where they are analyzed with tandem mass spectrometry (MS/MS) (Gevaert et al., 2007; Matallana-Surget, Leroy, & Wattiez, 2010; Walther & Mann, 2010). Nowadays, the LC and MS steps are coupled: the peptide mixture elutes directly off the column and into the mass spectrometer. The high degree of automation makes this experimental setup very suitable for high-throughput analysis of protein mixtures (Washburn, Wolters, & Yates, 2001). To identify the obtained fragmentation mass spectra, specialized algorithms called database search engines first generate theoretical spectra from in silico fragmentation of peptides obtained after in silico digestion of the protein sequence database of the sample (Frewen & MacCoss, 2007). To achieve maximal efficiency, this step ideally builds upon full knowledge of the specificity of the protease to allow optimal prediction of cleaved peptides. Finally, the measured spectra, which contain pairs of mass-to-charge (m/z) ratios and ion intensities, are matched against theoretically derived spectra to identify the parent peptide and infer the precursor protein with the aid of search algorithms such as Mascot (Perkins et al., 1999), OMSSA (Geer et al., 2004), and X!Tandem (Craig & Beavis, 2004). The outcome of the algorithm is a hit list based on scores that reflect the degree of resemblance between the observed and theoretical spectra (Sadygov, Cociorva, & Yates, 2004). To estimate the false discovery rate (FDR) of such a peptide or protein hit list, a decoy database can be used. The FDR is defined as the frequency at which spectra are matched against an entry from the decoy database at a given score threshold (Käll et al., 2008). Decoy databases can be obtained in various ways, but the most popular methods consist of a simple reversal of the original database or a shuffle of the original database (Feng, Naiman, & Cooper, 2007).

A. Trypsin as a Tool in Proteomics

Over the years, trypsin was, together with ovalbumin and hemoglobin, often among one of the first enzymes on which a new technique was tested. Trypsin owes this pioneering position to its ready availability. In the early years, protein purification was based on fractional precipitation with ammonium sulfate to produce a crystalline protein, and trypsin was one of the first proteins to be purified on an industrial scale via this method (Neurath, 1994).

In mass spectrometry-based proteomics, trypsin is by far the most popular protease. Trypsin owes this position to its high efficiency, cleavage-site specificity, and the fact that the tryptic peptides are easily amenable to mass spectrometry-based analysis. The relation between a theoretical tryptic digest and the observed peptides via mass spectrometry is illustrated in Figure 5 which represents the probability histograms of the peptide-length distribution of a trypsin digest of the human complement of UniProtKB/SwissProt (Jain et al., 2009; Uniprot Consortium, 2011) and the peptide-length distribution found from shotgun proteomics identifications from human samples, as obtained from the PRIDE database (Martens et al., 2005a; Vizcaíno et al., 2009). The in silico trypsin digest of the human proteome was performed with DBToolkit (Martens, Vandekerckhove, & Gevaert, 2005), which allowed for one missed cleavage. It is clear that the overall trypsin digest falls within the range of the mass spectrometer apart from very small and very large peptides. Finally, the presence of a basic residue at the C-terminus of the peptide leads to more-efficient positive ionization, and in concert with the basic nature of the *N*-terminus of the peptide, to overall favorable fragmentation characteristics.

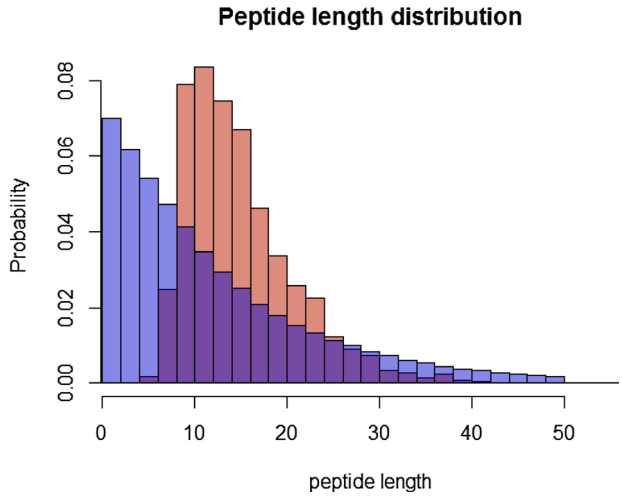

FIGURE 5. Superposition of the probability histograms of the peptide length distribution of a trypsin digest of the human complement of UniProt/SwissProt and the peptide length distribution found from shotgun proteomics identifications from human samples obtained from the PRIDE database. The peptide length distribution of the trypsin digest is colored in blue, the peptide length distribution from the peptides originating from the PRIDE database is colored in red. For clarity, the histogram only contains information about peptides up to a maximum length of 50 residues.

Trypsin plays a central role in two distinct but essentially parallel steps of a typical mass spectrometry-based proteomics experiment. Trypsin is used to digest the protein mixture of interest into a complex tryptic peptide mixture, but it is also mimicked in silico to perform the digest of the protein sequence database of the sample that is analyzed. As such, the in silico workflow aims to replicate faithfully the in vitro workflow. This in vitro workflow typically starts with solubilization and denaturation of proteins in the mixture by adding detergents such as sodium dodecyl sulfate (SDS) or sodium deoxycholate (SDC) and chaotropic agents such as urea or guanidinium hydrochloride. Typically, dithiotreitol (DTT) is added to reduce disulfide bridges, followed by alkylation of the reduced Cys with iodoacetic acid (IAA) or iodoacetamide (IAN). Finally, trypsin is added at a 1:30–1:50 trypsin: substrate molar ratio, sometimes complemented with 2 mM $CaCl_2$. The resulting mixture is incubated overnight at 37°C with shaking, typically

for approximately 16 hr, to obtain full digestion of the protein mixture (Klammer & MacCoss, 2006; Hervey, Strader, & Hurst, 2007). It is worth noting however, that digest protocols can vary substantially between labs.

The ideal denaturation and digestion protocol is one where complete digestion of all proteins in the sample takes place, because complete digestion provides the maximal possible number of observable peptides. Usually however, the digestion efficiency varies from protein to protein, and even from peptide to peptide to result in variations in the tryptic peptides. This variability in digestion efficiency affects the amount of peptide that is available for detection, and has an adverse effect on the reproducibility of analyses (Proc et al., 2010). To gain insight in the origins of this variability, a study of the different digestion protocols was performed by Proc et al. (2010) to determine the conditions in which the highest possible percentage of the proteins is digested in the most reproducible way. This study showed that digestion conditions are protein-dependent; there is no one common procedure that is optimal for all proteins. Therefore, alteration of the denaturation condition towards properties of the protein sample can improve digestion efficiency (Proc et al., 2010). Some proteins such as myoglobin and membrane proteins are difficult to digest. For these proteins, addition of SDC improves denaturation, solubilization, and subsequent trypsin digestion to result in improved identification and sequence coverage (Lin et al., 2008). RapiGest^{TM}, an acid-labile surfactant, is another reagent that improves solubilization and denaturation of proteins, while maintaining the activity of trypsin without an influence of its degradation products on MS analysis (Chen et al., 2007). The importance of the conservation of trypsin activity is in contrast with the common use of urea, which reduces trypsin activity and potential carbamylation of Lys and/or Cys. The result of RapiGest^{TM} thus resembles SDC, to lead to an increase in protein identifications by improving solubilization and denaturation of proteins without reduction of protease activity (Klammer & MacCoss, 2006). The traditional use of urea is to some extent attributable to the ease with which it can be removed from the peptide mixture (Gordon & Jencks, 1963; Greene & Pace, 1974). In combination with trypsin, however, urea can only be used at concentrations below 2 M to maintain sufficient trypsin activity; a similar effect is seen for guanidinium hydrochloride (Staes et al., 2004). When higher concentrations (6–8 M) of urea are needed to obtain sufficient protein denaturation, LysC should be the protease of choice because it is more stable to chaotropes (Raijmakers et al., 2010).

An ideal denaturation and digestion protocol should not only lead to complete digestion of a protein sample, but should also not interfere with LC-MS analysis. This latter motive provides an incentive to switch from SDS to RapiGest^{TM} because SDS is difficult to remove from the final peptide mixture (Wiśniewski et al., 2009) whereas RapiGest^{TM} does not seem to interfere with MS measurements (Chen et al., 2007) and can be left in the sample.

Because of the variety of ways in which digestion can be carried out, there is no single universal protocol available for protein digestion prior to a mass spectrometry. This lack of a universal protocol is not just because individual laboratories favor different, typically in-house protocols, but also because the best protocol for an experiment depends on the specifics of the sample, the properties of the proteins that should be analyzed, and the ultimate aim of the analysis. As a result, the literature abounds with protocols for digestion that are tailored to specific applications. However there is also an unmet need for a "gold standard," verified, simple yet efficient general purpose digestion protocol. The community would do well to invest some time and energy into development of such a protocol.

Beyond the influence from the denaturation and digestion protocol the efficiency of trypsin is also influenced by the brand of trypsin that is used. This was recently demonstrated by Burkhart and co-workers who analyzed the performance of six different trypsin preparations, retrieved from different vendors and belonging to different price ranges. They revealed considerable differences in protease performance which do not fully correlate with the price of trypsin (Burkhart et al., 2012). To meet a "gold standard" in proteomics, Burkhart and co-workers recommend some procedures as routine workflows for digestion and trypsin control. They propose monolithic column-based separations as a quality control for the digest. To evaluate the specificity of cleavage, SCX-enriched semi-tryptic fractions in conjunction with a three-step database search strategy are advised. The degree of missed cleavages and their reproducibility should then be evaluated by selected reaction monitoring (SRM) of selected peptide pairs (Burkhart et al., 2012).

B. The Evolution of Trypsin Into an Engineered Enzyme

Trypsin can be bought commercially from several companies, including Promega, Sigma-Aldrich, Roche, and Worthington Biochemical. The latter was among the first to provide crystalline trypsinogen. The application of trypsin in a proteomics experiment demands stringent specificity, and the autolysis of trypsin, which results in a broadened specificity, including chymotrypsin-like activity, therefore must be prevented as much as possible (Weil & Timasheff, 1966). This goal is achieved by modification of the native trypsin. The Lys residues are reductively methylated to yield a product that is highly active, stable, and most-importantly, resistant to autolysis (Rice, Means, & Brown, 1977). This resistance originates from the inability of trypsin to recognize and cleave methylated Lys (Joys & Kim, 1979; Poncz & Dearborn, 1983). Despite these alterations, autolysis of commercial trypsin can still occur; trypsin fragments can be used as internal mass calibrants (Harris, Janecki, & Reilly, 2002). This modified trypsin can further be treated with TPCK (*N*-alpha-tosyl-L-phenylalanyl chloromethyl ketone), an inhibitor for serine proteases with specificity for chymotrypsin and chymotrypsin-like proteases, to remove chymotryptic activity (Tobita & Folk, 1967; Rokhlin et al., 2004). Trypsin is subsequently purified via affinity chromatography and lyophilized to retrieve trypsin of high specificity and robustness.

At the end of the digestion phase, the activity of this heavily engineered trypsin must be stopped. This termination of activity is of particular importance for experiments where only limited proteolysis is required. This "shutdown" of trypsin can be obtained with acidification of the mixture [placing it on ice, and add formic acid (Smillie & Neurath, 1959) or trifluoroacetic acid]. Acidification is, however, not sufficient when the digest is used to label peptide C-termini with heavy oxygen (^{18}O) for quantitative analysis (a procedure carried out in $H_2{}^{18}O$)

(Desiderio & Kai, 1983), because residual tryptic activity will cause clearly noticeable back-exchange of the heavy isotopes with $H_2{}^{16}O$ once the digest is again transferred into unlabeled water, as shown by Staes et al. (2004). To completely terminate tryptic activity, Staes et al. reported the use of performic acid oxidation; however this acid caused chemical back-exchange of the heavy-labeled oxygen due to the very acidic pH. The final recommendation by these authors was to covalently modify the thiol groups of trypsin, and subject trypsin to 3 M guanidinium hydrochloride. At this concentration of the potent chaotrope, trypsin was sufficiently denatured to lose all activity, especially because the covalent modification of the trypsin thiol groups prevented (transient) reformation of the active structure of trypsin.

C. Current Knowledge on the Characteristics of Trypsin (Mis-)Cleavage

The reason behind the specificity for positively charged residues was elucidated early on during the co-crystallization of trypsin and an inhibitor, pancreatic trypsin inhibitor. Analysis of the structure of this complex revealed the salt bridge between the Lys from pancreatic trypsin inhibitor and the Asp 195 in the bottom of the S1 pocket of trypsin (Rühlmann et al., 1973). Contrary to expectation, trypsin cleavage is not always reproducible or predictable. Indeed, trypsin, like other proteases, can skip a seemingly cleavable residue. This skip results in a so-called miscleaved position. For trypsin, these miscleaved positions have been extensively analyzed to result in the Keil rules for miscleavage (summarized in Table 1; Keil, 1992). Some of these rules are also implemented in various database search engines to ensure more-accurate in silico cleavage of peptides. The best-known rule is the occurrence of a missed cleavage when Arg or Lys is followed by a Pro. This propensity for miscleavage is mechanistically explained by the atypical conformation imposed on the Arg or Lys by Pro (Olsen, Ong, & Mann, 2004). Another frequently occurring missed cleavage is found when two or more positively charged residues follow each other. In this situation, trypsin cleaves after the most C-terminal positively charged residue (Thiede et al., 2000). The last rule relates to the presence of negatively charged residues in close proximity to the Lys or Arg, a configuration that again promotes miscleavage. When the Asp or Glu are immediately next to the Arg or Lys, only one negatively charged residue is necessary to obtain miscleavage. When the acidic residue is located at position $j \pm 2$ or $j \pm 3$, at least one additional Asp or Glu must be nearby (see Table 1). These missed cleavages are most likely the result of salt bridges formed between the negatively charged residue and the Arg or Lys. The negative residues in close proximity to the Lys or Arg thus compete with the Asp at the bottom of the S1 pocket to competitively inhibit its interaction with the Arg or Lys (Thiede et al., 2000; Yen et al., 2006; Siepen et al., 2007).

The rules described so far all depend on residues in close proximity to the possible cleavage site; they all depend exclusively on sequence information. Hamady and co-workers introduced several structural parameters in their analysis of missed cleavages (Hamady et al., 2005). They focused on secondary structure, changes in secondary structure, and the exposed surface area of the residues that surround the cleavage site. They found that the average exposed surface area at the incorrect cleavage sites is greater than that of the correct cleavage sites; that result is at odds with the intuitive notion that residues buried inside the protein should be less accessible and therefore less readily cleaved. The authors propose that these miscleaved positions might occur in "nests" of acidic residues. This proposal is in agreement with the miscleavage rule that states that negatively charged residues that surround Arg or Lys result in miscleavage. When secondary structure is considered, Hamady et al. found that cleavage sites within unstructured regions are slightly more likely to be cleaved incorrectly, whereas cleavage sites within alpha-helices were slightly more likely to be cleaved correctly (Hamady et al., 2005).

Despite this established body of knowledge on the cleavage specifics of trypsin, it remains very difficult to accurately predict missed cleavage sites for a given protein. This difficulty primarily impacts targeted proteomics studies, where precise prediction of cleavage propensity can play a crucial role in the delineation of potential proteotypic peptides.

IV. THE FUTURE OF TRYPSIN IN PROTEOMICS

A. The Need for Expansion of Common Knowledge on Trypsin cleavage

Nowadays, mass spectrometry-based proteomics experiments are not merely dedicated to the characterization of protein mixtures. Indeed, they are more often performed to identify or quantify specific proteins or peptides at high throughput yet with high specificity. To propose optimal peptide targets, it is important to correctly predict the outcome of tryptic cleavage of the proteins of interest. The peptide to target must be cleaved reliably on both its tryptic termini in order for it to be routinely observable, and the presence of an unknown percentage of miscleaved variants will interfere with quantification. The currently known rules for missed cleavage cannot predict

TABLE 1. Summary of the Keil rules for miscleavage by trypsin

P4	P3	P2	P1	P1′	P2′	P3′
			[RK]	[P]		
			[RK]	[RK]		
		[DE]	[RK]			
			[RK]	[DE]		
			[RK]		[DE]	[DE]
[DE]	[DE]		[RK]			
	[DE]		[RK]		[DE]	

The different patterns of miscleavage are represented as regular expressions. Letters within a set of square brackets are considered to be only one symbol, equal to any one of the letters inside the brackets.

upwards of 10% of the observed missed cleavages. They also cannot stochastically predict cleavage or miscleavage (Thiede et al., 2000); a prediction that would be of great value because cleavage is rarely an all-or-nothing event; whereas some sites are always cleaved and other sites are never cleaved, many other sites are cleaved stochastically. The ability to predict stochastic cleavage would improve database search algorithms, aid more-reliable quantification, and provide the ability to select more accurate targets for SRM experiments.

Typical shotgun proteomics would perhaps not benefit dramatically from improved tryptic cleavage specificity because the current lack of precision can be circumvented by simultaneously considering one or more missed cleavages for each peptide in an in silico digest. Because all possibilities are covered, and whereas the search engine might waste some compute cycles, it will not lose information. The main impact can be expected in quantitative studies and targeted SRM analyses. The former will benefit because knowledge of stochastic miscleavage can help predict where a certain proportion of undetectable miscleaved variants of a peptide will cause an underestimation of the peptide quantification. Ideally, the reliability of the prediction would be such that a quantification shortfall can be corrected for. In the case of targeted proteomics with SRM, miscleavage prediction plays a crucial role. Indeed, the position of miscleavage and—ideally—the rate of miscleavage must be accurately predicted to correctly predict and quantify target peptides (Lange et al., 2008; Holman, Sims, & Eyers, 2012; Reker & Malmström, 2012).

Another application that requires full knowledge of the specificity of proteases is the use of accurate mass tags for protein identification in peptide mass fingerprinting. The current resolution of mass spectrometers allows such tags to be measured with very high mass accuracy to enable their use for high-throughput protein identification. The idea behind this approach is based on the observation that a particular proteome, if digested with a specific protease such as trypsin, will generate a peptide mixture in which most peptides can be uniquely classified based on their accurate mass and one or more additional parameters such as chromatographic retention times (Conrads et al., 2000; Smith et al., 2002; Pasa-Tolić et al., 2004).

Finally, a recent approach by the Yates group specifically aims to leverage the concentration-dependence of the speed of tryptic cleavage for any given protein, to reduce the footprint of highly abundant proteins in complex mixtures (Fonslow et al., 2013). The approach, called DeagDePr, relies on a partial initial digest of a complex proteome, in which highly abundant proteins are preferentially digested into peptides due to their faster digest kinetics, and a subsequent molecular-weight-cutoff filtration step is then used to remove these peptides. The remaining mixture is thus enriched for less abundant proteins that have escaped the initial digest. A downside of the method however, is that it requires more starting material (the published study used 10 times the normal protein mass) than a typical proteomics experiment.

B. Contradictions Within the Common Knowledge on Trypsin Cleavage

As stated above, the currently known rules to predict miscleavage cannot explain *ca.* 10% of observed miscleavages. Moreover, all these rules use only the local sequence as the information source. It is quite possible that the currently unexplained miscleaved positions are due to residual structure and local conformation. Whereas structural analysis of the specific substrates of trypsin and chymotrypsin showed that, for binding pockets S1, S2, S1', and S2', no structural constraints are placed on the residues that bind to these pockets; the S3 pocket does seem to enforce some structural constraints. The residue that binds this S3 pocket shows a sharp bend at its backbone, and the residues on each side of this bend are in a left-handed helical conformation that resembles the conformation of collagen. This bending results in the formation of a hydrogen bond between the main chain of trypsin and the main chain of the bound substrate. This conformation is consistent with the expectation that trypsin attacks the most-exposed parts of the molecular surface, which are mainly found in loops. From this analysis, it is clear that the specificity of the S3 pocket does not reside in the identity of a particular residue but in the presentation of hydrogen-bonding groups (Wright, 1977).

The general accepted rule of miscleavage is that trypsin cleaves next to Arg and Lys unless they are followed by a Pro. This rule was derived several decades ago and was based on a small number of experimentally confirmed cleavages. A more recent study re-examined this rule with the large data sets of available spectra. To limit the number of false discoveries, only doubly confirmed cleavages were retained in the data set. This analysis revealed that the presence of cleavages between (Arg/Lys) and Pro is not dramatically larger than the number of cleavages between (Arg/Lys) and Trp, and (Arg/Lys) and Cys, which are both generally accepted tryptic-cleavage sites. From these observations, Rodriguez and co-workers suggest that cleavage before Pro might also be a tryptic event but perhaps with less propensity for cleavage than other sites (Rodriguez et al., 2008) to highlight a shift from a deterministic non-cleavage to a stochastic rule.

Another generally accepted rule of miscleavage that can be questioned is the rule that trypsin does not cleave if the Arg or Lys is closely surrounded by negatively charged residues. This question is illustrated by a very salient example: trypsin can activate itself by autocatalysis when it cleaves the *N*-terminal peptide from trypsinogen. This autocatalysis involves cleavage of a Lys-Ile bond that is preceded by Val-Asp-Asp-Asp-Asp, and should have low cleavage propensity. Compared to the alternative preceding sequence of Val-Ala-Ala-Ala-Ala, the autocleavage indeed occurs at a lower rate that indicates that aspartyl residues do have a negative effect. Addition of high Ca^{2+} concentrations can reverse this effect. This reverse effect can be explained by the observation that Ca^{2+} binds to the aspartyl residues to shield their negative charges (Abita, Delaage, & Lazdunski, 1969). This slow autocatalysis is most likely a protection in case of an accidental activation. The stability of the zymogen system depends on a limit of the concentration of trypsin, which activates all pancreatic zymogens, to a minimum during intracellular transport, concentration and storage in the zymogen granule, and even for some time after extrusion into the pancreatic juice. Even small amounts of trypsin are sufficient to activate all available zymogens; in case of erroneous activation, acute pancreatitis might ensue (Hirota, Ohmuraya, & Baba, 2006; Sha, Ma, & Jha, 2009; Ji & Logsdon, 2011).

The Keil rule which states that miscleavage occurs when two or more successive positively charged residues follow each

other can also be questioned. Upon analysis of the human tryptic peptides available in PRIDE, we have found that 57% of the successive positively charged residues are correctly cleaved.

The above illustrates that, although implementation of these rules in search algorithms might reduce search time by limiting the search space, they can also result in wrongly identified peptides due to a mismatch between the peptides in the sample and the presumed peptides after in silico digest of proteins in the database. Obviously, improper use of these rules can also lead to a significant loss of identified peptides, whenever a cleaved peptide such as one that originates from an (Arg/Lys)-Pro cleavage, is not even considered by the search engine.

C. How to Overcome Shortcomings of Trypsin in MS Applications

Cleavage of a proteome with a protease such as trypsin results in many peptides of various lengths. For a mass spectrometry-based experiment, only peptides composed of 7–35 residues (corresponds roughly to a mass range of 600–4,000 Da) can typically be detected reliably due to limitations in the resolving power of the LC and MS instruments. Indeed, any one protease will only reveal certain parts of the proteome through the subset of detectable peptides that it generates. Other parts of the proteome will be cleaved into peptides that are either too small or too large to be detected, and leave this part of the proteome inaccessible to MS-based analysis. The limitations of one protease can be (partially) compensated for with another protease with different specificity in parallel on the same sample. Indeed, judicious use of different proteases in parallel within one experiment will result in a better sequence coverage and a higher number of identified proteins (Schlosser, Vanselow, & Kramer, 2005; Fischer & Poetsch, 2006; Swaney, Wenger, & Coon, 2010). Another drawback of the use of only one protease is that digestion conditions are protein-dependent. Hence, for a certain protein mixture, the protease of choice might not obtain its maximal proteolytic efficiency. Gatter and co-workers demonstrated that, to keep the number of missed cleavage in a trypsin digest low, an additional cleavage step with LysC can be very helpful (Glatter et al., 2012).

A particularly interesting analysis can also be carried out with regards to the resolution limitations of trypsin in a global human proteome analysis. With efforts to establish a human proteome project now underway (Hancock et al., 2011; Paik et al., 2012a,b), such an in silico analysis is of particular relevance, especially because there has been some debate about the relative merits and caveats of a gene-centric versus a protein (isoform) centric concept (Rabilloud, Hochstrasser, & Simpson, 2010). It is interesting to see from Figure 6 that the usefulness of parallel digests is relatively limited for a gene-centric proteome project, because trypsin alone yields unique, identifiable peptides that cover more than 98% of all human genes (top chart in Figure 6, dark circles). Use of two to six proteases in parallel does increase the coverage, but the gains are very small compared to the increased cost in complexity (second from bottom chart in Figure 6, open squares; complexity is counted in number of unique peptide sequences obtained, and is normalized to one for trypsin). The picture changes quite a bit when splice isoforms are taken into account because the maximal theoretical coverage by peptides unique to a single splice isoform drops to a little under 82% for trypsin (Fig. 6, top chart, open circles). As a result, parallel digests do yield important benefits here, with an optimal combination found at combination index 27 (overall *X*-axis) for the protease trio of Arg-C, Lys-C, and V8E at slightly more than 88% coverage. This optimal combination also clearly stands out in the cost-benefit chart (bottom chart in Figure 6, black circles; higher is better), where the cost in complexity is compared to the gain in coverage. Note that this combination essentially replaces a trypsin digest with parallel Arg-C and Lys-C digests to obtain longer and more uniquely mapping peptides and better coverage. If all six enzymes considered here were to be used in parallel (combination index 63 at the right-hand side of the chart), coverage of isoforms would be $<92\%$, to yield relatively little gain compared to the large increase in complexity, to result in one of the lowest cost-benefit scores for any combination of proteases. This chart makes it clear that unique identification of peptides for each known protein isoform in the human proteome is essentially impossible in theory, and that simple analysis of as many proteases in parallel as possible is clearly not the optimal strategy. It is wise to carefully plan experiments aimed to identification of a large amounts of protein isoforms uniquely, because achievement of this goal is already rendered surprisingly difficult by the digest alone. As such, a human proteome project is probably best planned as a gene-centric project at first, fostering research into improved methods and parallel digests to increase protein isoform coverage for a future expansion of the project.

D. The Role of Trypsin in the Future

In the trypsin applications described so far, the hydrolysis always takes place under denaturing conditions to expose and cleave as many peptide bonds as possible. Proteases such as trypsin can also be used with substrates in their native state, and it is the structure and not the sequence that will determine the sites of initial hydrolysis. The limited cleavage, also called limited proteolysis, is already used as a tool to investigate protein structure. The rationale of limited proteolysis is that peptide bonds buried in the core of the protein or located in secondary structure elements will be less accessible to the protease and are not cleaved as quickly as more accessible, surface-exposed residues. As such, limited digests of native proteins preferentially yields peptides located on the surface of the protein or in exposed loop regions (Fontana et al., 1997; Hubbard, 1998; Tsai et al., 2002).

Another future role for trypsin can be found in highly sensitive approaches to quantitative protein analysis that are based on a combination of proteolytic digest, antibody-based enrichment, and mass spectrometry. Protein cleavage followed by specific enrichment of proteotypic peptides with the aid of anti-peptide antibodies and subsequent identification of the captured peptides provides mass spectrometry with a central role in the field of protein biomarkers because even very small quantities of proteins can be measured with high sensitivity and absolute specificity (Meyer & Sthler, 2007; Alvarez-Llamas et al., 2008; Anderson et al., 2009; Poschmann et al., 2009). Finally, another specialized application of trypsin with an interesting application is so-called cell shaving, where trypsin is added to cultured cells to cleave off the extracellular parts of (trans-) membrane proteins (Tjalsma et al., 2008; Braconi

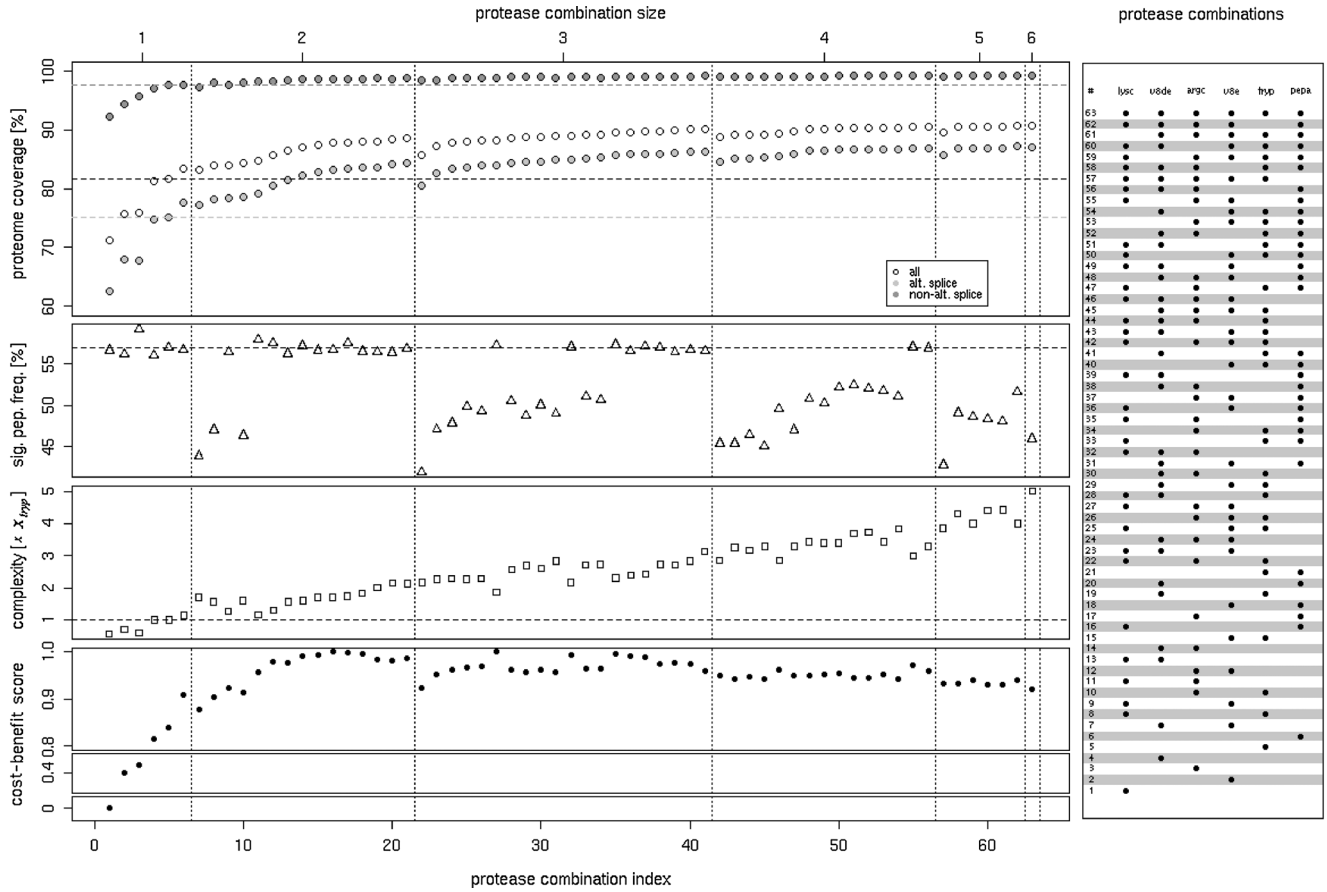

FIGURE 6. Cost-benefit analysis of proteases and protease combinations when analyzing the human proteome. The charts arranged from top to bottom are: (i) the proteome coverage for genes (dark circles), for genes with splice isoforms (light gray circles), and for all known proteins including splice isoforms (open circles); (ii) the signature peptide frequency, that is, the number of obtained peptides that are unique to a particular splice isoform; (iii) the complexity increase as measured by the total number of unique peptide sequences obtained; the number for trypsin is normalized to one; and (iv) the cost-benefit score (reflecting the normalized ratio of the coverage gain compared to trypsin over the complexity increase) for all 63 combinations of the six proteases considered. The protease combinations are numbered along the *X*-axis, and their composition can be read on the legend on the right.

et al., 2011; Vialás et al., 2012). This approach also allows the study of cellular interactions, as for instance found between a pathogen and a host cell (Dreisbach et al., 2011).

E. The Use of Trypsin Beyond Proteomics

Apart from being the main protease in proteomics, trypsin is also used in other biochemical and biomedical methods. One of these applications is its use in the isolation and culture of cells. Here, trypsin is used to cleave proteins that bind cells to a dish wall (Hentzer & Kobayasi, 1975; Kirkpatrick, Melzner, & Göller, 1985). A specific example is isolation and culture of mesenchymal stem cells from bone marrow, where trypsin is used to separate the different cell types (Soleimani & Nadri, 2009). Trypsin is also involved in the isolation of retinal endothelial cells, where it is used to remove pericyte contamination (Banumathi et al., 2009).

Apart from its use in biochemical methods, trypsin has also found its way into medicine as a means to remove blood clots (Grabe & Forsberg, 1986).

Trypsin also plays an important industrial role in the production of hydrolyzed cow milk formula for newborns (von Berg et al., 2003). This hydrolyzed cow milk is the preferred formula for newborns with cow-milk allergy; an allergy against bovine beta-lactoglobulin (Kattan, Cocco, & Järvinen, 2011). Hydrolysis of this protein with trypsin removes the epitopes of beta-lactoglobulin and prevents an allergic reaction (Ehn et al., 2005; Duan et al., 2012).

ACKNOWLEDGMENTS

The authors thank all PRIDE submitters for making their data publicly available. E.V. is supported by the IWT O&O "Kinase Switch" project, and L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"), and the PRIME-XS and ProteomeXchange projects funded by the European Union 7th Framework Program under grant agreement numbers 262067 and 260558, respectively. The work of M.M. was funded by the EMBL International PhD program.

REFERENCES

Abita JP, Delaage M, Lazdunski M. 1969. The mechanism of activation of trypsinogen. The role of the four N-terminal aspartyl residues. Eur J Biochem 8:314–324.

Alvarez-Llamas G, de la Cuesta F, Barderas MEG, Darde V, Padial LR, Vivanco F. 2008. Recent advances in atherosclerosis-based proteomics: New biomarkers and a future perspective. Expert Rev Proteomics 5:679–691.

Anderson NL, Anderson NG, Pearson TW, Borchers CH, Paulovich AG, Patterson SD, Gillette M, Aebersold R, Carr SA. 2009. A human proteome detection and quantitation project. Mol Cell Proteomics 8:883–886.

Banumathi E, Haribalaganesh R, Babu SSP, Kumar NS, Sangiliyandi G. 2009. High-yielding enzymatic method for isolation and culture of microvascular endothelial cells from bovine retinal blood vessels. Microvasc Res 77:377–381.

Berg J, Tymoczko J, Stryer L. 2002. Biochemistry, 5th edition. New York: WH Freeman and Company.

Blow DM, Birktoft JJ, Hartley BS. 1969. Role of a buried acid group in the mechanism of action of chymotrypsin. Nature 221:337–340.

Bode W, Schwager P. 1975. The refined crystal structure of bovine beta-trypsin at 1.8 a resolution. II. Crystallographic refinement, calcium binding site, benzamidine binding site and active site at PH 7.0. J Mol Biol 98:693–717.

Bode W, Fehlhammer H, Huber R. 1976. Crystal structure of bovine trypsinogen at 1–8 a resolution. I. Data collection, application of Patterson search techniques and preliminary structural interpretation. J Mol Biol 106:325–335.

Braconi D, Amato L, Bernardini G, Arena S, Orlandini M, Scaloni A, Santucci A. 2011. Surfome analysis of a wild-type wine *Saccharomyces cerevisiae* strain. Food Microbiol 28:1220–1230.

Burkhart JM, Schumbrutzki C, Wortelkamp S, Sickmann A, Zahedi RP. 2012. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. J Proteomics 75:1454–1462.

Chen EI, Cociorva D, Norris JL, Yates JR III. 2007. Optimization of mass spectrometry-compatible surfactants for shotgun proteomics. J Proteome Res 6:2529–2538.

Clark S, Eckardt G, Siddle K, Harrison LC. 1991. Changes in insulin-receptor structure associated with trypsin-induced activation of the receptor tyrosine kinase. Biochem J 276(Pt 1):27–33.

Conrads TP, Anderson GA, Veenstra TD, Pasa-Tolić L, Smith RD. 2000. Utility of accurate mass tags for proteome-wide protein identification. Anal Chem 72:3349–3354.

Craig R, Beavis RC. 2004. Tandem: Matching proteins with tandem mass spectra. Bioinformatics 20:1466–1467.

Déry O, Corvera CU, Steinhoff M, Bunnett NW. 1998. Proteinase-activated receptors: Novel mechanisms of signaling by serine proteases. Am J Physiol 274:C1429–C1452.

Desiderio DM, Kai M. 1983. Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. Biomed Mass Spectrom 10:471–479.

Di Cera E. 2009. Serine proteases. IUBMB Life 61:510–515.

Dreisbach A, van der Kooi-Pol MM, Otto A, Gronau K, Bonarius HPJ, Westra H, Groen H, Becher D, Hecker M, van Dijl JM. 2011. Surface shaving as a versatile tool to profile global interactions between human serum proteins and the staphylococcus aureus cell surface. Proteomics 11:2921–2930.

Duan C, Huo G, Yang L, Ren D, Chen J. 2012. Comparison of sensitization between beta-lactoglobulin and its hydrolysates. Asian Pac J Allergy Immunol 30:32–39.

Ehn B, Allmere T, Telemo E, Bengtsson U, Ekstrand B. 2005. Modification of IGE binding to beta-lactoglobulin by fermentation and proteolysis of cow's milk. J Agric Food Chem 53:3743–3748.

Ekici OD, Paetzel M, Dalbey RE. 2008. Unconventional serine proteases: Variations on the catalytic Ser/His/Asp triad configuration. Protein Sci 17:2023–2037.

Fehlhammer H, Bode W, Huber R. 1977. Crystal structure of bovine trypsinogen at 1–8 a resolution. II. Crystallographic refinement, refined crystal structure and comparison with bovine trypsin. J Mol Biol 111:415–438.

Feng J, Naiman DQ, Cooper B. 2007. Probability-based pattern recognition and statistical framework for randomization: Modeling tandem mass spectrum/peptide sequence false match frequencies. Bioinformatics 23:2210–2217.

Fischer F, Poetsch A. 2006. Protein cleavage strategies for an improved analysis of the membrane proteome. Proteome Sci 4:2.

Fonslow BR, Stein BD, Webb KJ, Xu T, Choi J, Park SK, Yates JR III. 2013. Digestion and depletion of abundant proteins improves proteomic coverage. Nat Methods 10:54–56.

Fontana A, Polverino de Laureto P, De Filippis V, Scaramella E, Zambonin M. 1997. Probing the partly folded states of proteins by limited proteolysis. Fold Des 2:R17–R26.

Freer ST, Kraut J, Robertus JD, Wright HT, Xuong NH. 1970. Chymotrypsinogen: 2.5-Angstrom crystal structure, comparison with alpha-chymotrypsin, and implications for zymogen activation. Biochemistry 9:1997–2009.

Frewen B, MacCoss MJ. 2007. Using bibliospec for creating and searching tandem ms peptide libraries. Curr Protoc Bioinformatics Chapter 13: Unit 13.7.

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. 2004. Open mass spectrometry search algorithm. J Proteome Res 3:958–964.

Gevaert K, Van Damme P, Ghesquière B, Impens F, Martens L, Helsens K, Vandekerckhove J. 2007. A la carte proteomics with an emphasis on gel-free techniques. Proteomics 7:2698–2718.

Glatter T, Ludwig C, Ahrne E, Aebersold R, Heck AJR, Schmidt A. 2012. Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. J Proteome Res 11:5145–5156.

Gordon JA, Jencks WP. 1963. The relationship of structure to the effectiveness of denaturing agents for proteins. Biochemistry 2:47–57.

Grabe M, Forsberg B. 1986. Retrograde trypsin instillation into the renal pelvis for the dissolution of obstructive blood clots. Eur Urol 12:69–70.

Greene RFJ, Pace CN. 1974. Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin. J Biol Chem 249:5388–5393.

Halfon S, Craik CS. 1998. Trypsin. In: Barrett AJ, Rawlings ND, Woessner JF, editors. Handbook of proteolytic enzymes. London: Academic Press. pp. 12–21.

Hamady M, Cheung THT, Tufo H, Knight R. 2005. Does protein structure influence trypsin miscleavage? Using structural properties to predict the behavior of related proteins. IEEE Eng Med Biol Mag 24:58–66.

Hancock W, Omenn G, Legrain P, Paik Y. 2011. Proteomics, human proteome project, and chromosomes. J Proteome Res 10:210.

Harris WA, Janecki DJ, Reilly JP. 2002. Use of matrix clusters and trypsin autolysis fragments as mass calibrants in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 16:1714–1722.

Hedstrom L. 2002. Serine protease mechanism and specificity. Chem Rev 102:4501–4524.

Hentzer B, Kobayasi T. 1975. Separation of human epidermal cells from fibroblasts in primary skin culture. Arch Dermatol Forsch 252:39–46.

Hervey WJ IV, Strader MB, Hurst GB. 2007. Comparison of digestion protocols for microgram quantities of enriched protein samples. J Proteome Res 6:3054–3061.

Hirota M, Ohmuraya M, Baba H. 2006. The role of trypsin, trypsin inhibitor, and trypsin receptor in the onset and aggravation of pancreatitis. J Gastroenterol 41:832–836.

Holman SW, Sims PFG, Eyers CE. 2012. The use of selected reaction monitoring in quantitative proteomics. Bioanalysis 4:1763–1786.

Hubbard SJ. 1998. The structural aspects of limited proteolysis of native proteins. Biochim Biophys Acta 1382:191–206.

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E. 2009. Infrastructure for the life sciences: Design and implementation of the UniProt website. BMC Bioinformatics 10:136.

Ji B, Logsdon CD. 2011. Digesting new information about the role of trypsin in pancreatitis. Gastroenterology 141:1972–1975.

Joys TM, Kim H. 1979. The susceptibility to tryptic hydrolysis of peptide bonds involving epsilon-*n*-methyllysine. Biochim Biophys Acta 581:360–362.

Käll L, Storey JD, MacCoss MJ, Noble WS. 2008. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 7:29–34.

Kattan JD, Cocco RR, Järvinen KM. 2011. Milk and soy allergy. Pediatr Clin North Am 58:407–426, x.

Kay J, Kassell B. 1971. The autoactivation of trypsinogen. J Biol Chem 246:6661–6665.

Keil B. 1992. Specificity of proteolysis. Berlin—Heidelberg—New York: Springer-Verlag.

Keuhne W. 1877. Uber das trypsin (enzym des pankreas). Verhanlungen des naturhistorisch-medicischen vereins zu Heidelberg, Vol. 1. pp. 194–198.

Kirkpatrick CJ, Melzner I, Göller T. 1985. Comparative effects of trypsin, collagenase and mechanical harvesting on cell membrane lipids studied in monolayer-cultured endothelial cells and a green monkey kidney cell line. Biochim Biophys Acta 846:120–126.

Klammer AA, MacCoss MJ. 2006. Effects of modified digestion schemes on the identification of proteins from complex mixtures. J Proteome Res 5:695–700.

Kraut J. 1971. The enzymes. New York and London: Academic Press. pp 165–183.

Kunitz M. 1939. Formation of trypsin from crystalline trypsinogen by means of enterokinase. J Gen Physiol 22:429–446.

Lange V, Picotti P, Domon B, Aebersold R. 2008. Selected reaction monitoring for quantitative proteomics: A tutorial. Mol Syst Biol 4:222.

Lin Y, Zhou J, Bi D, Chen P, Wang X, Liang S. 2008. Sodium-deoxycholate-assisted tryptic digestion and identification of proteolytically resistant proteins. Anal Biochem 377:259–266.

Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R. 2005. Pride: The proteomics identifications database. Proteomics 5:3537–3545.

Martens L, Vandekerckhove J, Gevaert K. 2005. DBToolkit: Processing protein databases for peptide-centric proteomics. Bioinformatics 21:3584–3585.

Matallana-Surget S, Leroy B, Wattiez R. 2010. Shotgun proteomics: Concept, key points and data mining. Expert Rev Proteomics 7:5–7.

Meyer HE, Stühler K. 2007. High-performance proteomics as a tool in biomarker discovery. Proteomics 7(Suppl. 1):18–26.

Neurath H. 1994. Proteolytic enzymes past and present: The second golden era. recollections, special section in honor of Max Perutz. Protein Sci 3:1734–1739.

Nystedt S, Emilsson K, Wahlestedt C, Sundelin J. 1994. Molecular cloning of a potential proteinase activated receptor. Proc Natl Acad Sci USA 91:9208–9212.

Olsen JV, Ong S, Mann M. 2004. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol Cell Proteomics 3:608–614.

Oppenheimer HL, Labouesse B, Hess GP. 1966. Implication of an ionizing group in the control of conformation and activity of chymotrypsin. J Biol Chem 241:2720–2730.

Paetzel M, Strynadka NC. 1999. Common protein architecture and binding sites in proteases utilizing a Ser/Lys dyad mechanism. Protein Sci 8:2533–2536.

Page MJ, Di Cera E. 2008. Evolution of peptidase diversity. J Biol Chem 283:30010–30014.

Paik Y, Jeong S, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee H, Na K, Choi E, Yan F, Zhang F, Zhang Y, Snyder M, Cheng Y, Chen R, Marko-Varga G, Deutsch EW, Kim H, Kwon J, Aebersold R, Bairoch A, Taylor AD, Kim KY, Lee E, Hochstrasser D, Legrain P, Hancock WS. 2012. The chromosome-centric human proteome project for cataloging proteins encoded in the genome. Nat Biotechnol 30:221–223.

Paik Y, Omenn GS, Uhlen M, Hanash S, Marko-Varga G, Aebersold R, Bairoch A, Yamamoto T, Legrain P, Lee H, Na K, Jeong S, He F, Binz P, Nishimura T, Keown P, Baker MS, Yoo JS, Garin J, Archakov A, Bergeron J, Salekdeh GH, Hancock WS. 2012. Standard guidelines for the chromosome-centric human proteome project. J Proteome Res 11:2005–2013.

Pasa-Tolić L, Masselon C, Barry RC, Shen Y, Smith RD. 2004. Proteomic analyses using an accurate mass and time tag strategy. Biotechniques 37:621–624, 626–633, 636 passim.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567.

Polgár L. 2005. The catalytic triad of serine peptidases. Cell Mol Life Sci 62:2161–2172.

Poncz L, Dearborn DG. 1983. The resistance to tryptic hydrolysis of peptide bonds adjacent to N epsilon, N-dimethyllysyl residues. J Biol Chem 258:1844–1850.

Poschmann G, Sitek B, Sipos B, Hamacher M, Vonend O, Meyer HE, Stühler K. 2009. Cell-based proteome analysis: The first stage in the pipeline for biomarker discovery. Biochim Biophys Acta 1794:1309–1316.

Proc JL, Kuzyk MA, Hardie DB, Yang J, Smith DS, Jackson AM, Parker CE, Borchers CH. 2010. A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin. J Proteome Res 9:5422–5437.

Rabilloud T, Hochstrasser D, Simpson RJ. 2010. Is a gene-centric human proteome project the best way for proteomics to serve biology? Proteomics 10:3067–3072.

Raijmakers R, Neerincx P, Mohammed S, Heck AJR. 2010. Cleavage specificities of the brother and sister proteases Lys-C and Lys-N. Chem Commun (Camb) 46:8827–8829.

Rawlings ND, Barrett AJ. 1994. Families of serine peptidases. Methods Enzymol 244:19–61.

Rawlings ND, Barrett AJ, Bateman A. 2012. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 40:D343–D350.

Reker D, Malmström L. 2012. Bioinformatic challenges in targeted proteomics. J Proteome Res 11:4393–4402.

Rice RH, Means GE, Brown WD. 1977. Stabilization of bovine trypsin by reductive methylation. Biochim Biophys Acta 492:316–321.

Rodriguez J, Gupta N, Smith RD, Pevzner PA. 2008. Does trypsin cut before proline? J Proteome Res 7:300–305.

Rokhlin OW, Guseva NV, Taghiyev AF, Glover RA, Cohen MB. 2004. Multiple effects of N-alpha-tosyl-L-phenylalanyl chloromethyl ketone (TPCK) on apoptotic pathways in human prostatic carcinoma cell lines. Cancer Biol Ther 3:761–768.

Rühlmann A, Kukla D, Schwager P, Bartels K, Huber R. 1973. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. Crystal structure determination and stereochemistry of the contact region. J Mol Biol 77:417–436.

Sadygov RG, Cociorva D, Yates JR III. 2004. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. Nat Methods 1:195–202.

Schechter I, Berger A. 1967. On the size of the active site in proteases. I. Papain. Biochem Biophys Res Commun 27:157–162.

Schlosser A, Vanselow JT, Kramer A. 2005. Mapping of phosphorylation sites by a multi-protease approach with specific phosphopeptide enrichment and NanoLC-Ms/Ms analysis. Anal Chem 77:5243–5250.

Sha H, Ma Q, Jha RK. 2009. Trypsin is the culprit of multiple organ injury with severe acute pancreatitis. Med Hypotheses 72:180–182.

Shotton DM, Watson HC. 1970. Three-dimensional structure of tosyl-elastase. Nature 225:811–816.

Siepen JA, Keevil E, Knight D, Hubbard SJ. 2007. Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. J Proteome Res 6:399–408.

Smillie LB, Neurath H. 1959. Reversible inactivation of trypsin by anhydrous formic acid. J Biol Chem 234:355–359.

Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, Conrads TP, Veenstra TD, Udseth HR. 2002. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. Proteomics 2:513–523.

Soleimani M, Nadri S. 2009. A protocol for isolation and culture of mesenchymal stem cells from mouse bone marrow. Nat Protoc 4:102–106.

Sprang S, Standing T, Fletterick RJ, Stroud RM, Finer-Moore J, Xuong NH, Hamlin R, Rutter WJ, Craik CS. 1987. The three-dimensional structure of Asn102 mutant of trypsin: Role of Asp102 in serine protease catalysis. Science 237:905–909.

Staes A, Demol H, Van Damme J, Martens L, Vandekerckhove J, Gevaert K. 2004. Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18. J Proteome Res 3:786–791.

Steinhoff M, Vergnolle N, Young SH, Tognetto M, Amadesi S, Ennes HS, Trevisani M, Hollenberg MD, Wallace JL, Caughey GH, Mitchell SE, Williams LM, Geppetti P, Mayer EA, Bunnett NW. 2000. Agonists of proteinase-activated receptor 2 induce inflammation by a neurogenic mechanism. Nat Med 6:151–158.

Swaney DL, Wenger CD, Coon JJ. 2010. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res 9:1323–1329.

Tabb DL, Huang Y, Wysocki VH, Yates JR III. 2004. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. Anal Chem 76:1243–1248.

Thiede B, Lamer S, Mattow J, Siejak F, Dimmler C, Rudel T, Jungblut PR. 2000. Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. Rapid Commun Mass Spectrom 14:496–502.

Tjalsma H, Lambooy L, Hermans PW, Swinkels DW. 2008. Shedding & shaving: Disclosure of proteomic expressions on a bacterial face. Proteomics 8:1415–1428.

Tobita T, Folk JE. 1967. Chymotrypsin c. 3. Sequence of amino acids around an essential histidine residue. Biochim Biophys Acta 147: 15–25.

Tsai C, Polverino de Laureto P, Fontana A, Nussinov R. 2002. Comparison of protein fragments identified by limited proteolysis and by computational cutting of proteins. Protein Sci 11:1753–1770.

Uniprot Consortium. 2011. Ongoing and future developments at the universal protein resource. Nucleic Acids Res 39:D214–D219.

Vaudel M, Sickmann A, Martens L. 2012. Current methods for global proteome identification. Expert Rev Proteomics 9:519–532.

Vialás V, Perumal P, Gutierrez D, Ximénez-Embún P, Nombela C, Gil C, Chaffin WL. 2012. Cell surface shaving of Candida albicans biofilms, hyphae, and yeast form cells. Proteomics 12:2331–2339.

Vizcaíno JA, Côté R, Reisinger F, Foster JM, Mueller M, Rameseder J, Hermjakob H, Martens L. 2009. A guide to the proteomics identifications database proteomics data repository. Proteomics 9:4276–4283.

Voet D, Voet J. 2004. Biochemistry. New York: John Wiley & Sons.

von Berg A, Koletzko S, Grübl A, Filipiak-Pittroff B, Wichmann H, Bauer CP, Reinhardt D, Berdel D. 2003. The effect of hydrolyzed cow's milk formula for allergy prevention in the first year of life: The German Infant Nutritional Intervention Study, a randomized double-blind trial. J Allergy Clin Immunol 111:533–540.

Vorob'ev MM, Dalgalarrondo M, Chobert J, Haertlé T. 2000. Kinetics of beta-casein hydrolysis by wild-type and engineered trypsin. Biopolymers 54:355–364.

Vu TK, Hung DT, Wheaton VI, Coughlin SR. 1991. Molecular cloning of a functional thrombin receptor reveals a novel proteolytic mechanism of receptor activation. Cell 64:1057–1068.

Walsh KA, Neurath H. 1964. Trypsinogen and chymotrypsinogen as homologous proteins. Proc Natl Acad Sci USA 52:884–889.

Walter J, Steigemann W, Singh T, Bartunik H, Bode W, Huber R. 1982. On the disordered activation domain in trypsinogen: Chemical labelling and low-temperature crystallography. Acta Cryst B38:1462–1472.

Walther TC, Mann M. 2010. Mass spectrometry-based proteomics in cell biology. J Cell Biol 190:491–500.

Washburn MP, Wolters D, Yates JR III. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19:242–247.

Weil L, Timasheff SN. 1966. The enzymic activity of trypsin autolysis products. Arch Biochem Biophys 116:252–254.

Whitcomb DC, Lowe ME. 2007. Human pancreatic digestive enzymes. Dig Dis Sci 52:1–17.

Wiśniewski JR, Zougman A, Nagaraj N, Mann M. 2009. Universal sample preparation method for proteome analysis. Nat Methods 6:359–362.

Wright HT. 1977. Secondary and conformational specificities of trypsin and chymotrypsin. Eur J Biochem 73:567–578.

Yen C, Russell S, Mendoza AM, Meyer-Arendt K, Sun S, Cios KJ, Ahn NG, Resing KA. 2006. Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: Protease cleavage and SCX elution rules from data mining of MS/MS spectra. Anal Chem 78:1071–1084.