The Pro in Proteomics

Saar van der Laarse

Cover design: Saar van der Laarse
Printed by Proefschriftmaken | proefschriftmaken.nl

The research in this thesis was performed in the Biomolecular Mass Spectrometry and Proteomics Group, Utrecht University, Utrecht, The Netherlands.

The Pro in Proteomics

De Pro in Proteomics
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 13 april 2022 des middags te 2.15 uur

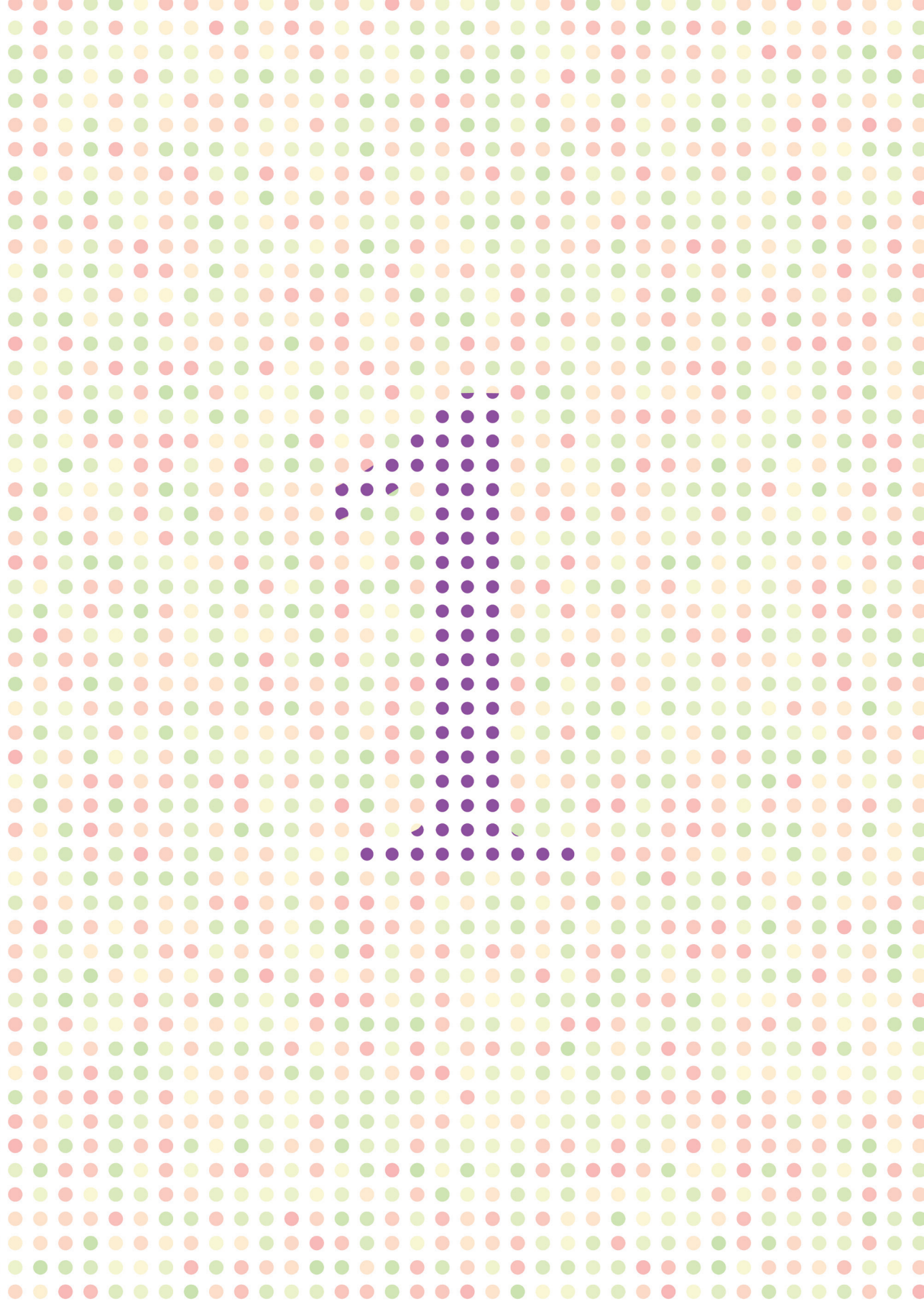
door

Saar Anna Margarita van der Laarse

geboren op 26 februari 1991
te Amsterdam

Promotor:
Prof. dr. A.J.R. Heck

Table of contents



Chapter 1

General Introduction

General introduction

Proteins play a central role in essential biological processes, ranging from the catalysis of biochemical reactions to the transport of nutrients, signal transduction and the maintenance of the cells mechanical structure[1]. In essence, each protein is a linear chain of amino acids, defining its primary structure. This chain is folded to give rise to a proteins' secondary and tertiary structure, characterized by the formation of key structural elements (i.e. alpha-helices and beta-sheets) and a fully structured protein, respectively. When a protein functions in a complex with others, or when it is composed of multiple, separate polypeptide chains, this complex is referred to as a proteins' quaternary structure[2].

When forming a proteins' primary structure, 20 different amino acids can be used for each position in the linear chain, see Figure 1. These amino acids all have their own distinct biochemical properties, which collectively form the properties of the protein by participating in enzymatic reactions or influencing the proteins' shape or charge at a certain position[4]. Two amino acids are joined together by a condensation reaction between the carboxylic acid of the first amino acid and the amine of the second. Through the release of a water molecule this gives rise to a peptidyl bond[2].

Each peptidyl bond can be stabilized by delocalization of the nitrogen lone pair towards the carbonyl oxygen, giving rise to two resonance structures: one with a double C=N bond, and one with the carbonyl intact (Figure 2A)[5]. This double bond can exist in two isomers, referred to as *cis* or *trans* (Figure 2B-D). These two isomers are interchangeable, which is known as *cis*/*trans* isomerization.

Chemically speaking, every combination of amino acids can be subjected to *cis*/*trans* isomerization. The more electron rich the nitrogen in the peptidyl bond, the stronger the double bond character is expected to be[6]. Since prolines are the only amino acids that have an electron-donating alkane substitution on their nitrogen, these amino acids are particularly restricted in their rotation leading to a high *cis* content in peptidyl bonds preceding prolines. When a protein is folded to produce its secondary and/or tertiary structure, the *cis*/*trans* isomerization of prolines is often the rate limiting step. To enhance this process many peptidyl prolyl isomerase enzymes (PPIases) are located around the ER to catalyze this isomerization step, helping newly synthesized proteins achieve their correct conformation[7].

Throughout a protein's lifetime, its shape is an important factor in the interactions it can achieve with other proteins and may also affect its activity. For example, accessibility of the active site of cyclin-dependent kinase (CDK) directly determines its activity. A modification on the kinase results in a conformational change, exposing the active site and hence activating the kinase[8]. Alternatively, a change in the conformation or shape of a protein

can affect the way two proteins fit together, resulting in interactions specific to a certain protein conformation[9,10].

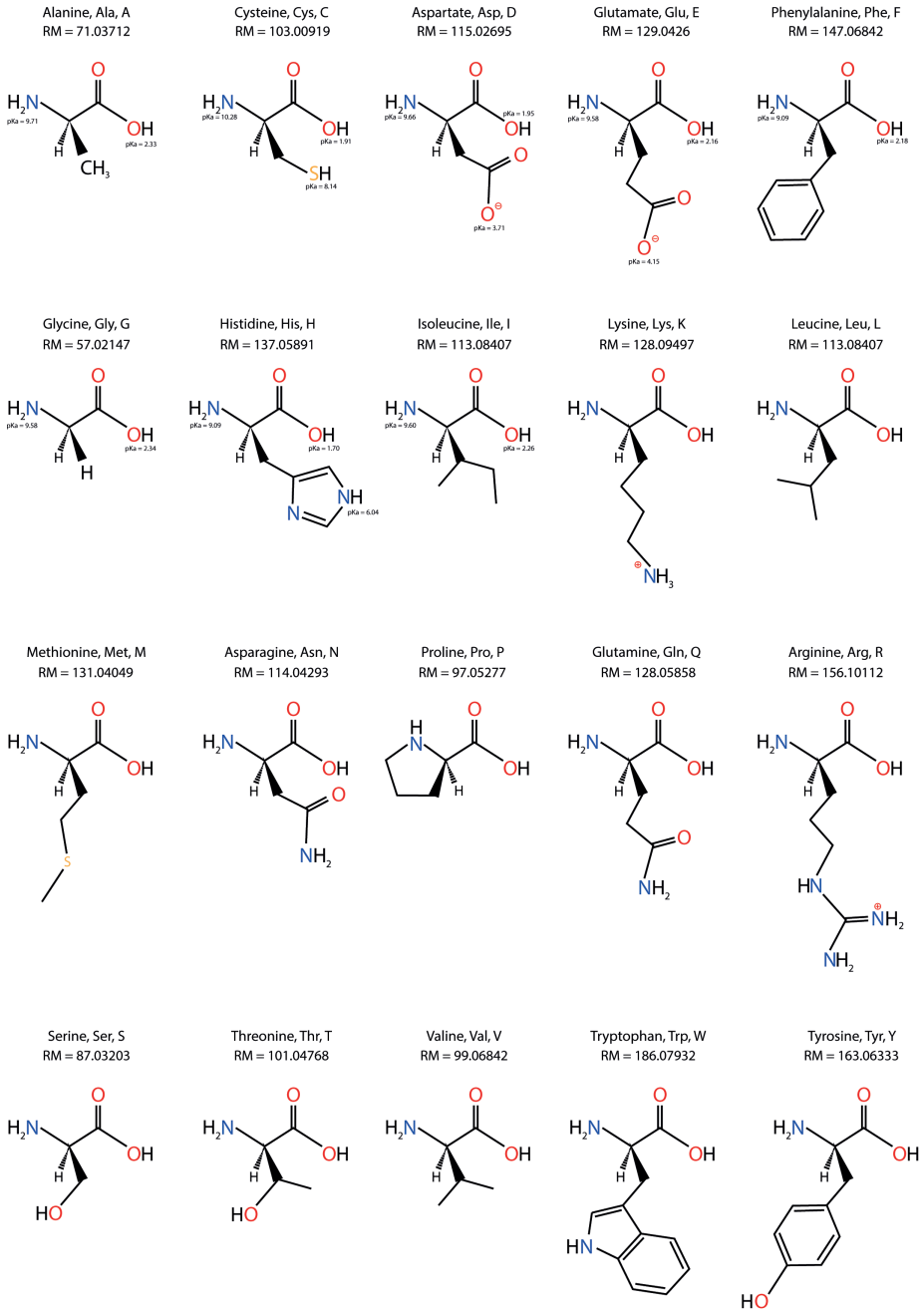


Figure 1 – Illustration of amino acid structures. Where applicable, pKa values are provided[3]. Charged side groups are depicted as they would be expected at pH = 7.4. RM indicates the residual mass of each amino acid, which is the actual mass minus the mass of a water molecule.

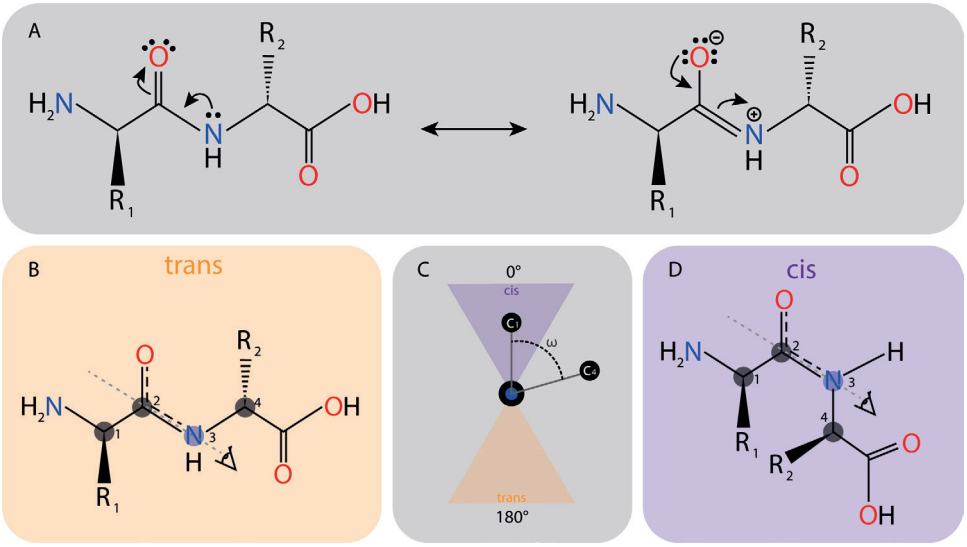


Figure 2 – Double bond character and *cis*/*trans* isomerization across a peptidyl bond. (A) Delocalization of electrons across the peptidyl bond reveal the origin of its partial double bond character. (B,D) Structure of *trans*- and *cis* isomers across the peptidyl bond, respectively. (C) A bond is termed *trans* when the angle between C_1 and C_4 viewed across the double bond (C_2=N_3) is approximately 180 degrees, and *cis* when this angle is around 0 degrees.

The human genome consists of about 20 000 different genes. In a too simplistic view, these genes lead to at least 20 000 different proteins, all involved in virtually every process occurring in our bodies[11]. To ensure that all these proteins perform the correct task at the appropriate time and in the right place, they are subjected to tight regulation[12,13]. This regulation can be achieved in several ways. Firstly, a proteins activity can be directly influenced, effectively turning a protein 'on' or 'off'. Secondly, a proteins location can be used to control its activity, by moving a protein to a certain site only when it is needed there. Thirdly, the interaction surface or conformation of a protein can be altered to render it incompatible or more compatible with certain interactors. Combinations of these regulatory mechanisms can occur and more often than not the modification of a protein and changes in its activity occur simultaneously, when a modification added or removed from a protein induces a conformational change[14].

The (reversible) modification of proteins after their expression and folding is known as post-translational modifications (PTMs) and these are a common mechanism to regulate protein structure, activity and subcellular localization[15]. To date, over 80,000 unique sites bearing PTMs have been reported, including acetylation, phosphorylation, glycosylation, methylation and ubiquitination[16]. Even though these modifications range from small additions (e.g. the addition of $-CH_3$ groups, or exchange of NH_2 for OH in deamidation) to the addition of one or more small proteins (e.g. ubiquitination

and SUMOylation), they are typically employed in a mechanistically similar way.

PTMs are usually dynamic. They can be added to proteins when needed, and usually can also be removed again when this need subsides. While present on a protein, they change the local environment there, which can be recognized (or prevent recognition) by other proteins. For example, Pin1, the enzyme catalyzing certain cis/trans isomerization reactions, recognizes its substrates via its WW-domain. Phosphorylation of Serine-16, located in the middle of this domain, prevents the binding of substrates and therefore reduces Pin1 activity while the phosphorylation is in place[17], as described further in Chapter 5.

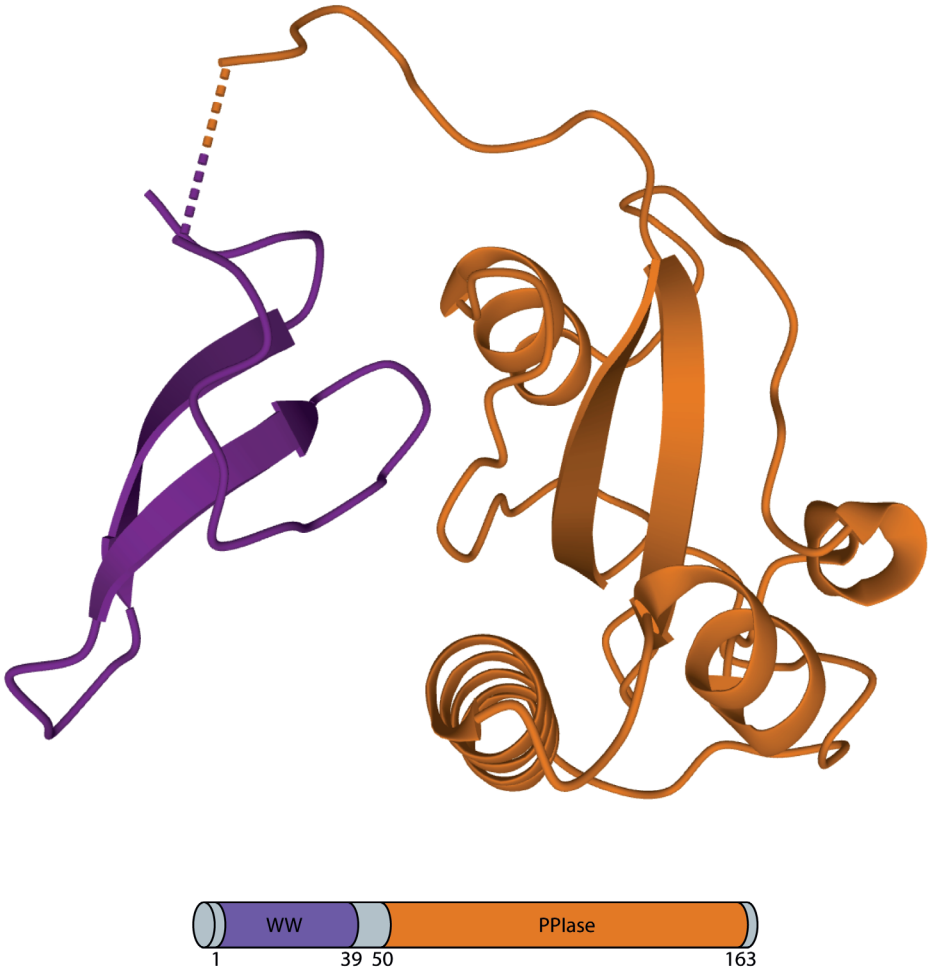


Figure 3 – Structural model of Pin1, featuring a WW-domain that binds Ser-Pro and Thr-Pro in a phosphorylation-dependent manner and a PPIase domain responsible for catalyzing its substrates *cis*/*trans* isomerization[22,26–28]. Figure adapted from [28].

The complexity of protein regulation is increased even further by the occurrence of combinations of PTMs[18]. Proteins carry a multitude of amino acids that can be modified and more than one PTM can be placed on a protein at any given time. These combinations can consist of multiple copies of the same type of PTMs or of two or more different types of PTMs. In addition to influencing the protein they were placed on, the changes induced by a PTM can also affect the placement of another PTM. This is termed PTM interplay or crosstalk[19].

An interesting kind of crosstalk is the direct interaction between phosphorylation and conformational change. This interaction can occur when a phosphorylation site is directly followed by a proline. The peptidyl bond between the serine, threonine or tyrosine and the proline is subject to *cis*/*trans* isomerization like any other amino acid pair. However, upon phosphorylation of the amino acid preceding the proline, the isomerization rate is drastically decreased, effectively locking the amino acid pair in either the *cis*- or *trans* conformation[6,20].

Once phosphorylated, the regular PPIases that catalyze isomerization of peptidyl bonds before prolines can no longer act, and isomerization is catalyzed solely by a PPIase called Pin1. In addition to the PPIase domain observed in regular PPIases, Pin1 also has a WW-domain which specifically binds serine proline and threonine proline motifs in a phosphorylation-dependent manner (Figure 3)[21–23]. Although the exact mechanism is still unclear, Pin1 is thought to efficiently catalyze the *trans*-to-*cis* isomerization[24,25].

An interesting kind of crosstalk is the direct interaction between phosphorylation and conformational change. This interaction can occur when a phosphorylation site is directly followed by a proline. The peptidyl bond between the serine, threonine or tyrosine and the proline is subject to *cis-trans* isomerization like any other amino acid pair. However, upon phosphorylation of the amino acid preceding the proline, the isomerization rate is drastically decreased, effectively locking the amino acid pair in either the *cis*- or *trans* conformation[6,20].

Once phosphorylated, the regular PPIases that catalyze isomerization of peptidyl bonds before prolines can no longer act, and isomerization is catalyzed solely by a PPIase called Pin1. In addition to the PPIase domain observed in regular PPIases, Pin1 also has a WW-domain which specifically binds serine proline and threonine proline motifs in a phosphorylation-dependent manner (Figure 3)[21–23]. Although the exact mechanism is still unclear, Pin1 is thought to efficiently catalyze the *trans*-to-*cis* isomerization[24,25].

The regulation of *cis*- or *trans* isomerization of proteins is interesting because these isomers of proteins each have their own biology. Several proteins were shown to act specifically on one of the two isomers and due to the large conformational change induced by *cis*/*trans* isomerization, many more – if

not all – proteins are expected to be isomer specific. For example, the proline-directed kinase ERK2 was shown to preferentially phosphorylate *trans*-substrates[9]. Similarly, Ssu72, a phosphatase that removes phosphorylation from Serine-5 in the C-terminal domain of RNA polymerase 2, was shown to be specific for the *cis* Serine-5 Proline-6 isomer[29].

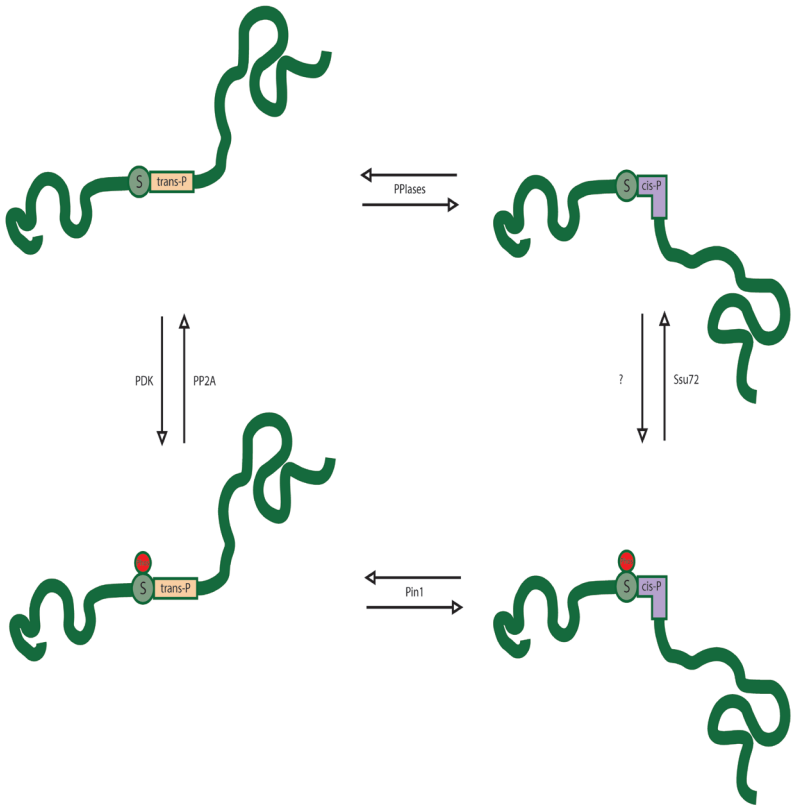


Figure 4 – Illustrative representation of the interplay between phosphorylation and protein conformation change. While unphosphorylated, many PPIases exist to catalyze *cis*/*trans* isomerization. After phosphorylation, only Pin1 has the affinity to catalyze the serine-proline or threonine-proline motif. Phosphorylation can be added or removed by isomer specific kinases and phosphatases, such as proline-dependent kinases (e.g. CDKs and MAPKs) and PP2A for the *trans* isomer and Ssu72 for the *cis* isomer[9,29,30].

Basically, this means that a protein is not just one species, but in this context can consist of four subsets of a protein which are connected through isomer-specific kinases and phosphatases and phosphorylation-specific PPIases (Figure 4). The abundance of each subset is controlled by these dualistic combinations of PTMs, allowing for a more sophisticated regulation of available binding partners, localization, activity and so on. Of course, these combinations of PTMs can exist for other PTMs as well and are not necessarily limited to binary combinations, further adding to the complexity of protein regulation.

The Mass Spectrometer

A valuable technique to study these PTMs, their combinations and their position on proteins, is mass spectrometry. A mass spectrometer (MS) is a specialized analytical instrument that can very accurately determine the mass-to-charge ratio of an ion. In its most basic form, it consists of an ion source, electronic components to guide and/or separate ions, known as analyzers, and a detector to determine the abundance of the ions[31].

The ion source forms the inlet of the MS. Here, your sample is ionized and – if not already gaseous – transferred into the gas phase. Many different ionization techniques exist, but for the analysis of proteins and peptides electrospray ionization (ESI) is typically the ionization method of choice[32]. During electrospray ionization, a solution containing your molecules and ions of interest is guided through a needle, see Figure 5A. Between the needle and the inlet, a potential difference is applied which brings protons to the edge of the droplets, generating a spray of fine droplets at the end of the needle. Because the ion source is heated, water molecules evaporate out of the droplets shrinking them even further. When the droplets become small enough, repulsion of protons in the droplet becomes big enough to overcome the surface tension of the droplet causing the droplet to burst apart. As a result, the compounds of interest are desolvated quite well and protons are attracted to electron rich areas on the compounds which aids the formation of ions[33,34].

From the ion source, ions are guided through the instrument while neutral species are pumped out of the MS[35]. The numbers and types of analyzers encountered between the ion source and the detector varies a lot across MS instruments and some instruments also combine analyzers and detectors to determine mass-to-charge and the corresponding intensity at the same time. Examples of mass analyzers are quadrupoles, ion traps, time-of-flight (TOF) and Orbitrap analyzers[32], see Figure 5. Other analyzers exist but are outside the scope of this thesis.

A quadrupole is basically an ion filter. It is made up out of four parallel placed circular or hyperbolic rods, see Figure 5C. On these rods, oscillating electric fields are applied. When a positively charged ion enters the quadrupole, it is attracted towards the rods with a negative potential and repelled by the rods with a positive potential, causing the ion to change trajectory towards the negatively charged rods. If the rods change potential sign before the ion hits one, the trajectory of the ion is changed towards the 'new' negative rods[36]. Using this principle, ions of a particular mass-to-charge ratio can be guided through the quadrupole center with a stabilized trajectory, while ions outside of this range are lost. Hence, the quadrupole acts as an ion filter, only allowing ions with a selected mass-to-charge ratio to pass.

An ion trap is similar to a quadrupole, except it also uses end caps to trap ions

in a three-dimensional space, see Figure 5B. To ensure that ions do not veer off course due to repulsion between the ions themselves, ion traps usually contain an inert gas to drain away energy via collisions (collisional cooling).

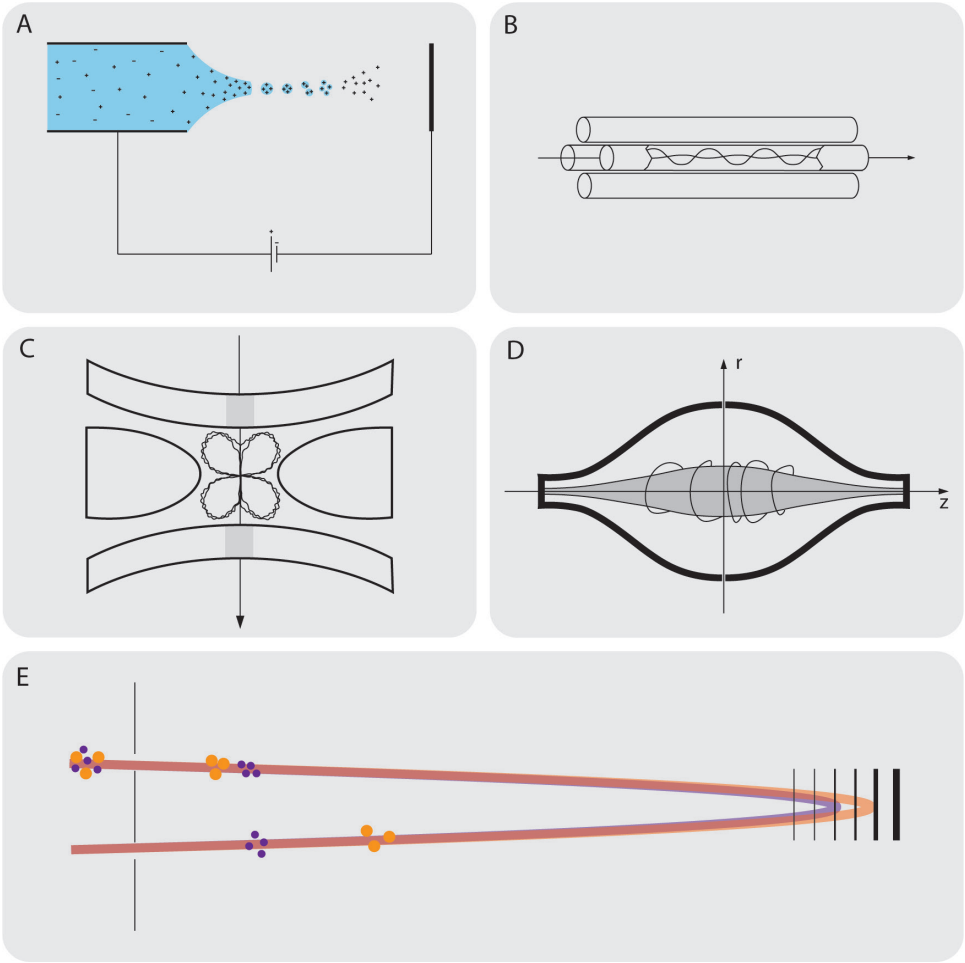


Figure 5 – Schematic representation of electrospray ionization (ESI) and the mass analyzers used in the work described in this thesis. A) During Electrospray Ionization a potential difference is set across a conductive capillary and the ion inlet of the mass spectrometer. When the voltage is high enough, small droplets are formed. When the droplets shrink, ion-ion repulsion overcomes the surface tension and the droplets explode, releasing ions into the gas phase. B) Simplified representation of a quadrupole, only ions with a stable trajectory will make it to the end of the quadrupole, others collide with the rods and never make it to the detector. C) Schematic of a three-dimensional ion trap, ion enter through the inlet end cap, oscillate in the trap and exit through the exit end cap towards the detector. D) Schematic of the orbitrap, ions rotate around the central spindle while oscillating in the z-direction. E) In time-of-flight detection ions are accelerated into a drift region of fixed distance. Since ions get the same momentum, their m/z determines the time they require to travel the distance. Reflectrons can be used to extend the drift region and refocus ions drifting apart due to repulsion between them. Adapted from [31].

By modifying the potential differences used, the ion trap can destabilize the trajectory of ions of a certain mass only, expelling them from the trap[37]. These ions can then be detected, resulting in a mass spectrum of a subset of the ions present in the ion trap.

Usually, ions ejected from an ion trap are detected or counted by an electron multiplier. Ions emerging from an analyzer are accelerated into a plate, known as the conversion dynode. Collision with this plate release several particles, such as ions and electrons. When detecting positive ions, the conversion dynode will be set to a large negative potential. Upon the release of electrons, these will be repelled from the dynode and attracted to the next dynode, which is held at higher potential. Upon impact, more electrons are released. This cascade of electrons produces a current that can be measured, resulting in a signal for the detected ion[38].

Nowadays, a widely used mass analyzer is the Orbitrap. In an Orbitrap, ions are injected with some velocity into the space between the central spindle and the outer electrode, see Figure 5D. An electrostatic potential is applied on the central spindle, causing ions to oscillate around it. The frequency of these oscillations is not related to the kinetic energy of the ions entering the trap, but is dependent on their mass-to-charge ratio. The oscillating ions produce an image current, which can be detected. Using a Fourier-transform, this alternating image current information is translated to mass-to-charge and intensity, giving rise to the final mass spectrum[40].

Alternatively, TOF analyzers determine the mass-to-charge ratio of an ion by accelerating the ion using an electric field and subsequently measuring how long it takes the ion to travel a fixed distance, see Figure 5E. Similar to detection after an ion trap, the actual detection of ions emerging from a TOF is usually performed by an electron multiplier. The corresponding mass-to-charge ratio of the ions is computed from the electric field used to accelerate them and their resulting speed, i.e. the time the ions need to travel the distance to the detector[39].

When analyzing ions, several characteristics are of importance. For example, we would like to be able to analyze ions with maximum accuracy, meaning that the mass-to-charge value we determine is close to the actual mass-to-charge ratio of the ion. This accuracy is usually expressed as follows[41]:

$$\frac{\left| m_{experiment} - m_{exact} \right|}{m_{exact}} \times 10^{6} \text{ [ppm]}$$

This means that a molecule with a mass of 1000 Da at 10 ppm accuracy, the determined mass must lie between 1000.01 and 999.99 Da. For the mass analyzers we discussed here, the mass accuracy is best for the orbitrap (typically < 5 ppm), closely followed by modern TOF instruments (10 ppm) and quadrupoles and ion traps (~100 ppm)[31]. Orbitrap analyzers also

champion in terms of mass resolution, which is very useful for instance when measuring a protein with various combinations of modifications. In terms of mass range, the TOFs have virtually no upper mass limit, however, measuring very big mass-to-charge ratio ions comes at the price of longer measurement times and reduced mass resolution[42]. Of course, the exact technical specifications can vary between instruments and manufacturers.

Tandem Mass Spectrometry

More complex mass spectrometers usually use combinations of analyzers to allow for more advanced experimental methods, such as the isolation of ions with a certain mass-to-charge (i.e. the precursor ions) ratio followed by induced fragmentation of these precursor ions. Measuring the mass-to-charge ratio of the resulting fragment ions gives extra information on the precursor. For example, when measuring peptides, fragmentation of a precursor peptide results in a series of ions fragmented between different amino acids, thereby giving information on the order of amino acids in, or the sequence of, the peptide[43].

Fragmenting ions requires the addition of energy to the ion, which has to be sufficient to break one (or more) chemical bonds. Over the years, many methods have been developed to add energy to an ion, including the use of collisions with inert gas molecules, irradiation with photons or via the capture of electrons[44]. Once an ion attains excess energy, this energy can be redistributed in several ways. Which way is most favorable will depend on the method of excitation and characteristics of the ion itself. In the ideal scenario, the added energy results in the breaking of a chemical bond, resulting in the release of one or more fragment ions. However, often the most favorable pathway is the release of a small, neutral molecule, which is known as neutral loss[45]. Neutral losses can be helpful as they can indicate certain modifications or moieties were present in the ion. For example, a neutral loss of 98 Da from a peptide is usually the result of the loss of phosphoric acid (H_3PO_4), indicative of the presence of phosphorylated serine or threonine residues[46–48]. Excited ions can also dissipate energy through bond vibrations or by release of photons, and combinations of these dissipation pathways usually occur.

In the laboratory where I performed my PhD research, we were capable of fragmenting ions in all aforementioned manners: by using collisions, photons or by creating radical species through capturing electrons. The most common method to fragment ions, not only in our lab, is by collisional activation, which is also known as collisional-induced dissociation (CID)[49,50]. Here, kinetic energy of an ion is transformed into internal energy by colliding them with neutral inert gas molecules, such as helium, nitrogen or argon. CID can be performed in an environment that can contain and refocus precursor and fragment ions, such as a quadrupole or an ion trap. When CID is performed

on an instrument with an orbitrap the term Higher-energy C-trap dissociation (HCD) is used. The collisional activation is performed in an ion trap (termed the HCD cell) and the resulting ions are cooled and stored in the C-trap prior to injection into the orbitrap for analysis. In the context of protein and peptide analysis, CID gives rise to fragment ions resulting from broken peptidyl bonds, also known as b- and y-ions (Figure 6)[51]. The technique is very useful for the production of ions series that can be used to determine the amino acid sequence of a peptide/protein, but tends to lead to a lot of small neutral losses when leaving groups are available, such as PTMs[52].

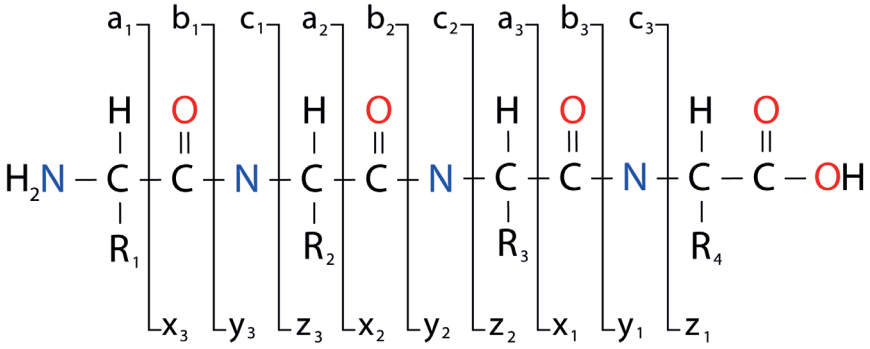


Figure 6 – Explanation of peptide fragmentation ion nomenclature. Fragment ions are assigned a letter to indicate whether they carry the N- or C-terminus of the polypeptide precursor as well as which bond broke to give rise to the fragment. The subscript refers to the number of amino acids present in the fragment ion, as described previously[51].

An alternative method to fragment ions is by generating a radical species. Since a radical has an unpaired electron, they tend to be unstable molecules which can be exploited in tandem mass spectrometry. Two fragmentation techniques that proceed via the production of radicals in the mass spectrometer are electron capture induced dissociation (ECD) and electron transfer induced dissociation (ETD). In ECD, free electrons are generated using a filament and cooled by collisions with an inert gas to reduce their energy. These electrons can be captured by positively charged proteins or peptides, neutralizing one of their charges and simultaneously introducing a radical species[53]. Due to the presence of the radical, cleavage of the N-C_α bond of amino acid backbones becomes favorable. Hence, ECD preferentially generates c- and z-ions in the resulting fragmentation spectrum from large, multiply charged biomolecules.

ETD works in a similar manner, except anions are used to transfer electrons towards the positively charged ions. Because anions with very low electron affinities are used, these anions readily transfer an electron to the peptide or protein cations when they get in close proximity. Because anions have far more mass than individual electrons, they can be present in ion trap-type surroundings far longer than electrons alone, allowing more time for

electrons to reach the cations[54]. After electron transfer, a radical species is formed $[M + nH](n - 1)^{+\bullet}$, which is often also observed in the resulting mass spectra[55]. When it decays, c- and z-ion fragments are formed. Similar to ECD, ETD is most efficient with highly positively charged precursor ions. The ETD technique requires a bit more time than CID/HCD fragmentation[44], but neutral loss of PTMs are a lot less common, enabling ECD/ETD to accurately pinpoint the position of PTMs on the peptides identified[54].

To reduce the amount of unfragmented radical ions observed in ETD, a little supplemental energy can be added to the ions by means of HCD. This hybrid fragmentation method was developed in the Hecklab and is termed EThcD[56]. The method allows for more efficient fragmentation, generating information rich spectra that contain next to the c- and z-ions, also b- and y-ions, often better maintaining the localization information of PTMs[57].

In addition to CID, HCD, ETD and EThcD, photon-based fragmentation techniques are evaluated for their capabilities in fragmenting peptides or proteins. In these methods, ions are irradiated with photons to increase their internal energy, hopefully resulting in fragmentation. Some examples are ultraviolet photodissociation (UVPD) and Infrared Multiple Photon Dissociation (IRMPD)[58,59]. As for all fragmentation methods, the challenge is to generate fragments that cover the whole sequence of the peptides or proteins under investigation, while maintaining ion origin information (i.e. creating recognizable fragments) and localization information of PTMs.

Studying proteins with mass spectrometry

The use of mass spectrometers to study proteins comes in many different flavors. Which instrument or method to use is often a trade off between time, detail and sample complexity. For example, we can investigate a single protein very detailed or aim to analyze all proteins expressed in a cell or tissue simultaneously.

An important distinction in the study of mass spectrometry is whether a protein is kept in its native fold during analysis, referred to as native MS, or if the protein is unfolded (denatured). Native MS is a very powerful technique to study protein complexes and determine their stoichiometry and assembly mechanisms[60]. Because proteins are still folded, a lot of basic residues are shielded from the buffer environment, which results in lower charges on the proteins after ionization. Because of the reduced number of charges, any given protein will have a higher mass-to-charge ratio in native MS than after unfolding. This is challenging for the electronics inside the MS, as higher mass-to-charge ions need to be transmitted efficiently through the instrument[61]. The higher mass-to-charge ratio also comes with a benefit: it leads to more spaced out peaks in the mass spectrum, making it easier to resolve complex PTM combinations or other biological variants of a certain

protein[62]. Since overlapping peaks make native MS spectra hard to interpret, samples are usually purified quite extensively prior to analysis. The method is often combined with direct infusion, as organic solvents or pH values far from the proteins normal environment tend to result in denatured proteins.

When a protein is denatured but otherwise intact, we speak of top-down MS approaches. Top-down MS can still give information on biological variants of proteins and combinations of PTMs occurring on the same protein copy, but any information on complexation or their stoichiometry is lost[63]. Since top-down MS is compatible with separation techniques such as chromatography, it can be performed on more complex samples. The main challenges here are to distinguish overlapping peak distributions and to fragment enough of a protein to be able to identify it accurately. Since intact proteins have many degrees of freedom, fragmenting bonds though an increase in internal energy tends to be more successful on the termini on the proteins than throughout the middle of the sequence[64,65].

To reduce the sample complexity and generate somewhat more manageable ions, proteins can also be cut in a few relatively large pieces. This way, quite a lot of potential interplay on PTMs can still be studied, while generating valuable fragment ions becomes easier, enabling elucidation of the protein sequence. This method is termed middle-down, and is very popular in the field of antibody analysis[66,67].

In proteomics, we often attempt to get information on as many proteins as possible from a given tissue or cell line. One of the major challenges to achieve that, is the enormous dynamic range and chemical diversity of the proteome[68,69]. To address these challenges, proteomics samples are usually digested with a protease to generate manageable peptides with somewhat similar characteristics. This approach is termed bottom-up MS[70]. The community preferred protease is Trypsin, which cleaves C-terminal to lysine and arginine residues[71]. This way, trypsin generates peptides very suitable for MS-based fragmentation, because they can carry protons on the C-terminal arginine or lysine and the N-terminal amine residue. Since both termini are charged, the ion series from either end of the peptide as produced by fragmentation can carry charge and thus be detected in the MS. For example, both the b- and complementary y-ion series of a given peptide might be observed upon HCD fragmentation[72,73].

Of course, generating peptides to give information on the parent proteins also has downsides. The protease chosen to perform the digestion and its cleavage specificities will largely determine which proteins are visible or invisible in your experiment[74]. The reduced size of peptides makes them easier to handle, however, due to their short length a lot of information on co-occurring PTMs is lost. Assuming the peptides of interest were fragmented properly, you can map PTMs back on the parent protein after the experiment, but unless PTMs were identified on the same peptide it is impossible to know

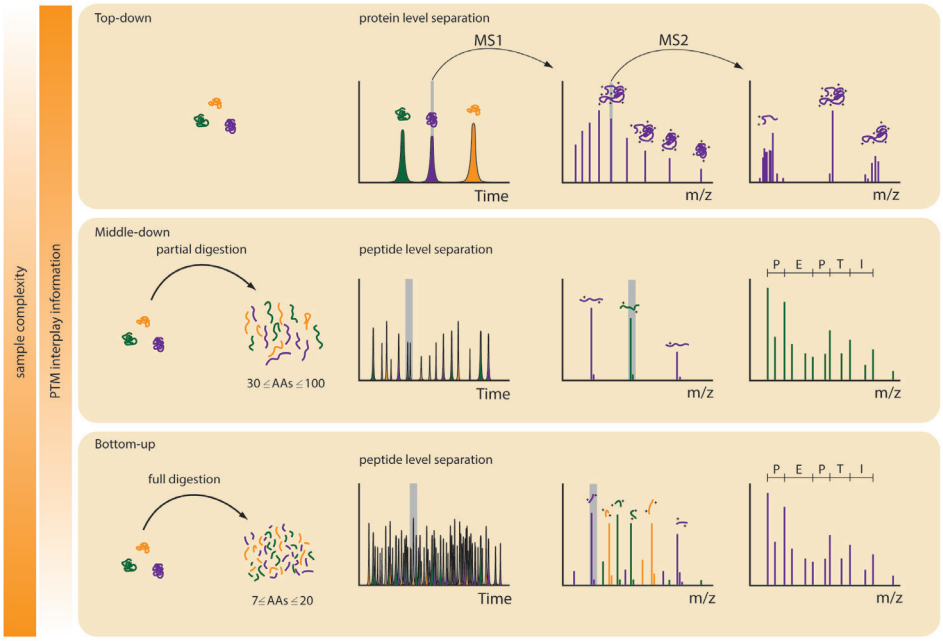


Figure 7 – Illustrative example of various kinds of denatured MS used in proteomics. In the top panel Top-down MS is shown, starting with one or several purified proteins subjected to protein level separation. MS1 reveals an envelope of charge states for a protein and in MS2 protein fragments can be recorded to help identifying the protein. In the bottom panel, bottom-up MS is illustrated, where one fully digests complex protein samples (or whole proteomes), separates them at the peptide level and typically selects the most abundant peptides in MS1 to fragment in MS2. In the middle panel the middle/down approach is illustrated, which can be viewed as a compromise between top-down and bottom-up. A partial digestion is performed, giving rise to large peptides which are separated and fragmented. This way, more information on cooccurring PTMs or can be maintained without the need to fragment whole proteins.

if the modifications were added to separate copies of the protein or if both were present on the same protein at the same time.

In summary, these advanced mass spectrometric techniques can all provide valuable information about the studied sample, albeit on different levels. The research question under investigation determines which technique is best suitable. Generally, each technique will always trade off demands on sample purity and complexity against the detail of the gathered information, see Figure 7.

A typical proteomics experiment

In a general proteomics workflow, the first step is to release proteins from the cells or tissues they were expressed in by a lysis step. This step needs to be optimized based on the sample: gram-negative bacteria benefit from the

bead beating or other types of mechanical stress to be properly lysed. Protease inhibitors are included to prevent nonspecific digestion by endogenous proteases. Following lysis, the samples are spun down to remove all insoluble proteins. This step illustrated the critical importance of your choice of lysis buffer. All the proteins that cannot be sufficiently solubilized will be lost upon centrifugation and never detected by MS.

Next, proteins are reduced to break any sulfur bridges that might have been present. Typically, this is achieved by incubating the proteins at 60 degrees Celsius for approximately one hour in the presence of a reducing agent (e.g. Dithiothreitol (DTT) or tris(2-carboxyethyl)phosphine (TCEP)). Immediately after reduction, all free cysteines are alkylated to prevent the reformation of sulfur bridges as a consequence of oxidation. Alkylation can be performed by incubating with Iodoacetamide (IAA) for 30 minutes. This incubation step is performed in the dark, as IAA is easily degraded by light. Any access IAA is quenched with reducing agent to prevent over alkylation (i.e. the alkylation of amino acids other than cysteine).

At this stage, the sample is ready to be digested by the protease of choice. Many candidates exist, but as mentioned above Trypsin is the most popular protease. It is crucial to optimize the protease activity, as any missed cleavage events will complicate your data analysis after the experiment. For Trypsin, pH is usually kept between 8.0 and 8.5 and the concentration of chaotropic agents (e.g. Urea or Guanidinium hydrochloride) needs to be reduced to 2M or less.

Following digestion, peptides need to be separated from incompletely digested proteins and other MS incompatible compounds, such as salts. This is typically achieved by using some form of solid phase extraction. The protocol to follow will depend on the solid phase material used to bind the peptides. After elution, peptides are dried down and transferred to a MS compatible buffer, ready for MS analysis.

In the Hecklab, peptides are analyzed by online coupling of liquid chromatography (LC) and tandem MS analysis (i.e. LC-MS/MS). In our standard approach, peptides are dissolved in a 20 mM citric acid buffer with 5% formic acid (FA) to ensure all basic residues are protonated. To reduce complexity, the peptide mixture is separated using an UHPLC 1290 (Agilent Technologies) using a trap column and analytical column filled in-house with C18 material. Chromatographic separation is followed by MS/MS analysis. From each full MS scan (375 to 1500 m/z) the most intense precursor ions are selected, isolated and fragmented. Because the 'best' ions in each precursor scan are selected for fragmentation, this is known as data-dependent acquisition. Depending on the MS of choice, different fragmentation methods can be chosen or combined to fragment selected peptides. Fragmentation spectra are collected for data analysis.

In a typical full proteome run (analyzed with a three-hour LC gradient) 150 000 to 170 000 MS2 spectra are recorded. To figure out which peptides belong to which fragmentation spectra, a database search is performed. From a given list of proteins, a theoretical digestion is made based on settings such as the indicated protease specificity and the allowed number of missed cleavage events (i.e. sites that match a proteases cleavage specificity, but were not hydrolyzed). From the theoretical digestion, options are selected that match the precursor mass and the theoretical fragmentation of these peptides are compared to ions observed in the actual MS/MS spectrum. The scoring algorithms differ depending on software used for the database search, but generally overlap of ions is scored, as well as the observation of complementary ion pairs (i.e. b- and y-ions resulting from the same bond being broken) and subsequent ion series (i.e. the observation of y_3, y_4, y_5 and y_6 make the match more likely than the observation of four random ions matching).

In this thesis

In the work described in this thesis, the central focus has been on the influence of proline amino acids on protein structure, protein regulation and its place in proteomics experiments.

In Chapter 2, we will evaluate how different PTMs can influence the placement or removal of other PTMs. We discuss the different mechanisms possible for both positive and negative interplay and look at some well-known examples when available. Most of this chapter is aimed at the interplay between O-GlcNac and phosphorylation, an interesting situation since both modifications occur on serine and threonine residues. Due to its enrichment near serine and threonine phosphorylation sites, proline is also involved and its role is briefly discussed.

In Chapter 3, we focus on the potential benefit of using proline residues as a cleavage site in proteomics. Since trypsin does not cleave when arginine or lysine are followed by a proline, the peptides resulting from cleavage at proline are expected to be highly complementary to the much more commonly analyzed tryptic peptides. In this chapter we touch upon some of the limitations of trypsin, the benefits of using alternative proteases and we compare the peptides and proteins found with either trypsin or EndoPro, the proline specific protease we explored.

We focus more on the characteristics of proline in Chapter 4. Since it can stably exist in both *cis* and *trans* isomers, proline adds an interesting, dynamic dimension to protein structure and the resulting protein biology. In this chapter we employ a bioinformatics approach to evaluate the abundance of *cis*/*trans* isomers of proline in protein structures and we investigate the proline-specific protease EndoPro for its isomer specificity.

In Chapter 5, we studied the full-length protein Pin1, the enzyme catalyzing the *cis*/*trans* isomerization of proline residues when they are directly adjacent to a phosphorylated serine or threonine residues. Following its expression and purification, we employed affinity based pull-down methods aimed to distinguish between different interactors of Pin1. We were able to purify and detect many known and novel Pin1 interactors, but still need further validation of the interactions and their functional relevance.

Finally, in Chapter 6, I close with a personal view on typical choices in proteomics workflows and how these may be suboptimal when interested in certain aspects of the proteome. My main point here is that we should always be aware of the choices we make and how these might affect the outcome of our experiments.

Bibliography

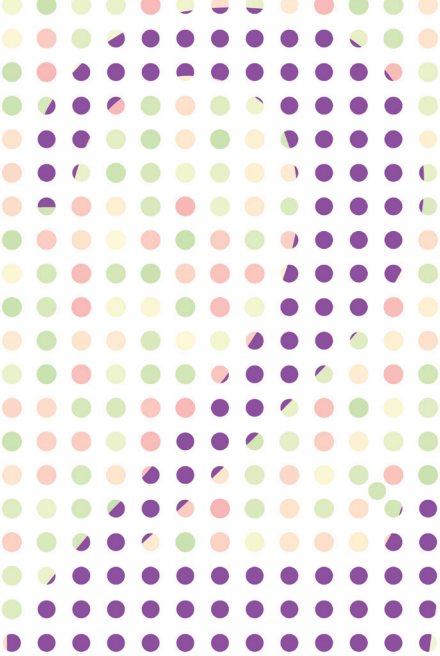
use of lysozyme, whereas organisms with thicker cell walls might require
1 Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA-K, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist P-H, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, Feilitzen K von, Forsberg M, Persson L, Johansson F, Zwahlen M, Heijne G von, Nielsen J & Pontén F (2015) Tissue-based map of the human proteome. Science (80-) 347, 394–407.
2 Branden C & Tooze J (1991) Introduction to Protein Structure.
3 Hausman R & Cooper G (2004) The Cell: a molecular approach ASM Press, Washington DC.
4 Garnier J, Gibrat J-F & Robson B (1996) [32] GOR method for predicting protein secondary structure from amino acid sequence. In Computer Methods for Macromolecular Sequence Analysis pp. 540–553.
5 Steward D, Sakar A & Wampler J (1990) Occurrence and Role of Cis Peptide bonds in Protein Structures. J Mol Biol 214, 253–260.
6 Lu KP, Finn G, Lee TH & Nicholson LK (2007) Prolyl cis-trans isomerization as a molecular timer. Nat Chem Biol 3, 619–629.
7 Rein T (2020) Peptidylprolylisomerases, Protein Folders, or Scaffolders? The Example of FKBP51 and FKBP52. BioEssays, 1900250.
8 Russo AA, Jeffrey PD & Pavletich NP (1996) Structural basis of cyclin-dependent kinase activation by phosphorylation. Nat Struct Mol Biol 3, 696–700.
9 Weiwad M, Küllertz G, Schutkowski M & Fischer G (2000) Evidence that the substrate backbone conformation is critical to phosphorylation by p42 MAP kinase. FEBS Lett 478, 39–42.
10 Reimer U & Fischer G (2002) Local structural changes caused by peptidyl–prolyl cis/trans isomerization in the native state of proteins. Biophys Chem 96, 203–212.
11 Harper JW & Bennett EJ (2016) Proteome complexity and the forces that drive proteome imbalance. Nature 537, 328–38.
12 Nishi H, Shaytan A & Panchenko AR (2014) Physicochemical mechanisms of protein regulation by phosphorylation. Front Genet 5, 270.
13 Nadaraia S, Yohrling GJ, Jiang GC-T, Flanagan JM & Vrana KE (2007) Enzyme Activity: Control. In Encyclopedia of Life Sciences John Wiley & Sons, Ltd, Chichester, UK.
14 Alberts B, Bray D, Hopkin K, Johnson AD, Lewis J & Roberts K (1997) Essential Cell Biology.
15 Deribe YL, Pawson T & Dikic I (2010) Post-translational modifications in signal integration. Nat Struct Mol Biol 17, 666–672.
16 Khoury GA, Baliban RC & Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci Rep 1, 1–5.
17 Lee Y-C, Que J, Chen Y-C, Lin J-T, Liou Y-C, Liao P-C, Liu Y-P, Lee K-H, Lin L-C, Hsiao M, Hung L-Y, Huang C-Y & Lu P-J (2013) Pin1 acts as a negative regulator of the G2/M transition by interacting with the Aurora-A-Bora complex. J Cell Sci 126, 4862–4872.

18 Venne AS, Kollipara L & Zahedi RP (2014) The next level of complexity: Crosstalk of posttranslational modifications. Proteomics 14, 513–524.
19 van der Laarse SAM, Leney AC & Heck AJR (2018) Crosstalk between phosphorylation and O-GlcNAcylation: friend or foe. FEBS J 285, 3152–3167.
20 Schutkowski M, Bernhardt A, Zhou XZ, Shen M, Reimer U, Rahfeld JU, Lu KP & Fischer G (1998) Role of phosphorylation in determining the backbone dynamics of the serine/threonine-proline motif and Pin1 substrate recognition. Biochemistry 37, 5566–75.
21 Lippens G, Landrieu I & Smet C (2007) Molecular mechanisms of the phospho-dependent prolyl cis/trans isomerase Pin1. FEBS J 274, 5211–5222.
22 Sudol M, Sliwa K & Russo T (2001) Functions of WW domains in the nucleus. Fed Eur Biochem Soc Lett 490, 190–195.
23 Yaffe MB (1997) Sequence-Specific and Phosphorylation-Dependent Proline Isomerization: A Potential Mitotic Regulatory Mechanism. Science (80-) 278, 1957–1960.
24 Lee YM & Liou Y-C (2018) Gears-In-Motion: The Interplay of WW and PPIase Domains in Pin1. Front Oncol 8.
25 Vöhringer-Martinez E, Duarte F & Toro-Labbé A (2012) How Does Pin1 Catalyze the Cis–Trans Prolyl Peptide Bond Isomerization? A QM/MM and Mean Reaction Force Study. J Phys Chem B 116, 12972–12979.
26 Lu P-J, Zhou XZ, Shen M & Lu KP (1999) Function of WW Domains as Phosphoserine- or Phosphothreonine-Binding Modules. Science (80-) 283, 1325–1328.
27 Yaffe MB, Schutkowski M, Shen M, Zhou XZ, Stukenberg PT, Rahfeld JU, Xu J, Kuang J, Kirschner MW, Fischer G, Cantley LC & Lu KP (1997) Sequence-Specific and Phosphorylation-Dependent Proline Isomerization: A Potential Mitotic Regulatory Mechanism. Science (80-) 278, 1957–1960.
28 Ranganathan R, Lu KP, Hunter T & Noel JP (1997) Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. Cell 89, 875–86.
29 Werner-Allen JW, Lee C-J, Liu P, Nicely NI, Wang S, Greenleaf AL & Zhou P (2011) cis -Proline-mediated Ser(P) 5 Dephosphorylation by the RNA Polymerase II C-terminal Domain Phosphatase Ssu72. J Biol Chem 286, 5717–5726.
30 Lu KP (2004) Pinning down cell signaling, cancer and Alzheimer's disease. Trends Biochem Sci 29, 200–2009.
31 de Hoffmann E & Stroobant V (2007) Mass Spectrometry Principles and Applications Wiley.
32 Yates III JR (2011) A century of mass spectrometry: from atoms to proteomes. Nat Methods 8, 633–637.
33 Fenn J, Mann M, Meng C, Wong S & Whitehouse C (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science (80-) 246, 64–71.
34 Cole RB (2000) Some tenets pertaining to electrospray ionization mass spectrometry. J Mass Spectrom 35, 763–772.
35 Manisali I, Chen DDY & Schneider BB (2006) Electrospray ionization source geometry for mass spectrometry: past, present, and future. Trends Anal Chem 25, 243–256.

36 March R An introduction to quadrupole ion trap mass spectrometry. 32, 351–369.
37 Douglas DJ, Frank AJ & Mao D (2005) Linear ion traps in mass spectrometry. Mass Spectrom Rev 24, 1–29.
38 Barnes JH & Hieftje GM (2004) Recent advances in detector-array technology for mass spectrometry. Int J Mass Spectrom 238, 33–46.
39 Boesl U (2017) Time-of-flight mass spectrometry: Introduction to the basics. Mass Spectrom Rev 36, 86–109.
40 Makarov A (2000) Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. Anal Chem 72, 1156–1162.
41 Brenton AG & Godfrey AR (2010) Accurate Mass Measurement: Terminology and Treatment of Data. J Am Soc Mass Spectrom 21, 1821–1835.
42 Lesur A & Domon B (2015) Advances in high-resolution accurate mass spectrometry application to targeted proteomics. Proteomics 15, 880–890.
43 CHONG KF & LEONG HW (2012) TUTORIAL ON DE NOVO PEPTIDE SEQUENCING USING MS/MS MASS SPECTROMETRY. J Bioinform Comput Biol 10, 1231002.
44 Guthals A & Bandeira N (2012) Peptide Identification by Tandem Mass Spectrometry with Alternate Fragmentation Modes. Mol Cell Proteomics 11, 550–557.
45 Hung C-W, Schlosser A, Wei J & Lehmann WD (2007) Collision-induced reporter fragmentations for identification of covalently modified peptides. Anal Bioanal Chem 389, 1003–1016.
46 Keller BO, Sui J, Young AB & Whittal RM (2008) Interferences and contaminants encountered in modern mass spectrometry. Anal Chim Acta 627, 71–81.
47 Lehmann WD, Krüger R, Salek M, Hung C-W, Wolschin F & Weckwerth W (2007) Neutral Loss-Based Phosphopeptide Recognition: A Collection of Caveats. J Proteome Res 6, 2866–2873.
48 Palumbo AM, Smith SA, Kalcic CL, Dantus M, Stemmer PM & Reid GE (2011) Tandem mass spectrometry strategies for phosphoproteome analysis. Mass Spectrom Rev 30, 600–625.
49 McLafferty FW & Bryce TA (1967) Metastable-ion characteristics: characterization of isomeric molecules. Chem Commun, 1215.
50 Jennings KR (1968) Collision-induced decompositions of aromatic molecular ions. Int J Mass Spectrom Ion Phys 1, 227–235.
51 Roepstorff P & Fohlman J (1984) Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. Biol Mass Spectrom 11, 601–601.
52 Solari FA, Dell'Aica M, Sickmann A & Zahedi RP (2015) Why phosphoproteomics is still a challenge. Mol Biosyst 11, 1487–1493.
53 Zubarev RA, Kelleher NL & McLafferty FW (1998) Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. J Am Chem Soc 120, 3265–3266.
54 Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J & Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci 101, 9528–9533.
55 Shliaha P V., Gibb S, Gorshkov V, Jespersen MS, Andersen GR,

Bailey D, Schwartz J, Eliuk S, Schwämmle V & Jensen ON (2018) Maximizing Sequence Coverage in Top-Down Proteomics By Automated Multimodal Gas-Phase Protein Fragmentation. Anal Chem 90, 12519–12526.
56 Frese CK, Altelaar AFM, van den Toorn H, Nolting D, Griep-Raming J, Heck AJR & Mohammed S (2012) Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. Anal Chem 84, 9668–9673.
57 Frese CK, Zhou H, Taus T, Altelaar AFM, Mechtler K, Heck AJR & Mohammed S (2013) Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). J Proteome Res 12, 1520–5.
58 Woodin RL, Bomse DS & Beauchamp JL (1978) Multiphoton dissociation of molecules with low power continuous wave infrared laser radiation. J Am Chem Soc 100, 3248–3250.
59 Fort KL, Dyachenko A, Potel CM, Corradini E, Marino F, Barendregt A, Makarov AA, Scheltema RA & Heck AJR (2016) Implementation of Ultraviolet Photodissociation on a Benchtop Q Exactive Mass Spectrometer and Its Application to Phosphoproteomics. Anal Chem 88, 2303–2310.
60 Leney AC & Heck AJR (2017) Native Mass Spectrometry: What is in the Name? J Am Soc Mass Spectrom 28, 5–13.
61 Fort KL, van de Waterbeemd M, Boll D, Reinhardt-Szyba M, Belov ME, Sasaki E, Zschoche R, Hilvert D, Makarov AA & Heck AJR (2018) Expanding the structural analysis capabilities on an Orbitrap-based mass spectrometer for large macromolecular complexes. Analyst 143, 100–105.
62 van de Waterbeemd M, Fort KL, Boll D, Reinhardt-Szyba M, Routh A, Makarov A & Heck AJR (2017) High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. Nat Methods 14, 283–286.
63 Cui W, Rohrs HW & Gross ML (2011) Top-down mass spectrometry: Recent developments, applications and perspectives. Analyst 136, 3854.
64 Catherman AD, Skinner OS & Kelleher NL (2014) Top Down proteomics: Facts and perspectives. Biochem Biophys Res Commun 445, 683–693.
65 Toby TK, Fornelli L & Kelleher NL (2016) Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem 9, 499–519.
66 Resemann A, Jabs W, Wiechmann A, Wagner E, Colas O, Evers W, Belau E, Vorwerg L, Evans C, Beck A & Suckau D (2016) Full validation of therapeutic antibody sequences by middle-up mass measurements and middle-down protein sequencing. MAbs 8, 318–330.
67 Fornelli L, Ayoub D, Aizikov K, Beck A & Tsybin YO (2014) Middle-Down Analysis of Monoclonal Antibodies with Electron Transfer Dissociation Orbitrap Fourier Transform Mass Spectrometry. Anal Chem 86, 3005–3012.
68 Zubarev RA (2013) The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 13, 723–726.
69 Corthals GL, Wasinger VC, Hochstrasser DF & Sanchez J-C (2000) The dynamic range of protein expression: A challenge for

proteomic research. Electrophoresis 21, 1104–1115.
70 Zhang Y, Fonslow BR, Shan B, Baek M-C & Yates JR (2013) Protein analysis by shotgun/bottom-up proteomics. Chem Rev 113, 2343–94.
71 Altelaar AFM, Munoz J & Heck AJR (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 14, 35–48.
72 Vandermarliere E, Mueller M & Martens L (2013) Getting intimate with trypsin, the leading protease in proteomics. Mass Spectrom Rev, 000–000.
73 Michalski A, Neuhauser N, Cox J & Mann M (2012) A Systematic Investigation into the Nature of Tryptic HCD Spectra. J Proteome Res 11, 5479–5491.
74 Tsiatsiani L & Heck AJR (2015) Proteomics beyond trypsin. FEBS J Doi: 10.1111/febs.13287



Chapter 2

Crosstalk between Phosphorylation and O-GlcNAcylation: friend or foe

Saar A.M. van der Laarse[a,b]+, Aneika C. Leney[a,b]+, Albert J. R. Heck[a,b]

[a] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584CH, Utrecht (The Netherlands)
[b] Netherlands Proteomics Centre, Padualaan 8, 3584CH, Utrecht (The Netherlands)
+ These authors contributed equally.

Published in The FEBS Journal, Volume 285, Issue 17, pages 3152 - 3167, DOI: 10.1111/febs.14491

Abstract

A wide variety of protein post-translational modifications (PTMs) decorate cellular proteins, regulating their structure, interactions and ultimately their function. The density of co-occurring PTMs on proteins can be very high, where multiple PTMs can positively or negatively influence each other's actions, nowadays termed PTM crosstalk. In this review we highlight recent progress in the area of PTM crosstalk, whereby we focus on crosstalk between protein phosphorylation and O-GlcNAcylation. These two PTMs largely target identical (i.e. Ser and Thr) amino acids in proteins. Phosphorylation/O-GlcNAcylation crosstalk comes in many flavors, for instance by competition for the same site/residue (reciprocal crosstalk), as well as by modifications influencing each other in proximity or even distal on the protein sequence. PTM crosstalk is observed on the writers of these modifications (i.e. kinases and O-GlcNAc transferase), on the erasers (i.e. phosphatases/O-GlcNAcase) and on the readers and the substrates. Although likely not all-inclusive, we describe examples of all these different flavors of crosstalk, and additionally the methods that are emerging to better investigate in particular phosphorylation/O-GlcNAcylation crosstalk.

Introduction

Proteins play key roles in essential biological processes, ranging from the catalysis of biochemical reactions to the transport of nutrients, signal transduction and the maintenance of the cells mechanical structure[1]. Many proteins are modified post-translationally to regulate their structure, activity, subcellular localization and overall function. To date, over 80,000 unique sites bearing post-translational protein modifications (PTMs) have been reported, including acetylation, phosphorylation, glycosylation, methylation and ubiquitination[2]. Many of these chemically different PTMs act mechanistically somewhat similar. Most PTMs are dynamic, whereby the modification is added to the protein by an enzyme, termed a writer, and removed again by another protein termed an eraser. The modifications can then be recognized by other proteins, referred to as readers[3,4].

The most studied and, therefore, probably best characterized protein PTM is phosphorylation, a modification whereby a phosphoryl group is covalently linked predominantly to the hydroxyl groups of selected serine, threonine and tyrosine residues[5]. Phosphorylation, however, has also been reported on histidine, lysine and arginine or aspartate and glutamate[6,7]. In the case of protein phosphorylation, the writers are called kinases, the erasers phosphatases. At any given moment, approximately 30% of the human proteome is thought to be phosphorylated[8]. This phosphorylation can activate or deactivate a proteins function, affect its conformation or influence its localization[5,8]. There are many readers of phosphorylation, such as those containing SH2 or WW-domains that can interact in a phosphorylation dependent manner[9,10]. For example, several ubiquitin ligase complexes recognize specific, phosphorylated sequences on proteins known as phosphodegrons, that, once phosphorylated, target these proteins for degradation[11].

Another highly prevalent, yet lesser studied PTM, is O-GlcNAcylation. O-GlcNAcylation is a unique type of glycosylation whereby a single sugar moiety, O-linked N-acetylglucosamine (O-GlcNAc), is transferred typically to the hydroxyl groups of serine and threonine residues of proteins[12]. In sharp contrast to protein phosphorylation, for which hundreds of writers/kinases exist[13–15], there is only one writer for protein O-GlcNAcylation, namely the O-GlcNAc transferase (OGT). Moreover, in contrast to the plethora of co-existing protein phosphatases, there is also only one so-called eraser for O-GlcNAc, namely the O-GlcNAcase (OGA), and to date, although hypothesized to exist[16], no readers of O-GlcNAcylation have been identified. Analogous to phosphorylation, O-GlcNAcylation plays an important role in cell signaling and gene regulation[17–20] and more than 50% of all human proteins are expected to be glycosylated[21,22].

Most protein PTMs occur on specific amino acids. Thus, evidently, a single protein can harbor multiple of the same PTMs (e.g. multiple phosphorylations), whereby the first modification may affect the addition or removal of the next.

This is illustrated in the well-known example of substrates of the glycogen synthase kinase-3β (GSK3β), where pre-phosphorylation is needed, four amino acids upstream of the GSK3β phosphorylation site, before GSK3β will act to add an extra phosphate moiety to the substrate[23]. Similarly, a certain PTM can also affect a chemically different type of PTM, which has been coined as PTM crosstalk. This phenomenon is thought to be highly prevalent in cells, yet is still challenging to investigate. In general, one PTM can affect another in two ways, either promoting (i.e. a friend) or hampering (i.e. a foe) the presence of the other modification. This interplay between modifications can occur in many different ways, as summarized in Figure 1. Crosstalk can be on the same writer, reader or eraser, directly influencing the placement of a second PTM nearby or even one located much further away in sequence or in space. The placement of the second modification may then either enhance or hamper the activity of the targeted writer/eraser/reader. In recent years, several examples of crosstalk have been described for a wide array of proteins, ranging from kinases to transcription factors and from histone deacetylases to filament proteins[24–27]. Particularly well-studied examples of PTM crosstalk occur on histones/nucleosomes, highly conserved proteins that function in genome packaging. Histones are modified by at least eleven different PTMs, thus the number of types of PTM crosstalk on histones is immense, whereby the full extent of modifications and their crosstalk has been termed the histone code[28–31]. O-GlcNAcylation was also reported on histones, suggesting that O-GlcNAc is directly involved in the histone code[32–34]. However, detecting O-GlcNAcylation on histones is not straightforward and thus concerns have been raised as to the abundance of this PTM on histones in vivo [35].

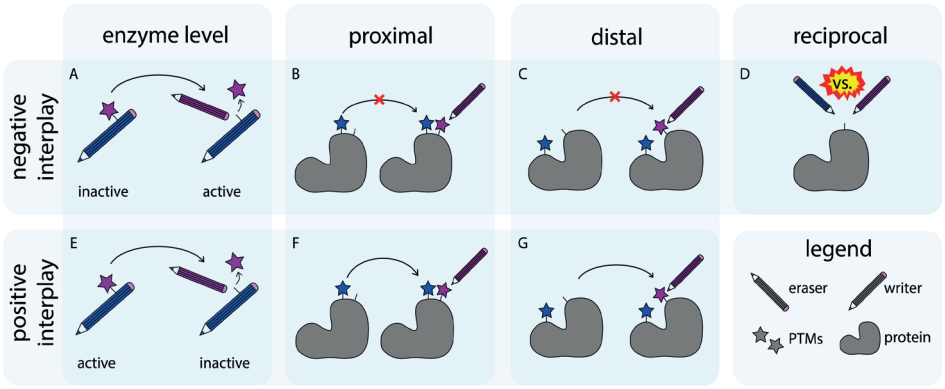


Figure 1 - Types of PTM Crosstalk. PTMs can affect each other in both a negative (A - D) and a positive manner (E - G), on the enzyme and the substrate level. The modification by a PTM on writers, readers and erasers can lead to their deactivation (A) or activation (E), hence affecting the regulation of that particular PTM (shown here in the case of erasers). When two potential PTM-sites occur in close proximity, modification of one of them can prevent modification of the other (B) or modification of the first site can promote modification of the second site (F). Crosstalk can also occur at distal sites whereby the modification of a protein on one side may prevent or enable the modification on another side (C and G, respectively). Finally, reciprocal crosstalk can occur whereby one modification competes with another to modify the same amino acid (D).

In this review, we focus on the so far most documented form of PTM crosstalk, that between phosphorylation and O-GlcNAcylation. As both phosphorylation and O-GlcNAcylation occur mainly on serine and threonine residues, it is not very surprising that these two PTMs undergo crosstalk. Indeed, when over 800 phosphorylation sites were monitored upon the inhibition of OGA, phosphorylation on 280 sites decreased in abundance at the same time as phosphorylation on 148 sites increased. Also, changes in the O-GlcNAcylation pattern were observed upon inhibition of serine/threonine phosphatases [31]. In addition, the characterization of phosphorylated and O-GlcNAcylated proteins located at the Murine synapse revealed that 52% of the identified protein phosphatases were phosphorylated and 8% was O-GlcNAcylated. In addition, 66% of the identified kinases was phosphorylated whereas 16% were O-GlcNAcylated, indicating that the modification of kinases and phosphatases is quite common[36]. These findings indicate that the relationship between phosphorylation and O-GlcNAcylation is complex and widespread. Here, we first describe in detail how O-GlcNAcylation is known to be regulated in cells. Secondly, we describe the effect that phosphorylation has on O-GlcNAcylation, detailing how these two modifications can affect each other. Indeed, many examples of crosstalk between phosphorylation and O-GlcNAcylation are emerging either occurring at the substrate level, or on the writers and erasers themselves. Most of these modifications directly modulate protein function. Finally, our review includes a description of the recent technological advances in the analysis and prediction of crosstalk between phosphorylation and O-GlcNAcylation.

Regulation of O-GlcNAcylation

Protein O-GlcNAcylation plays essential roles in the regulation of transcriptional activity[37–40], neuronal function[40], epigenetic regulation[41], modulation of protein-protein interactions[16] and the response to external stress factors[42]. Moreover, protein O-GlcNAcylation is very sensitive to the availability of uridine diphospho-N-acetylglucosamine (UDP-GlcNAc), the GlcNAc donor for O-GlcNAcylation[43] and downstream metabolite of glucose, hence O-GlcNAcylation is often referred to as a nutrient sensor in cells[20,43,44]. Loss of O-GlcNAcylation was shown to result in loss of cellular function or even cell death[45] and the disturbance of normal O-GlcNAc function has been linked to many diseases, including Alzheimer's disease, diabetes and other chronic illnesses[46,47]. Therefore, the proper functioning of O-GlcNAcylation in cells is essential, however, how this regulation is fine-tuned is still not well understood. Only a single OGT enzyme is responsible for O-GlcNAcylation of over a thousand protein substrates[48]. OGT has three isoforms that differ in the number of tetratricopeptide repeats (TPRs) they harbor at their N-terminus, a well-described protein-protein interaction-mediating domain[49]. Regulation may in part, therefore, be governed by (the number of) these TPR repeats. The OGT isoforms localize to different compartments of the cell. Short OGT (sOGT) and nucleocytoplasmic OGT (ncOGT) are found in the nucleus and cytoplasm,

whereas mitochondrial OGT (mOGT) is found exclusively in mitochondria[50]. Localization of mOGT and ncOGT is directed by localization sequences. mOGT has an N-terminal mitochondrial localization sequence, whereas ncOGT contains a nuclear localization sequence between its TPRs and the catalytic domain located at the C-terminus[51,52]. Deletion of this sequence prevents transport of ncOGT into the nucleus, resulting in an increase of cytosolic O-GlcNAcylation[51]. The localization sequence in ncOGT is also present in the other isoforms, thus sOGT likely also has a localization sequence whose relevance is yet to be reported.

sOGT, mOGT and ncOGT have 2.5, 9 and 11.5 TPRs, respectively. These TPRs are thought to be involved in substrate recognition[53], and removal of all TPRs abolished OGT activity towards protein substrates. Interestingly, the activity towards small peptide substrates in vitro remains unaffected upon TPR removal, indicating that the TPR repeats could play a specific role in mediating OGT substrate specificity[17,54]. Indeed, interaction of OGT with myosin phosphatase target subunit 1 influenced OGT substrate specificity in vitro[55]. This principle is illustrated in Figure 2, whereby OGT substrate specificity is defined by the initial complex formed with so-called adapter proteins. For example, OGT is recruited to specific promotors by the paired amphipathic helix protein Sin3A, resulting in transcriptional repression[56]. Furthermore, the interaction between OGT and p38, a mitogen-activated protein kinase, increases OGT activity by targeting OGT to Neurofilament H, as well as other targets[57]. Many other proteins have been shown to interact with OGT, however, the biological relevance of these interactions, and whether they play a regulatory role is often not yet fully understood[53,55,57–61].

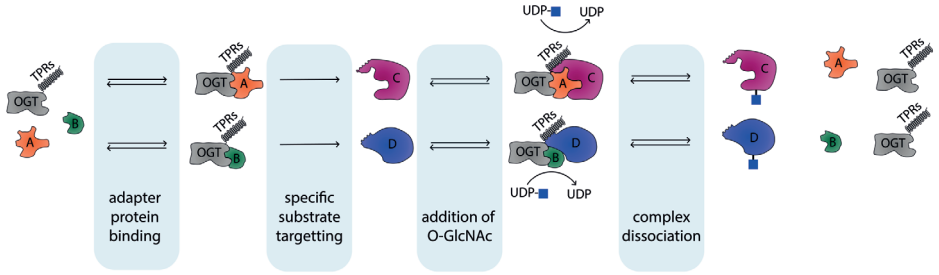


Figure 2 - Targeting of OGT to subsets of its substrates. OGT specificity can be governed by its interaction with so-called adapter proteins, directing OGT activity to a subset of substrates. OGT can form complexes with adaptor proteins A and B. Proteins C and D are both substrates of OGT, but C can only be O-GlcNAcylated by the OGT-A complex, whereas D can only be O-GlcNAcylated by the OGT-B complex.

Both writers and erasers act to regulate the presence of PTMs. The O-GlcNAc eraser, OGA, was identified by Hart et al. in 1994 and is mainly present in the cytosol[12,62]. OGA catalyzes the removal of O-GlcNAc from proteins via a substrate-assisted mechanism, utilizing the acetamido group of the sugar

moiety[63,64]. Thus, in addition to OGT, the regulation of OGA needs to be tightly controlled to prevent disease. Similar to OGT, the single OGA gene can encode two isoforms, namely a short isoform which appears inactive and the full length isoform[65]. The short isoform lacks a putative acetyl transferase domain located at the C-terminus, however, OGA does not have acetyl transferase activity[12,52,66]. In addition, full length OGA contains a caspase-3 cleavage site, although cleavage at this site does not reduce the enzymatic activity of OGA[67,68]. Interestingly, OGA seems to form dimers and the intersubunit interactions differ between catalytically active and compromised OGA variants, indicating that the dimerization of OGA could also influence the binding of substrates and affect OGA activity[68–71].

The writers and readers of O-GlcNAcylation themselves, however, are not the only enzymes responsible for the regulation of O-GlcNAcylation in cells. Moreover, the interaction of OGT and OGA with various kinases and phosphatases hints that phosphorylation may additionally play a role in O-GlcNAc regulation. Indeed, as described above, it is known that writers and erasers can be modified themselves to increase or decrease their own enzymatic activity (Figure 1A,E).

Phosphorylation/O-GlcNAcylation crosstalk at the enzyme level

There are ample examples reported of phosphorylation/O-GlcNAcylation crosstalk whereby the modification of the writers and erasers of phosphorylation and O-GlcNAcylation themselves affect their own enzymatic capability (Table 1). For example, the Calcium/Calmodulin-dependent Kinase IV (CaMKIV) was found to be phosphorylated at multiple phosphorylation sites, including Thr200[72]. Phosphorylation at this site is required for CaMKIV activation[73,74]. Dias *et al.* showed that CaMKIV is also O-GlcNAcylated on at least five specific residues, including Ser137 and Ser189. Mutation of Ser137 into an alanine resulted in a slight increase in CaMKIV phosphorylation at Thr200, however, mutation of Ser189 to alanine resulted in a significant increase of Thr200 phosphorylation. Moreover, a Ser189 CaMKIV mutant showed a five-fold increase in activity, indicating that O-GlcNAcylation of CaMKIV on Ser189 has an inhibitory effect on the kinase activity[72].

Another example was reported by Tarrant *et al.*, who demonstrated that phosphorylated Casein kinase 2α (CK2α) can readily phosphorylate Copine 1 isoform a, progesterone receptor membrane component 1 and the Golgin subfamily A member 4 protein, whereas O-GlcNAcylated CK2α could no longer phosphorylate these substrates. Inversely, O-GlcNAcylated CK2α could phosphorylate death-associated protein kinase 2, which was not phosphorylated by phosphorylated CK2α[75]. Hence, the modification of CK2α by phosphorylation or O-GlcNAcylation affects CK2α substrate activity and thus specificity. Utilizing a more global approach, Hart and coworkers

revealed that O-GlcNAcylation of kinases is a widespread phenomenon. They screened a protein array containing 152 kinases, confirming that 42 of them could be O-GlcNAcylated *in vitro*[76]. Recently, Shi et al. showed that protein kinase B is O-GlcNAcylated at Thr308 and Ser473[77], which are known to be phosphorylated by phosphoinositide-dependent protein kinase 1 and mechanistic target of rapamycin complex 2, respectively[78]. Phosphorylation on these residues is required for protein kinase B activation[78,79]. The overexpression of OGT revealed a decrease of phosphorylation at these sites, whereas overexpression of OGA led to an increase of phosphorylation, indicating a reciprocal relationship between O-GlcNAcylation and phosphorylation at these sites[77]. The actual relationship between OGT and protein kinase B is likely more complex than mere reciprocality, however, as OGT and OGA knockdowns have been reported to decrease and increase the level of phosphorylated protein kinase B, respectively [80,81].

In addition to the extensive crosstalk on kinases induced by O-GlcNAcylation, the O-GlcNAc cycling enzymes OGT and OGA can also be post-translationally modified themselves, altering their activity. OGT can be phosphorylated on at least four different sites. Kaasik *et al.* showed that GSK3β can phosphorylate OGT at Ser3 and/or Ser4, whereby these phosphorylations enhance OGT activity[82]. Recently, Li *et al.* showed that Checkpoint kinase 1 can phosphorylate OGT on Ser20, which seemingly improves OGT stability[83]. In addition, AMPK was found to phosphorylate OGT on Thr444, which induced changes in OGT substrate specificity[34,84]. Finally, tyrosine phosphorylation was reported for the rat homologue of OGT, possibly also occurring on Tyr989 on the human homologue[85]. Whelan *et al.* showed that tyrosine phosphorylation of OGT is increased after insulin stimulation and that the isolated insulin receptor complex, as well as proto-oncogene tyrosine-protein kinase Src could phosphorylate OGT, resulting in an increase of OGT activity[86]. Interestingly, OGT can also be O-GlcNAcylated on Ser3 and Ser4[82], hinting at a potential reciprocal occupancy of these sites by O-GlcNAc and phospho modifications. OGA can be O-GlcNAcylated on Ser405[87,88], however, the functional effect of this modification is not yet known.

Phosphorylation/O-GlcNAcylation crosstalk at the substrate level

In addition to PTM crosstalk occurring on the 'writers' or 'erasers' themselves, crosstalk has also been reported whereby the same residue within a substrate protein sequence or residues in close proximity are modified by distinct PTMs (Figure 1, Table 1)[89,90]. Crosstalk whereby competition occurs for the same residue has been termed reciprocal crosstalk (Figure 1D). Early evidence for the existence of reciprocal crosstalk came shortly after the discovery of O-GlcNAcylation by Hart *et al*. in 1984 when it was noted that O-GlcNAcylation occurs on known phosphorylation sites[91,92]. The idea that such a reciprocal crosstalk mechanism could be more generic was validated by

the fact that activation of protein kinase A and protein kinase C individually resulted in a global decrease of O-GlcNAcylation in cellular neurons and, inversely, inhibition of protein kinase A resulted in an overall increase of O-GlcNAcylation[93]. In line with these findings, the inhibition of GSK3β also resulted in an increase of O-GlcNAcylation in COS7 cells[94]. Now, many cases of reciprocal crosstalk between O-GlcNAcylation and phosphorylation have been identified, whereby several individual sites of modification have been mapped. Noteworthy examples include the modification of Thr58 by either phosphorylation or O-GlcNAcylation on the transactivation domain of c-Myc[95–97] and the modification of Ser16 on the N-terminus of the murine estrogen receptor beta which functions to regulate its activity[98] (Table 1).

Phosphorylation/O-GlcNAcylation crosstalk also occurs on Ser/Thr residues that are in close proximity to each other along a protein sequence (Figure 1B,F). In these cases, the O-GlcNAcylation or phosphorylation may have specific effects on the proteins function, whereby the PTM crosstalk can act to prevent or enhance one of these actions. For example, O-GlcNAcylation of p53 at Ser149 hampers phosphorylation at Thr155, a site targeted by the COP9 signalosome, finally resulting in p53 ubiquitination and degradation[99]. Indeed, this "proximity" type of crosstalk features in many biological systems fine-tuning activity. For example, in the circadian clock, O-GlcNAcylation of the period circadian regulator 2 (PER2) at Ser662 in its regulatory region blocks casein kinase I dependent PER2 phosphorylation[82]. Also in NF-κB signaling, p65 is O-GlcNAcylated at Thr305 and Ser319, where phosphorylation at Thr308 may prevent O-GlcNAcylation at Thr305[100]. In addition, crosstalk was hypothesized to also occur between tyrosine phosphorylation and O-GlcNAcylation[101], which sites were later validated using a peptide microarray screening study by Pieters and co-workers[102]. One specific example occurs on STAT5, where O-GlcNAcylation of STAT5 controls tyrosine phosphorylation[103]. These disjointed examples show that phosphorylation/O-GlcNAcylation crosstalk at nearby sites in proteins is a generic feature in cellular signaling. The distance between the involved PTM sites in phosphorylation/O-GlcNAcylation crosstalk can be anything from a single amino acid apart[104] to six residues apart in the case of p53[100].

Phosphorylation/O-GlcNAcylation crosstalk can even occur whereby the two PTMs are situated at quite distal sites with respect to one another (Figure 1C,G). Although far away in terms of protein sequence, these modifications could still be spatially in close proximity, altering the ability of the 'writers' to interact with their substrates. A prime example of this comes from the insulin receptor substrate 1, whereby phosphorylation occurs at the N-terminus and O-GlcNAcylation at its C-terminus[86,105,106]. Alternatively, a PTM in one loop or domain may cause a conformational change in the protein effecting the whole protein. As in all types of crosstalk, distal crosstalk can both catalyze PTM addition, for example in the case of the enzyme endothelial nitric oxide synthase, where phosphorylation at Ser615 enhances phosphorylation at Ser1177, and hamper PTM addition, for example the O-GlcNAcylation of Ser615

on endothelial nitric oxide synthase prevents phosphorylation at Ser1177[107].

Prediction of phosphorylation/O-GlcNAcylation crosstalk

With serine and threonine being amongst the most frequently occurring amino acids in vertebrates, the number of residues that potentially could be modified by O-GlcNAcylation or phosphorylation, and thus be involved in PTM crosstalk, is in theory almost indefinite. Elucidating which residues can be modified, and when, is imperative for understanding cellular signaling and thus how cells function. Indeed, not all serine and threonine residues are modified. Most protein kinases target specific Ser/Thr residues to phosphorylate. Each kinase has evolved to recognize a specific amino acid sequence or structural motif. Developments in mass spectrometry (MS) and phosphopeptide enrichment methods[108,109] have enabled thousands of phosphosites to be identified and site-localized located within a single experiment[110]. These large datasets have been inferred to elucidate specific kinase motifs[111]. Thus, due to the wealth of information available on phosphosites, potential Ser/Thr phosphorylation on proteins can now be predicted with some degree of accuracy[112].

When it comes to O-GlcNAcylation, however, the substrate motif is more diffusely defined. The first prediction software for O-GlcNAcylation sites, termed YinOYang, was based on the observations that proline residues commonly occurred at the -4, -3 and -2 positions, with valines at the -1, +2, +4 and +5 positions and there was a high frequency of serine residues in the targeted stretch, typically at positions +1, +4 and +7[113,114]. It was also noted that leucine and glutamine residues were less frequently observed within the substrate motif. Although a reasonable start for O-GlcNAc prediction, this software was based on only 40 experimentally determined O-GlcNAcylation sites. A decade later, the number of O-GlcNAcylation sites identified on proteins had dramatically increased, primarily due to advances in MS based detection of O-GlcNAcylation. From this data, a database of O-GlcNAcylated proteins (dbOGAP) was generated and a new prediction software emerged, termed O-GlcNAc scan, whose algorithm is based on ~800 experimentally determined O-GlcNAcylated proteins[115]. Although more accurate, both these methods are prone to false positive and negative predictions[116]. Most recently, by using machine learning to predict O-GlcNAc sites[117–119], O-GlcNAc sites identification has further improved with O-GlcNAc sites being predicted with supposedly up to 84% accuracy[119].

Table 1. Key examples of proteins identified to be involved in different types of phosphorylation/O-GlcNAcylation crosstalk.

Crosstalk		Type	Protein	Mod. sites	Kinase	Role	Reference
Enzyme level		negative	CaMKIV	gSer189 pThr200	CaMKK [74]	Phosphorylation required for activation, O-GlcNac inhibits phosphorylation	[72,73,152]
		modulating	CK2α	gSer347 pThr344	CDK1 [153]	Differential substrate affinity	[75]
		positive	AMPK	pThr172	LKB1 [154]	Phosphorylation at Thr172 enhances O-GlcNAcylation of the kinase	[84]
		negative	IKKbeta	gSer733 pSer733	Plk1 [155]	Phosphorylation inactivates kinase	[156]
		modulating	OGT	gSer3 pSer3 gSer4 pSer4 pThr444	GSK3β [82] AMPK [84]	Phosphorylation of Ser3/4 enhanced activity, phosphorylation at Thr444 is associated with OGT nuclear localization and altered substrate specificity	[82,84]
Substrate level	Proximal	negative	p53	gSer149 pThr150 pThr155	CK2α [99,157]	O-GlcNAcylation prevents phosphorylation, which is required for p53 ubiquitination and subsequent degradation	[99]
		negative	c-myc	gThr58 pThr58	GSK3β [158,159]	unknown	[95–97]
		negative	ER-beta	gSer16 pSer16	unknown	Conformational change, protein stabilization	[98]
		reciprocal	β-catenin	gThr41 pThr41	GSK3β [160]	O-GlcNAcylation stabilizes β-catenin by preventing phosphorylation of its destruction box	[160]
		reciprocal	POLR2A	pThr1616 gThr1618 pThr1618 pThr1619	BRD4 [161] CDK9 [162] CDK12 [163] DYRK1A [164]	Reciprocal behavior altering the PTM code on the C-terminal domain.	[129,130]
	Distal	positive	eNOS	pSer615 pSer1177	PKB [165,166]	Phosphorylation at Ser615 enhances phosphorylation at Ser1177	[107]
		negative	eNOS	gSer615 pSer1177	PKB [165,166]	O-GlcNAcylation of Ser615 prevents phosphorylation at Ser1177	[107]
		unknown	FoxO1	pSer256 gThr317 pSer319 gSer550 gThr646 gThr648 pThr649 gSer654	PKB [167,168] AMPK [169]	Interplay observed at global level, exact effects not yet known.	[170,171]
		unknown	IRS1	gSer984 gSer985 gSer1011	unknown	Interplay observed at global level, involved sites are not yet known.	[172,173]

As the number of O-GlcNAcylation sites identified experimentally continues to rise, the motif (or motifs) for OGT will undoubtedly become clearer. To fast track the identification of OGT substrates, peptide microarrays have been developed whereby thousands of proteins can be screened simultaneously for their ability to be OGT substrates[120,121]. These arrays have helped to define many new important OGT substrates in vitro, which can subsequently be verified in vivo, although that is still much more laborious. For example, using such approaches OTX2, a transcription factor critical for brain development, whose misregulation is associated with the most malignant brain tumors in children, was found to contain multiple O-GlcNAcylation sites[121]. Observation of an O-GlcNAc site alone, however, does not provide information as to whether the protein of interest is a good or bad substrate of OGT. Indeed, good motifs for OGT should not only be based on sequence and detection alone, but also on the efficiency at which these sequences can be O-GlcNAcylated on the native protein, ideally in vivo. Elegant work by van Aalten and co-workers addressed the O-GlcNAcylation efficiency, taking 720 peptides harboring a putative OGT site. They render a motif based on only the 32 most efficient O-GlcNAcylation substrates; giving [TS][PT][VT]S/T[RLV][ASY][122]. Obviously, there are proteins O-GlcNAcylated that do confer to this rather strict O-GlcNAcylation motif. For example, Sox2, a transcription factor that functions in embryonic stem cell differentiation, has been found O-GlcNAcylated in mice at residue 248 within the sequence $_{244}$SVVKSEASS$_{252}$[39,123], a sequence remarkably different to this 'best substrate' motif proposed. Such motif deviations hint at a more intricate involvement of the TPR domains of OGT in substrate specificity[53,124,125].

Although exceptions exist, sequence motifs can be defined wherein O-GlcNAcylation and phosphorylation can co-occur. Thus, crosstalk can be predicted by simply noting amino acid sequences within proteins that contain both the O-GlcNAcylation and phosphorylation motifs. Along these lines, Hart and co-workers developed software whereby sequences containing an O-GlcNAc motif were subjected to NetPhos, a computational software whereby phosphorylation sites are predicted based on the phosphorylation motifs of different kinases[126]. When a single residue had a positive hit for both O-GlcNAcylation and phosphorylation, i.e. where reciprocal crosstalk could occur, it was termed a Yin-Yang site[113,114]. This type of prediction works well, with Yin-Yang sites most commonly being predicted within PEST regions of proteins (peptide sequences rich in proline, glutamic acid, serine and threonine)[114]. However, since the motif for O-GlcNAcylation is somewhat 'fuzzy', this Yin-Yang software likely also identifies false positive and false negative crosstalk sites. Thus, efforts more recently have turned to search for a reciprocal O-GlcNAcylation/phosphorylation crosstalk motif. By combining all experimentally identified PTMs, 3 motifs were extracted for O-GlcNAcylation/phosphorylation crosstalk; Pxx[S], Txxx[S] and [T]xxxxxxxxxP, whereby O-GlcNAcylation and phosphorylation occur on the same Ser/Thr site[127]. Interestingly, proteins containing these motifs were enriched in specific gene ontology terms such as nuclear transport, cytoskeleton and structure molecular

activity, suggesting specific biological functions for PTM crosstalk[127]. Since many kinases are proline directed, a somewhat surprising observation came when Heck and coworkers observed that the presence of a proline residue at the P+1 position compared to the O-GlcNAcylation site hampers O-GlcNAcylation, suggesting no reciprocal O-GlcNAcylation/phosphorylation crosstalk can occur on sites targeted by proline-directed kinases[128]. This work hints that reciprocal crosstalk likely only occurs with specific non-proline directed kinases. However, there are notable exceptions. For example, c-Myc is O-GlcNAcylated at Thr58, a residue before a proline[95]. In addition, O-GlcNAcylation also occurs on the heptad repeats (YSPTSPS) in the C-terminal domain of RNA polymerase. This O-GlcNAcylation on RNA polymerase, however, only occurs efficiently when more than ~10 heptad repeats are present[129,130], thus factors in addition to sequence alone likely govern which proteins can undergo reciprocal O-GlcNAcylation/phosphorylation crosstalk.

Predicting crosstalk on proximate sites is significantly more challenging than reciprocal crosstalk since the observation of O-GlcNAc and phosphate on a peptide in close proximity does not necessarily mean that crosstalk is occurring. Moreover, there are likely multiple factors governing whether crosstalk exists such as the location of the O-GlcNAc/phosphorylation site with respect to the other, and the order in which the specific sequence is post-translationally modified. Additionally, as mentioned previously, both positive and negative crosstalk can occur. Aside from this, patterns whereby crosstalk occurs on adjacent sites are starting to emerge. However, defining a motif for crosstalk on adjacent sites will take more effort since each site of modification must first be exactly site-localized and then the nature of the crosstalk (positive/negative) determined in a controlled manner. Kinetic-based MS assays have helped accelerate this process[128]. Leney *et al.* noted that the most common O-GlcNAcylation motif also contained a kinase motif whereby the serine/threonine at the -3 position with respect to the O-GlcNAcylation site could be phosphorylated. Thus, the rate of O-GlcNAcylation with and without a phosphate residue at the -3 position (and vice versa, i.e. the rate of phosphorylation with and without an O-GlcNAc residue at the +3 position with respect to the phosphorylation site) was monitored using MS on several OGT substrate-mimicking peptides. In all cases, negative crosstalk was observed within the sequence [S/T]P[V/A/T][S/T]$X_{\text{-P}}$ where $X_{\text{-P}}$ represents any amino acid except proline. Indeed, this type of negative crosstalk is likely not uncommon with about one thousand sequences in the PhosphoSitePlus database containing precisely this putative crosstalk motif[128].

Future Directions

The majority of experiments that investigate phosphorylation/O-GlcNAcylation PTM crosstalk to date either focus on how one protein's PTM affects another protein's structure/function or look more globally at the cellular level on how phosphorylation affects O-GlcNAcylation or *vice versa*. In the case where

PTM crosstalk is monitored extensively on a single protein or a few proteins, the proteins of interest are made recombinantly or over-expressed in large quantities enabling biochemical experiments such as enzymatic assays or structural studies using X-ray crystallography or nuclear magnetic resonance to be carried out. These data together provide a wealth of information on how each specific PTM alters the proteins' structure alone and how this may be altered by PTM crosstalk. Moreover, this information is key to decipher how PTMs alter protein function. Indeed, there are no generic mechanisms on how a specific PTM alters protein function. For example phosphorylation can both activate (for example in the case of some kinases whereby an activation loop needs to be phosphorylated for the kinase to be active[131,132]) or inactivate proteins (for example in the case of Src tyrosine kinase whereby its phosphorylation results in a conformational change and its deactivation[133]). Although fundamentally important, expressing and characterizing each protein individually can be challenging and time consuming, and thus it will take decades to monitor each protein in the human genome. In addition, proteins act differently when removed from their cellular context where potential cofactors are absent. Thus, to accelerate the identification of PTMs and their roles in the context of the cell, we need to start to look at PTMs and PTM crosstalk at a system-wide level.

MS is an ideal method to detect multiple PTMs as each PTM can be uniquely detected based on a change in mass or fragmentation pattern[134,135]. Using this technology, thousands of PTMs can now be detected within a single experiment[136,137]. Unfortunately, though, the detection of some PTMs by MS is not straightforward. For example, on peptides the O-GlcNAc moiety is labile in the gas phase and thus O-GlcNAcylation went undetected for a long time before fragmentation techniques such as electron transfer dissociation and electron capture dissociation came about[138–140], as extensively reviewed[141–143]. This aside, PTMs are also dynamic modifications and are often present in low stoichiometric amounts hampering their analysis. To circumvent this, enrichment strategies are often employed. Thus, the development of methodology goes hand in hand with the identification of novel PTMs and PTM crosstalk sites. Also, enrichment methodologies employed in case of O-GlcNAcylation or phosphorylation have been extensively reviewed[108,144].

To detect instances of where positive PTM crosstalk might occur on adjacent sites, peptides would ideally be detected harboring both modifications simultaneously. Enrichment techniques, however, are currently specialized to target a single PTM of interest. Thus, the chances of finding both modifications especially if they have different chemical properties are slim. In addition, proteomics methods commonly use enzymes such as trypsin for digestion that produce relatively short peptides. Thus, relatively few peptides that encompasses sequences that could contain multiple PTM sites are identified. In addition, through digestion, even if multiple PTMs are detected, information is lost as to whether these PTMs were co-localized on the same protein molecule. Thus, scientists need to think about developing methodologies to not only

detect PTMs but that also aims to look specifically for PTM crosstalk.

To increase the chance of finding multiple PTMs on the same peptide and maintain information with regard to PTM co-occurrence, one could employ MS techniques that analyze longer peptides, protein subunits or even protein complexes[145–148]. In recent work by Sidoli *et al.*, middle-down proteomics was used to analyze histone peptides of roughly 60 amino acids in length, about three times larger than regular tryptic peptides. These longer peptides enable the researchers to reveal many instances of co-occurring methylations and assess whether these PTMs are observed on proximal sites more or less frequently than expected based on their abundance and thus whether the PTMs undergo interplay[149,150]. The analysis of intact proteins using top-down MS has also been successfully applied to help unravel histone code[151]. These techniques maintain valuable information on modifications co-occurring on the same protein, but also require more purified samples than required in traditional, bottom-up approaches.

These examples show that MS is a very powerful tool to detect multiple, chemically distinct PTMs simultaneously and it can even reveal PTM crosstalk by the detection of modifications co-occurring on the same protein, by relying on less conventional MS based techniques. The analysis of positive crosstalk, however, represents the simplest of cases. Negative PTM crosstalk adds yet another level of complexity as the lack of detection of one modification does not mean it does not exist. Hence, negative crosstalk can only be identified if the individual modifications are present at sufficient abundance in the sample and therefore it is currently the most challenging type of PTM crosstalk to study.

Summary

In conclusion, crosstalk between PTMs is a general phenomenon, which occurs on many different proteins and protein complexes[29,30,82]. Crosstalk can occur on different levels, ranging from the modification of writers and readers (i.e. enzyme level) to proximal and distal sites on modified proteins, resulting in astounding combinatorial PTM complexity. Modifications can influence each other positively (friends) or negatively (foes)[29,90]. Here, we have focused on phosphorylation and O-GlcNAcylation, briefly reviewed the regulation of these modifications and discussed examples of the different types of crosstalk observed between these PTMs. Finally, we have addressed the advancements in the prediction of crosstalk and highlighted techniques that may aid in the future search for PTM crosstalk and its functions. Overall, we believe that researchers should not only look to control the presence of a single PTM in the prevention of disease, but be mindful that looking elsewhere for other PTMs involved in crosstalk either in close proximity or distally to the PTM of interest could offer an alternative therapeutic target. Indeed, if one particular PTM is proving challenging to control, why not target another whose mechanism is

fully established?

Acknowledgements

This work was supported by the Roadmap Initiative Proteins@Work (project number 184.032.201), funded by the Netherlands Organisation for Scientific Research (NWO). AJRH acknowledges further support by the NWO TOP-Punt Grant 718.015.003 and the Spinoza Award SPI.2017.028.

References

1 Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA-K, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist P-H, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, Feilitzen K von, Forsberg M, Persson L, Johansson F, Zwahlen M, Heijne G von, Nielsen J & Pontén F (2015) Tissue-based map of the human proteome. Science (80-.). 347, 394–407.

2 Khoury GA, Baliban RC & Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci. Rep. 1, 1–5.

3 Jin J & Pawson T (2012) Modular evolution of phosphorylation-based signalling systems. Philos. Trans. R. Soc. B Biol. Sci. 367, 2540–2555.

4 Beltrao P, Bork P, Krogan NJ & van Noort V (2013) Evolution and functional cross-talk of protein post-translational modifications. Mol. Syst. Biol. 9, 714.

5 Nishi H, Shaytan A & Panchenko AR (2014) Physicochemical mechanisms of protein regulation by phosphorylation. Front. Genet. 5, 1–10.

6 Ciesla J, Fraczyk T & Rode W (2011) Phosphorylation of basic amino acid residues in proteins: important but easily missed. Acta Biochim. Pol. 58, 137–147.

7 Potel CM, Lin M-H, Heck AJR & Lemeer S (2018) Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based Proteomics. Nat. Methods, DOI: 10.1038/nmeth.4580.

8 Cohen P (2000) The regulation of protein function by multisite phosphorylation. Trends Biochem. Sci. 25, 596–601.

9 Filippakopoulos P, Muller S & Knapp S (2009) SH2 domains: modulators of nonreceptor tyrosine kinase activity. Curr. Opin. Struct. Biol. 19, 643–649.

10 Lu P-J, Zhou XZ, Shen M & Lu KP (1999) Function of WW Domains as Phosphoserine- or Phosphothreonine-Binding Modules. Science (80-.). 283, 1325–1328.

11 Holt LJ (2012) Regulatory modules: Coupling protein stability to phopshoregulation during cell division. FEBS Lett. 586, 2773–2777.

12 Gao Y, Wells L, Comer FI, Parker GL & Hart GW (2001) Dynamic O-Glycosylation of Nuclear and Cytosolic Proteins. J. Biol. Chem. 276, 9838–9845.

13 Milanesi L, Petrillo M, Sepe L, Boccia A, D'Agostino N, Passamano M, Nardo S Di, Tasco G, Casadio R & Paolella G (2005) Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity. BMC Bioinformatics 6, 1–11.

14 Brown NR, Noble MEM, Endicott JA & Johnson LN (1999) The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. Cell Biol. 1, 438–443.

15 Roskoski Jr R (2012) ERK1/2 MAP kinases: Structure, function, and regulation. Pharmacol. Res. 66, 105–143.

16 Tarbet HJ, Toleman CA & Boyce M (2018) A Sweet Embrace: Control of Protein-Protein Interactions by O-Linked β-N-Acetylglucosamine.

Biochemistry 57, 13–21.
17 Yang X & Qian K (2017) Protein O-GlcNAcylation: emerging mechanisms and functions. Nat. Rev. Mol. Cell Biol. 18, 452–465.
18 Zachara NE & Hart GW (2004) O-GlcNAc a sensor of cellular state: the role of nucleoplasmic glycosylation in modulating cellular function in response to nutrition and stress. Biochim. Biophys. Acta 1673, 13–28.
19 Lefebre T, Guinez C, Dehennaut V, Beseme-Dekeyser O, Morelle W & Michalski JC (2005) Does O-GlcNAc play a role in neurogenerative diseases? Expert Rev. Proteomics 2, 265–275.
20 Hardiville S & Hart GW (2014) Nutrient Regulation of Signaling, Transcription and Cell Physiology by O-GlcNAcylation. Cell Metab. 20, 208–213.
21 Adamczyk B, Tharmalingam T & Rudd PM (2012) Glycans as cancer biomarkers. Biochim. Biophys. Acta 1820, 1347–1353.
22 Apweiler R, Hermjakob H & Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim. Biophys. Acta 1473, 4–8.
23 Beurel E, Grieco SF & Jopea RS (2015) Glycogen synthase kinase-3 (GSK3): regulation, actions, and diseases. Pharmacol. Ther., 114–131.
24 Snider NT & Omary MB (2014) Post-translational modifications of intermediate filament proteins: mechanisms and functions. Nat. Rev. Mol. Cell Biol. 15, 163–177.
25 Segré C V. & Chiocca S (2011) Regulating the Regulators: The Post-Translational Code of Class I HDAC1 and HDAC2. J. Biomed. Biotechnol. 2011, 1–15.
26 Rust HL & Thompson PR (2011) Kinase Consensus Sequences: A Breeding Ground for Crosstalk. ACS Chem. Biol. 6, 881–892.
27 Benayoun BA & Veitia RA (2009) A post-translational modification code for transcription factors: sorting through a sea of signals. Trends Cell Biol. 19, 189–197.
28 Zhang T, Cooper S & Brockdorff N (2015) The interplay of histone modifications - writers that read. EMBO Rep. 16, 1467–1481.
29 Venne A, Kollipara L & Zahedi RP (2014) The next level of complexity: Crosstalk of posttranslational modifications. Proteomics 14, 513–524.
30 Jenuwein T & Allis CD (2001) Translating the Histone Code. Science (80-.). 293, 1074–1080.
31 Wang Z, Gucek M & Hart GW (2008) Cross-talk between GlcNAcylation and phosphorylation: Site-specific phosphorylation dynamics in response to globally elevated O-GlcNAc. PNAS 105, 13793–13798.
32 Sakabe K, Wang Z & Hart GW (2010) Beta-N-acetylglucosamine (O-GlcNAc) is part of the histone code. Proc. Natl. Acad. Sci. 107, 19915–19920.
33 Fujiki R, Hashiba W, Sekine H, Yokoyama A, Chikanishi T, Ito S, Imai Y, Kim J, He HH, Igarashi K, Kanno J, Ohtake F, Kitagawa H, Roeder RG, Brown M & Kato S (2011) GlcNAcylation of histone H2B facilitates its monoubiquitination. Nature 480, 557–560.
34 Xu Q, Yang C, Du Y, Chen Y, Liu H, Deng M, Zhang H, Zhang L, Liu T, Liu Q, Wang L, Lou Z & Pei H (2014) AMPK regulates histone H2B O-GlcNAcylation. Nucleic Acids Res. 42, 5594–5604.
35 Gagnon J, Daou S, Zamorano N, Iannantuono NV, Hammond-

Martel I, Mashtalir N, Bonneil E, Wurtele H, Thibault P & Affar EB (2015) Undetectable histone O-GlcNAcylation in mammalian cells. Epigenetics 10, 677–691.
36 Trinidad JC, Barkan DT, Gulledge BF, Thalhammer A, Sali A, Schoepfer R & Burlingame AL (2012) Global identification and characterization of both O-GlcNAcylation and phosphorylation at the murine synapse. Mol. Cell. Proteomics 11, 215–229.
37 Constable S, Lim JM, Vaidyanathan K & Wells L (2017) O-GlcNAc transferase regulates transcriptional activity of human Oct4. Glycobiology 27, 927–937.
38 Han J, Valdez J, Ho D, Lee C, Kim H, Wang X, Huang L & Chan J (2017) Nuclear factor-erythroid-2 related transcription factor-1 (Nrf1) is regulated by O-GlcNAc transferase. Free Radic. Biol. Med. 110, 196–205.
39 Myers S, Peddala S, Chatterjee N, Friedrich T, Tomoda K, Krings G, Thomas S, Maynard J, Broeker M, Thomson M, Pollard K, Yamanaka S, Burlingame AL & Panning B (2016) Sox2 O-GlcNAcylation alters its protein-protein interactions and genomic occupancy to modulate gene expression in pluripotent cells. Elife 5, 1–20.
40 Rexach JE, Clark PM & Hsieh-Wilson LC (2008) Chemical approaches to understanding O-GlcNAc glycosylation in the brain. Nat. Chem. Biol. 4, 97–106.
41 Wu D, Cai Y & Jin J (2017) Potential coordination role between O-GlcNAcylation and epigenetics. Protein Cell 8, 713–723.
42 Sohn KC, Lee KY, Park J & Do SL (2004) OGT functions as a catalytic chaperone under heat stress response: a unique defense role of OGT in hyperthermia. Biochem. Biophys. Res. Commun. 322, 1045–1051.
43 Ruan HB, Singh JP, Li MD, Wu J & Yang X (2013) Cracking the O-GlcNAc code in metabolism. Trends Endocrinol. Metab. 24, 301–309.
44 Bond MR & Hanover JA (2013) O-GlcNAc Cycling: A Link Between Metabolism and Chronic Disease. Annu. Rev. Nutr. 33, 205–229.
45 O'Donnell N, Zachara NE, Hart GW & Marth J. (2004) Ogt-Dependent X-Chromosome-Linked Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability. Mol. Cell. Biol. 24, 1680–1690.
46 Ma X, Li H, He Y & Hao J (2017) The emerging link between O-GlcNAcylation and neurological disorders. Cell. Mol. life Sci. 74, 3667–3686.
47 Banerjee P, Lagerlof O & Hart GW (2016) Roles of O-GlcNAc in chronic diseases of aging. Mol. Aspects Med. 51, 1–15.
48 Vercoutter-Edouart A-S, El Yazidi-Belkoura I, Guinez C, Baldini S, Leturcq M, Mortuaire M, Mir A-M, Steenackers A, Dehennaut V, Pierce A & Lefebvre T (2015) Detection and identification of O -GlcNAcylated proteins by proteomic approaches. Proteomics 15, 1039–1050.
49 Lazarus MB, Nam Y, Jiang J, Sliz P & Walker S (2011) Structure of human O-GlcNAc transferase and its complex with a peptide substrate. Nature 469, 564–567.
50 Love DC, Kochan J, Cathey R, Shin S & Hanover JA (2003) Mitochondrial and nucleocytoplasmic targeting of O-linked GlcNAc transferase. J. Cell Sci. 116, 647–654.
51 Seo HG, Kim HB, Kang MJ, Ryum JW, Yi EC & Cho JW (2016) Identification of the nuclear localization signal of

O-GlcNAc transferase and its nuclear import regulation. Sci. Rep. 6, 1–12.
52 Butkinaree C, Park K & Hart GW (2010) O-linked beta-N-acetylglucosamine (O-GlcNAc): Extensive crosstalk with phosphorylation to regulate signaling and transcription in response to nutrients and stress. Biochim. Biophys. Acta 1800, 96–106.
53 Rafie K, Raimi O, Ferenbach A, Borodkin V, Kapuroa V & van Aalten D (2017) Recognition of a glycosylation substrate by the O-GlcNAc transferase TPR repeats. Open Biol. 7, 1–9.
54 Iyer S & Hart G (2003) Roles of the tetratricopeptide Repeat Domain in O-GlcNAc Transferase Targeting and Protein Substrate Specificity. J. Biol. Chem. 278, 24608–24616.
55 Cheung W, Sakabe K, Housley M, Dias W & Hart G (2008) O-linked beta-N-acetylglucosaminetransferase Substrate Specificity Is Regulated by Myosin Phosphatase Targetting and Other Interacting Proteins. J. Biol. Chem. 283, 33935–33941.
56 Yang X, Zhang F & Kudlow J (2002) Recruitment of O-GlcNAc Transferase to Promotors by Corepressor mSin3A: Coupling Protein O-GlcNAcylation to Transcriptional Repression. Cell 110, 69–80.
57 Cheung W & Hart G (2008) AMP-activated Protein Kinase and p38 MAPK activate O-GlcNAcylation of Neuronal Proteins during Glucose Deprivation. J. Biol. Chem. 283, 13009–13020.
58 Iyer S, Akimoto Y & Hart G (2003) Identification and Cloning of a Novel Family of Coiled-coil Domain Proteins that Interact with O-GlcNAc Transferase. J. Biol. Chem. 278, 5399–5409.
59 Marz P, Stetefeld J, Bendfeldt K, Nitsch C, Reinstein J, Shoeman R, Dimitriades-Schmutz B, Schwager M, Leiser D, Ozcan S, Otten U & Ozbek S (2006) Ataxin-10 Interacts with O-linked beta-N-acetylglucosamine Transferase in the Brain. J. Biol. Chem. 281, 20263–20270.
60 Deng R, He X, Guo S, Liu W, Tao Y & Tao S (2014) Global Identification of O-GlcNAc Transferase (OGT) interactors by human proteome microarray and the construction of an OGT interactome. Proteomics 14, 1020–1030.
61 Zhang Z, Costa F, Tan E, Bushue N, DiTacchio L, Costello C, McComb M, Whelan S, Peterson K & Slawson C (2016) O-linked N-acetylglucosamine (O-GlcNAc) Transferase and O-GlcNAcase Interact with Mi2beta Protein at the gamma-Globulin Promotor. J. Biol. Chem. 291, 15628–15640.
62 Dong D & Hart GW (1994) Purification and Characterization of an 0-GlcNAc Selective N-Acetyl-pa-glucosaminidase from Rat Spleen Cytosol. J. Biol. Chem. 269, 19321–19330.
63 Zhang R, Zhou J, Song Z & Huang Z (2018) Enzymatic properties of β-N-acetylglucosaminidases. Appl. Microbiol. Biotechnol. 102, 93–103.
64 Macauley MS, Whitworth GE, Debowski AW, Chin D & Vocadlo DJ (2005) O-GlcNAcase uses substrate-assisted catalysis: kinetic analysis and development of highly selective mechanism-inspired inhibitors. J. Biol. Chem. 280, 25313–25322.
65 Comtesse N, Maldener E & Meese E (2001) Identification of a nuclear variant of MGEA5, a cytoplasmic hyaluronidase and a beta-N-acetylglucosaminidase. Biochem. Biophys. Res. Commun. 283, 634–640.
66 Rao F, Schuettelkopf A, Dorfmueller H, Ferenbach A, Navratilova I & van

Aalten D (2013) Structure of a bacterial putative acetyltransferase defines the fold of the human O-GlcNAcase C-terminal domain. Open Biol. 3, 1–8.
67 Butkinaree C, Cheung W, Park S, Park K, Barber M & Hart GW (2008) Characterization of beta-N-acetylglucosaminidase cleavage by caspase-3 during apoptosis. J. Biol. Chem. 283, 23557–23566.
68 Wells L, Gao Y, Mahoney J, Vosseller K, Chen C, Rosen A & Hart GW (2002) Dynamic O-glycosylation of nuclear and cytosolic proteins - Further characterization of the nucleocytoplasmic beta-N-acetylglucosaminidase, O-GlcNAcase. J. Biol. Chem. 277, 1755–1761.
69 Roth C, Chan S, Offen WA, Hemsworth GR, Willems LI, King DT, Varghese V, Britton R, Vocadlo DJ & Davies GJ (2017) Structural and functional insight into human O-GlcNAcase. Nat. Chem. Biol. 13, 610–612.
70 Elsen NL, Patel SB, Ford RE, Hall DL, Hess F, Kandula H, Kornienko M, Reid J, Selnick H, Shipman JM, Sharma S, Lumb KJ, Soisson SM & Klein DJ (2017) Insights into activity and inhibition from the crystal structure of human O-GlcNAcase. Nat. Chem. Biol. 13, 613–615.
71 Li B, Li H, Hu C-W & Jiang J (2017) Structural insights into the substrate binding adaptability and specificity of human O-GlcNAcase. Nat. Commun. 8, 666.
72 Dias W, Cheung W, Wang Z & Hart GW (2009) Regulation of Calcium/Calmodulin-dependent Kinase IV by O-GlcNAc Modification. J. Biol. Chem. 284, 21327–21337.
73 Tokumitsu H & Soderling TR (1996) Requirements for Calcium and Calmodulin in the Calmodulin Kinase Activation Cascade. J. Biol. Chem. 271, 5617–5622.
74 Chatila T, Anderson KA, Ho N & Means AR (1996) A Unique Phosphorylation-dependent Mechanism for the Activation of Ca2+/Calmodulin-dependent Protein Kinase Type IV/GR. J. Biol. Chem. 271, 21542–21548.
75 Tarrant MK, Rho HS, Xie Z, Jiang YL, Gross C, Culhane JC, Yan G, Qian J, Ichikawa Y, Matsuoka T, Zachara NE, Etzkorn FA, Hart GW, Jeong JS, Blackshaw S, Zhu H & Cole PA (2012) Regulation of CK2 by phosphorylation and O-GlcNAcylation revealed by semisynthesis. Nat. Chem. Biol. 8, 262–269.
76 Dias W, Cheung W & Hart GW (2012) O-GlcNAcylation of kinases. Biochem. Biophys. Res. Commun. 422, 224–228.
77 Shi J, Gu J, Dai C, Gu J, Jin X, Sun J, Iqbal K, Liu F & Gong C (2015) O-GlcNAcylation regulates ischemia-induced neuronal apoptosis through AKT signaling. Sci. Rep. 5, 1–14.
78 Risso G, Blaustein M, Pozzi B, Mammi P & Srebrow A (2015) Akt/PKB: one kinase, many modifications. Biochem. J. 468, 203–214.
79 Sarbassov DD, Guertin DA, Ali SM & Sabatini DM (2005) Phosphorylation and Regulation of Akt/PKB by the Rictor-mTOR Complex. Science (80-.). 307, 1098–1101.
80 Perez-Cervera Y, Dehennaut V, Aquino Gil M, Guedri K, Solórzano Mata CJ, Olivier-Van Stichelen S, Michalski J-C, Foulquier F & Lefebvre T (2013) Insulin signaling controls the expression of O-GlcNAc transferase and its interaction with lipid microdomains. FASEB J. 27, 3478–3486.
81 Sekine O, Love DC, Rubenstein DS & Hanover JA (2010) Blocking O-Linked GlcNAc Cycling in Drosophila

Insulin-producing Cells Perturbs Glucose-Insulin Homeostasis. J. Biol. Chem. 285, 38684–38691.
82 Kaasik K, Kivimäe S, Allen JJ, Chalkley RJ, Huang Y, Baer K, Kissel H, Burlingame AL, Shokat KM, Ptáček LJ & Fu Y-H (2013) Glucose sensor O-GlcNAcylation coordinates with phosphorylation to regulate circadian clock. Cell Metab. 17, 291–302.
83 Li Z, Li X, Nai S, Geng Q, Liao J, Xu X & Li J (2017) Checkpoint kinase 1–induced phosphorylation of O-linked β-N-acetylglucosamine transferase regulates the intermediate filament network during cytokinesis. J. Biol. Chem. 292, 19548–19555.
84 Bullen JW, Balsbaugh JL, Chanda D, Shabanowitz J, Hunt DF, Neumann D & Hart GW (2014) Cross-talk between Two Essential Nutrient-sensitive Enzymes. J. Biol. Chem. 289, 10592–10606.
85 Kreppel LK, Blomberg MA & Hart GW (1997) Dynamic Glycosylation of Nuclear and Cytosolic Proteins. J. Biol. Chem. 272, 9308–9315.
86 Whelan SA, Lane MD & Hart GW (2008) Regulation of O-linked beta-N-Acetylglucosamine Transferase by Insulin Signaling. J. Biol. Chem. 283, 21411–21417.
87 Lazarus BD, Love DC & Hanover JA (2006) Recombinant O-GlcNAc transferase isoforms: identification of O-GlcNAcase, yes tyrosine kinase, and tau as isoform-specific substrates. Glycobiology 16, 415–421.
88 Khidekel N, Ficarro SB, Clark PM, Bryan MC, Swaney DL, Rexach JE, Sun YE, Coon JJ, Peters EC & Hsieh-Wilson LC (2007) Probing the dynamics of O-GlcNAc glycosylation in the brain using quantitative proteomics. Nat. Chem. Biol. 3, 339–348.
89 Hart GW, Slawson C, Ramirez-Correa G & Lagerlof O (2011) Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. Annu. Rev. Biochem. 80, 825–858.
90 Hu P, Shimoji S & Hart GW (2010) Site-specific interplay between O-GlcNAcylation and phosphorylation in cellular regulation. FEBS Lett. 584, 2526–2538.
91 Torres CR & Hart GW (1984) Topography and Polypeptide Distribution of Terminal N-Acetylglucosamine Residues on the Surfaces of Intact Lymphocytes. J. Biol. Chem. 259, 3308–3317.
92 Dong D, Xu Z-S, Chevrier M, Cotter R, Cleveland D & Hart GW (1993) Glycosylation of Mammalian Neurofilaments. J. Biol. Chem. 268, 16679–16687.
93 Griffith LS & Schmitz B (1999) O-linked N-acetylglucosamine levels in cerebellar neurons respond reciprocally to pertubations of phosphorylation. FEBS J. 262, 824–831.
94 Wang Z, Pandey A & Hart GW (2007) Dynamic Interplay between O-Linked N-Acetylglucosaminylation and Glycogen Synthase Kinase-3-dependent Phosphorylation. Mol. Cell. Proteomics 6, 1365–1379.
95 Chou TY, Hart GW & Dang CV. (1995) c-Myc is glycosylated at threonine 58, a known phosphorylation site and a mutational hot spot in lymphomas. J. Biol. Chem. 270, 18961–18965.
96 Chou TY, Dang C V. & Hart GW (1995) Glycosylation of the c-Myc transactivation domain. Proc. Natl. Acad. Sci. USA 92, 4417–4421.
97 Kamemura K, Hayes BK, Comer FI & Hart GW (2002) Dynamic interplay between O-glycosylation and O-phosphorylation of nucleocytoplasmic proteins: alternative glycosylation/

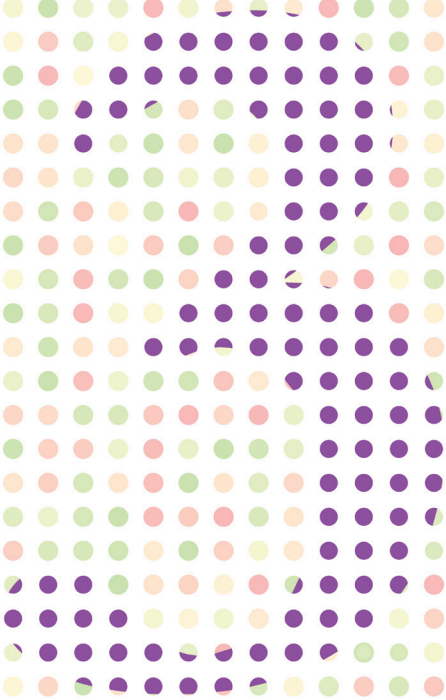
phosphorylation of THR-58, a known mutational hot spot of c-Myc in lymphomas, is regulated by mitogens. J. Biol. Chem. 277, 19229–19235.
98 Chen Y, Du J, Zhou L, Liu X, Zhao Y, Nakanishi H & Li Y (2006) Alternative O-GlcNAcylation/O-phosphorylation of Ser16 induce different conformational disturbances to the N terminus of murine estrogen receptor beta. Chem. Biol. 13, 937–944.
99 Yang WH, Kim JE, Nam HW, Ju JW, Kim HS, Kim YS & Cho JW (2006) Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. Nat. Cell Biol. 8, 1074–1083.
100 Ma Z, Chalkley RJ & Vosseller K (2017) Hyper- O -GlcNAcylation activates nuclear factor κ-light-chain-enhancer of activated B cells (NF-κB) signaling through interplay with phosphorylation and acetylation. J. Biol. Chem. 292, 9150–9163.
101 Mishra S, Ande SR & Salter NW (2011) O-GlcNAc modification: why so intimately associated with phosphorylation? Cell Commun. Signal. 9, 1.
102 Shi J, Tomasic T, Sharif S, Brouwer AJ, Anderluh M, Ruijtenbeek R & Pieters RJ (2017) Peptide microarray analysis of the cross-talk between O-GlcNAcylation and tyrosine phosphorylation. FEBS Lett. 591, 1872–1883.
103 Freund P, Kerenyi MA, Hager M, Wagner T, Wingelhofer B, Pham HTT, Elabd M, Han X, Valent P, Gouilleux F, Sexl V, Krämer OH, Groner B & Moriggl R (2017) O-GlcNAcylation of STAT5 controls tyrosine phosphorylation and oncogenic transcription in STAT5-dependent malignancies. Leukemia 31, 2132–2142.
104 Wang Z, Udeshi ND, Slawson C, Compton PD, Sakabe K, Cheung WD, Shabanowitz J, Hunt DF & Hart GW (2010) Extensive Crosstalk Between O-GlcNAcylation and Phosphorylation Regulates Cytokinesis. Sci. Signal. 3, 1–13.
105 Gual P, Le Marchand-Brustel Y & Tanti J-F (2005) Positive and negative regulation of insulin signaling through IRS-1 phosphorylation. Biochimie 87, 99–109.
106 Yang X, Ongusaha PP, Miles PD, Havstad JC, Zhang F, So WV, Kudlow JE, Michell RH, Olefsky JM, Field SJ & Evans RM (2008) Phosphoinositide signalling links O-GlcNAc transferase to insulin resistance. Nature 451, 964–969.
107 Aulak K, Barnes J, Tian L, Mellor N, Comhair S, Willard B & Dweik R (2017) Specific O-Glcnac Modification At Ser-615 Modulates Ser-1177 Phosphorylation Of Enos And Leads To Vascular Dysfunction In Pulmonary Arterial Hypertension. Am. J. Respir. Crit. Care Med. 195.
108 Dunn JD, Reid GE & Bruening ML (2010) Techniques for phosphopeptide enrichment prior to analysis by mass spectrometry. Mass Spectrom. Rev. 29, 29–54.
109 Riley NM & Coon JJ (2016) Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. Anal. Chem. 88, 74–94.
110 Lemeer S & Heck AJ (2009) The phosphoproteomics data explosion. Curr. Opin. Chem. Biol. 13, 414–420.
111 Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota SG & Pandey A (2007) A curated compendium of phosphorylation motifs. Nat. Biotechnol. 25, 285–286.
112 Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S & Brunak S (2004) Prediction of post-translational glycosylation and phosphorylation

of proteins from the amino acid sequence. Proteomics 4, 1633–1649.
113 Gupta R (2001) Prediction of glycosylation sites in proteomes: from post-translational modifications to protein function.
114 Gupta R & Brunak S (2002) Prediction of glycosylation across the human proteome and the correlation to protein function. Pac. Symp. Biocomput., 310–322.
115 Wang J, Torii M, Liu H, Hart GW & Hu Z-Z (2011) dbOGAP - An Integrated Bioinformatics Resource for Protein O-GlcNAcylation. BMC Bioinformatics 12, 91.
116 Jochmann R, Holz P, Sticht H & Stürzl M (2014) Validation of the reliability of computational O-GlcNAc prediction. Biochim. Biophys. Acta 1844, 416–421.
117 Jia C-Z, Liu T & Wang Z-P (2013) O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. Mol. Biosyst. 9, 2909–2913.
118 Zhao X, Ning Q, Chai H, Ai M & Ma Z (2015) PGlcS: Prediction of protein O-GlcNAcylation sites with multiple features and analysis. J. Theor. Biol. 380, 524–529.
119 Kao H-J, Huang C-H, Bretaña NA, Lu C-T, Huang K-Y, Weng S-L & Lee T-Y (2015) A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. BMC Bioinformatics 16 Suppl 1, S10.
120 Shi J, Sharif S, Ruijtenbeek R & Pieters RJ (2016) Activity Based High-Throughput Screening for Novel O-GlcNAc Transferase Substrates Using a Dynamic Peptide Microarray. PLoS One 11, e0151085.
121 Ortiz-Meoz RF, Merbl Y, Kirschner MW & Walker S (2014) Microarray discovery of new OGT substrates: the medulloblastoma oncogene OTX2 is O-GlcNAcylated. J. Am. Chem. Soc. 136, 4845–4848.
122 Pathak S, Alonso J, Schimpl M, Rafie K, Blair DE, Borodkin VS, Schuttelkopf AW, Albarbarawi O & van Aalten D (2015) The active site of O-GlcNAc transferase imposes constraints on substrate sequence. Nat. Struct. Mol. Biol. 22, 744–750.
123 Myers SA, Panning B & Burlingame AL (2011) Polycomb repressive complex 2 is necessary for the normal site-specific O-GlcNAc distribution in mouse embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. 108, 9490–9495.
124 Iyer SPN & Hart GW (2003) Roles of the tetratricopeptide repeat domain in O-GlcNAc transferase targeting and protein substrate specificity. J. Biol. Chem. 278, 24608–24616.
125 Lazarus MB, Jiang J, Kapuria V, Bhuiyan T, Janetzko J, Zandberg WF, Vocadlo DJ, Herr W & Walker S (2013) HCF-1 Is Cleaved in the Active Site of O-GlcNAc Transferase. Science (80-.). 342, 1235–1239.
126 Blom N, Gammeltoft S & Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. 294, 1351–1362.
127 Yao H, Li A & Wang M (2015) Systematic Analysis and Prediction of In Situ Cross Talk of O-GlcNAcylation and Phosphorylation. Biomed Res. Int. 2015, 279823.
128 Leney AC, El Atmioui D, Wu W, Ovaa H & Heck AJR (2017) Elucidating crosstalk mechanisms between phosphorylation and O-GlcNAcylation. PNAS, E7255–E7261.
129 Lu L, Fan D, Hu C-W, Worth M, Ma Z-X & Jiang J (2016) Distributive O-GlcNAcylation on the Highly

Repetitive C-Terminal Domain of RNA Polymerase II. Biochemistry 55, 1149–1158.
130 Comer FI & Hart GW (2001) Reciprocity between O-GlcNAc and O-phosphate on the carboxyl terminal domain of RNA polymerase II. Biochemistry 40, 7845–7852.
131 Adams JA (2003) Activation Loop Phosphorylation and Catalysis in Protein Kinases: Is There Functional Evidence for the Autoinhibitor Model? †. Biochemistry 42, 601–607.
132 Johnson LN, Noble ME & Owen DJ (1996) Active and inactive protein kinases: structural basis for regulation. Cell 85, 149–158.
133 Cole PA, Shen K, Qiao Y & Wang D (2003) Protein tyrosine kinases Src and Csk: a tail's tale. Curr. Opin. Chem. Biol. 7, 580–585.
134 Mann M & Jensen O (2003) Proteomic analysis of post-translational modifications. Nat. Biotechnol. 21, 255–261.
135 Witze E, Old W, Resing K & Ahn N (2007) Mapping protein post-translational modifications with mass spectrometry. Nat. Methods 4, 798–806.
136 Doll S & Burlingame A (2015) Mass spectrometry-based detection and assignment of protein posttranslational modifications. ACS Chem. Biol. 10, 63–71.
137 Olsen J & Mann M (2013) Status of large-scale analysis of post-translational modifications by mass spectrometry. Mol. Cell. Proteomics 12, 3444–3452.
138 Chalkley RJ, Thalhammer A, Schoepfer R & Burlingame AL (2009) Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. Proc. Natl. Acad. Sci. 106, 8894–8899.
139 Huang T-Y, Emory JF, O'Hair RAJ & McLuckey SA (2006) Electron-Transfer Reagent Anion Formation via Electrospray Ionization and Collision-Induced Dissociation. Anal. Chem. 78, 7387–7391.
140 Mirgorodskaya E, Roepstorff P & Zubarev RA (1999) Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. Anal. Chem. 71, 4431–4436.
141 Wang S, Yang F, Camp D, Rodland K, Qian W, Liu T & Smith R (2014) Proteomic approaches for site-specific O-GlcNAcylation analysis. Bioanalysis 6, 2571–2580.
142 You X, Qin H & Ye M (2018) Recent advances in methods for the analysis of protein o-glycosylation at proteome level. J. Sep. Sci. 41, 248–261.
143 Ma J & Hart G (2014) O-GlcNAc profiling: from proteins to proteomes. Clin. Proteomics 11, 1–8.
144 Worth M, Li H & Jiang J (2017) Deciphering the Functions of Protein O-GlcNAcylation with Chemisty. JACS Chem. Biol. 23, 326–335.
145 Riley NM & Coon JJ (2018) The Role of Electron Transfer Dissociation in Modern Proteomics. Anal. Chem. 90, 40–64.
146 van de Waterbeemd M, Lössl P, Gautier V, Marino F, Yamashita M, Conti E, Scholten A & Heck AJR (2014) Simultaneous Assessment of Kinetic, Site-Specific, and Structural Aspects of Enzymatic Protein Phosphorylation. Angew. Chemie Int. Ed. 53, 9660–9664.
147 Leney AC & Heck AJR (2017) Native Mass Spectrometry: What is in the Name? J. Am. Soc. Mass Spectrom. 28, 5–13.
148 Brunner AM, Lössl P, Liu F, Huguet R, Mullen C, Yamashita M, Zabrouskov V, Makarov A, Altelaar

AFM & Heck AJR (2015) Benchmarking Multiple Fragmentation Methods on an Orbitrap Fusion for Top-down Phospho-Proteoform Characterization. Anal. Chem. 87, 4152–4158.
149 Sidoli S, Lu C, Coradin M, Wang X, Karch KR, Ruminowicz C & Garcia BA (2017) Metabolic labeling in middle-down proteomics allows for investigation of the dynamics of the histone code. Epigenetics Chromatin 10, 34.
150 Sidoli S & Garcia BA (2017) Middle-down proteomics: a still unexploited resource for chromatin biology. Expert Rev. Proteomics 14, 617–626.
151 Zhou M, Paša-Tolić L & Stenoien DL (2017) Profiling of Histone Post-Translational Modifications in Mouse Brain with High-Resolution Top-Down Mass Spectrometry. J. Proteome Res. 16, 599–608.
152 Anderson KA, Means RL, Huang QH, Kemp B, Goldstein E, Selbert M, Fremeau R & Means A (1998) Components of a calmodulin-dependent protein kinase cascade. Molecular cloning, functional characterization and cellular localization of Ca2+/calmodulin-dependent protein kinase kinase beta. J. Biol. Chem. 273, 31880–31889.
153 Bosc DG, Slominski E, Sichler C & Litchfield D (1995) Phosphorylation of Casein Kinase II by p34cdc2. J. Biol. Chem. 270, 25872–25878.
154 Shaw RJ, Kosmatka M, Bardeesy N, Hurley RL, Witters LA, DePinho RA & Cantley LC (2003) The tumor suppressor LKB1 kinase directly activates AMP-activated kinase and regulates apoptosis in response to energy stress. PNAS 101, 3329–3335.
155 Higashimoto T, Chan N, Lee Y & Zandi E (2008) Regulation of IκB Kinase Complex by Phosphorylation of γ-Binding Domain of IκB Kinase β by Polo-like Kinase 1. J. Biol. Chem. 283, 35354–35367.
156 Kawauchi K, Araki K, Tobiume K & Tanaka N (2009) Loss of p53 enhances catalytic activity of IKKbeta through O-linked beta-N-acetyl glucosamine modification. Proc. Natl. Acad. Sci. USA 106, 3431–3436.
157 Bech-Otschir D, Kraft R, Huang X, Henklein P, Kapelari B, Pollmann C & Dubiel W (2001) COP9 signalosome-specific phosphorylation targets p53 to degradation by the ubiquitin system. EMBO 20, 1630–1639.
158 Welcker M, Orian A, Jin J, Grim JE, Harper JW, Eisenman RN & Clurman BE (2004) The Fbw7 tumor suppressor regulates glycogen synthase kinase 3 phosphorylation-dependent c-Myc protein degradation. Proc. Natl. Acad. Sci. USA 101, 9085–9090.
159 Sears R, Nuckolls F, Haura E, Taya Y, Tamai K & Nevins JR (2000) Multiple Ras-dependent phosphorylation pathways regulate Myc protein stabilit. Genes Dev. 14, 2501–2514.
160 Olivier-Van Stichelen S, Dehennaut V, Buzy A, Zachayus J-L, Guinez C, Mir A-M, El Yazidi-Belkoura I, Copin M-C, Boureme D, Loyaux D, Ferrara P & Lefebvre T (2014) O -GlcNAcylation stabilizes β-catenin through direct competition with phosphorylation at threonine 41. FASEB J. 28, 3325–3338.
161 Devaiah BN, Lewis BA, Cherman N, Hewitt MC, Albrecht BK, Robey RG, Ozato K, Sims RJ & Singer DS (2012) BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain. Proc. Natl. Acad. Sci. USA 109, 6927–6932.
162 Hsin J, Sheth A & Manley J (2011) RNAP II CTD phosphorylated on

threonine-4 is required for histone mRNA 3' end processing. Science (80-.). 334, 683–686.
163 Blazek D, Kohoutek J, Bartholomeeusen K, Johansen E, Hulinkova P, Luo Z, Cimerancic P, Ule J & Peterlin BM (2011) The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. Genes Dev. 25, 2158–2172.
164 Di Vona C, Bezdan D, Islam AB, Salichs E, Lopez-Bigas N, Ossowski S & de la Luna S (2015) Chromatin-wide Profiling of DYRK1A Reveals a Role as a Gene-Specific RNA Polymerase II CTD Kinase. Mol. Cell 57, 506–520.
165 Dimmeler S, Fleming I, Fisslthaler B, Hermann C, Busse R & Zeiher AM (1999) Activation of nitric oxide synthase in endothelial cells by Akt-dependent phosphorylation. Nature 399, 601–605.
166 Michell BJ, Harris MB, Chen Z, Ju H, Venema VJ, Blackstone MA, Huang W, Venema RC & Kemp BE (2002) Identification of Regulatory Sites of Phosphorylation of the Bovine Endothelial Nitric-oxide Synthase at Serine 617 and Serine 635. J. Biol. Chem. 277, 42344–42351.
167 Matsuzaki H, Daitoku H, Hatta M, Tanaka K & Fukamizu A (2003) Insulin-induced phosphorylation of FKHR (Foxo1) targets to proteasomal degradation. Proc. Natl. Acad. Sci. USA 100, 11285–11290.
168 Rena G, Guo S, Cichy S, Unterman T & Cohen P (1999) Phosphorylation of the transcription factor forkhead family member FKHR by protein kinase B. J. Biol. Chem. 274, 17179–17183.
169 Yun H, Park S, Kim M, Yang W, Im DU, Yang KR, Hong J, Choe W, Kang I, Kim SS & Ha J (2014) AMP-activated protein kinase mediates the antioxidant effects of resveratrol through regulation of the transcription factor FoxO1. FEBS J. 281, 4421–4438.
170 Housley M, Rodgers JT, Udeshi N, Kelly TJ, Shabanowitz J, Hunt DF, Puigserver P & Hart GW (2008) O-GlcNAc Regulates FoxO Activation in Response to Glucose. J. Biol. Chem. 283, 16283–16292.
171 Fardini Y, Perez-Cervera Y, Camoin L, Pagesy P, Lefebvre T & Issad T (2015) Regulatory O-GlcNAcylation sites on FoxO1 are yet to be identified. Biochem. Biophys. Res. Commun. 462, 151–158.
172 Klein AL, Berkaw MN, Buse MG & Ball LE (2009) O-linked N-acetylglucosamine modification of insulin receptor substrate-1 occurs in close proximity to multiple SH2 domain binding motifs. Mol. Cell. Proteomics 8, 2733–2745.
173 Jahangir Z, Ahmad W & Shabbiri K (2014) Alternate Phosphorylation/O-GlcNAc Modification on Human Insulin IRSs: A Road towards Impaired Insulin Signaling in Alzheimer and Diabetes. Adv. Bioinformatics 2014, 1–18.



Chapter 3

Targeting Proline in (Phospho)Proteomics

Saar A.M. van der Laarse[a,b,†], Charlotte A.G.H. van Gelder[a,b,†], Marshall Bern[c], Michiel Akeroyd[d], Maurien M.A. Olsthoorn[d], and Albert J.R. Heck[a,b]

a Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
b Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands
c ProteinMetrics, USA
d DSM Biotechnology Center, PO Box 1, 2600 MA Delft, the Netherlands

† These authors contributed equally

Published in The FEBS Journal, Volume 287, Issue 14, 2018, pages 2979-2997
DOI: 10.1111/febs.15190

Abstract

Mass spectrometry-based proteomics experiments typically start with the digestion of proteins using Trypsin, chosen because of its high specificity, availability and ease of use. It has become apparent that the sole use of Trypsin may impose certain limits on our ability to grasp the full proteome, missing out particular sites of post-translational modifications, protein segments or even subsets of proteins. To tackle this problem, alternative proteases have been introduced and shown to lead to an increase in the detectable (phospho) proteome. Here we argue that there may be further room for improvement and explore the protease EndoPro. For optimal peptide identification rates we explored multiple peptide fragmentation techniques (HCD, ETD and EThcD) and employed Byonic as search algorithm. We obtain peptide IDs for about 40% of the MS2 spectra (66% for Trypsin). EndoPro cleaves with high specificity at the C-terminal site of Pro and Ala residues and displays activity in a broad pH range, where we focused on its performance at pH = 2 and pH = 5.5. The proteome coverage of EndoPro at these two pH values is rather distinct, and also complementary to the coverage obtained with Trypsin. As about 40% of mammalian protein phosphorylations are proline-directed, we also explored the performance of EndoPro in phosphoproteomics. EndoPro extends the coverable phosphoproteome substantially, whereby both the at pH = 2 and pH = 5.5 acquired phosphoproteomes are complementary to each other and to the phosphoproteome obtained using Trypsin. Hence, EndoPro is a powerful tool to exploit in (phospho)proteomics applications.

Introduction

Proteins are involved in nearly all biological processes. Their functionality can be regulated extensively, through the formation of complexes, changes in expression levels, and widespread posttranslational modifications, such as acetylation and phosphorylation. Proteins must be tightly regulated as undesired changes at the protein level can cause disease and other unintended biological effects[1,2]. Owing to their crucial role, identifying and quantifying proteins present in certain biological states is of great importance and can improve our understanding of the mechanisms underlying health and disease. To this end, the field of proteomics aims to measure all proteins expressed by a certain organism or cell type[3,4]. Proteomics comes in a range of different workflows[3]. In the more standard bottom-up workflow, proteins are extracted from the material of interest and subjected to proteolysis, which results in a complex mixture of peptides that originated from the proteins present in the targeted cells. Through LC-MS/MS analysis, these peptides are separated, fragmented and analyzed. Then, the collected empirical spectra are correlated to peptide and thus protein sequences through the use of theoretical in silico fragmentation spectra[5,6]. However, due to the indirect nature of this assignment, how peptides are generated from the intact protein is of critical importance.

Most proteomic studies use Trypsin for the protein digestion as it cleaves with very high specificity proteins C-terminally to only Arginine (Arg) and Lysine (Lys) residues. As both amino acids are basic, the resulting peptides have basic C-termini. This, combined with the free amine at the peptide N-terminus, ensures that tryptic peptides carry a positive charge at either end of the peptide, making them very suitable for fragmentation-based sequencing[6]. In addition, Trypsin's high specificity reduces the complexity of the subsequent database searches as they can be restricted to peptides ending with Arg or Lys, which reduces computational requirements of the search. However, the use of Trypsin also has limitations and is not the optimal enzyme for all analyses. Owing to the high specificity of Trypsin, the spacing of Arg and Lys amino acids across the proteome dictate the length of peptides, and thus the number of unique peptides. For standard intracellular proteins, Arg and Lys occur at a high frequency (5.6% and 5.7%, respectively) which leads to the fact that roughly 50% of the peptides produced by Trypsin are too short (<6 amino acids) to be nicely fragmented and uniquely assigned to a protein[7]. Conversely, some proteins, notably membrane proteins, exhibit few tryptic cleavage sites and extreme hydrophobicity, resulting in poor coverage of this class of proteins in Trypsin based proteomics[8]. These combined effects all contribute to undetected, less visible areas of the proteome. To illustrate this, we have performed an *in silico* digestion of the human proteome using the specificity listed in Table 1, and asked what the upper limit of detection was for each protease using the search and mass spectrometry settings employed in this study (Table 1). For Trypsin, a maximum of 87% of the proteome would in theory be detectable using this proteomics setup, assuming every peptide

of suitable characteristics is actually fragmented and identified. To improve on this boundary, efforts have been made to utilize different proteases within bottom-up workflows. Several groups have shown that by using proteases that cleave at different amino acid motifs, the number of unique peptides identified, and thus the proteome coverage, can be substantially improved[7,9–12].

Table 1 – Cleavage specificities reported for some of the most commonly used proteases in bottom up proteomics[9,10,13–16]. X_{np} indicates any amino acid except proline.

Protease	Cleavage site (|)	Optimal pH	Max Proteome Coverage
Asp-N	|D	8	78%
LysargiNase	|K/R	7.5	87%
Lys-C	K|X_{np}	8	79%
Lys-N	|KX_{np}	8	78%
Arg-C	R|X_{np}	8	82%
Glu-C	E|	8	86%
Chymotrypsin	F/W/Y|	8	87%
Trypsin	K/R|X_{np}	8-9	87%
Sap-9	K/R|X_{np}	6-7	87%

Numerous alternative proteases have been used for the digestion of proteins from a lysate, whereby each has its own cleavage specificity and optimal conditions (Table 1). By combining the proteases either in parallel or sequentially, one is able to improve the proteome coverage through combining the results of individual proteases together. For instance, work by Swaney *et al.* nicely illustrated that expanding beyond a single protease can yield a roughly 20% increase in protein identifications and achieved double the proteome sequence coverage[7]. Similarly, our group has shown that the use of multiple proteases in parallel for phosphoproteomics gives rise to highly complementary sets of phosphosites, where only 27% of all identified sites were found in more than one protease dataset[17].

While the combination of proteases has already been shown to aid in expanding the proteome sequence coverage, the presence of (multiple) proline residues presents a particular challenge for many proteases. Proline is a unique amino acid in peptides/proteins as it is the only cyclic amino acid, giving rise to a tertiary amide, limiting hydrogen donating properties and imposing rigid structural constraints on peptide bonds [18,19]. Because of its unique properties, proline often leads to missed proteolytic events during digestion[9], increasing the resultant peptide length and database search complexity. Moreover, proline also effects the fragmentation step during mass analysis, known as the "proline effect"[20], where fragmentation shows enhanced production of y-ions spanning from the proline to the peptide C-terminus due to the enhanced basicity of the proline nitrogen, restricting the peptide sequence coverage[20–22].

To overcome these limitations, research efforts have been directed toward finding a proline-directed protease, as such a protease would decrease database search complexity by well-defining the proline position, as well as substantially improve proteome sequence coverage due to its high complementarity to Arg and Lys directed proteases. In 2009, Šebela *et al.* evaluated an acidic prolyl endoprotease from Aspergillus niger, called An-PEP, for its use in proteomics and found that the enzyme has potential for in-solution digestion studies[23]. Moreover, our lab showed that An-PEP, also termed EndoPro, exhibited maximum activity at pH = 2 and is active at moderately high urea concentrations and low temperatures, making it very suitable for use in mass spectrometry based hydrogen-deuterium exchange experiments[24]. In addition, work published on another prolyl endopeptidase originally from Nepenthes ventrata, termed Neprosin, showed that almost half of the sequence coverage achieved by the proline-directed protease on proteins detected in both tryptic and Neprosin digests were not observed when digestion was performed with Trypsin[25]. Collectively, these works suggest huge potential of proline-directed proteases to shed light on previously undetectable areas of the proteome. In phosphoproteomics, however, proline-induced complications are even more prevalent as in eukaryotic systems around 40% of the phosphorylation events detected are proline-directed, dominated by so-called SP or TP motifs[26]. Hence, in most eukaryotic phosphoproteomics experiments, prolines are highly enriched and even more prevalent than in a standard proteomics analysis.

Here we extend substantially on previous work using proline-directed proteases. We first benchmark EndoPro versus Trypsin, thereby generating large proteomics datasets on HeLa lysates digested by EndoPro at pH = 2, EndoPro at pH = 5.5 and Trypsin at pH = 8.5. We optimize the peptide ID rates using multiple peptide fragmentation techniques, and the search engine Byonic, allowing us to increase the ID rate substantially to about 40% and 66% of all PSMs for EndoPro and Trypsin, respectively. When using EndoPro at these two different pH values, we find the specificities and activities to be similar. However, our datasets reveal a substantial difference between the peptides generated with EndoPro at pH = 2, EndoPro at pH = 5.5 and Trypsin, indicating the cleavage of different proteins and/or sites at different pH values. Overall, EndoPro enabled us to detect over 2200 unique proteins not observed in our tryptic digests and contributed 49% of the total unique phosphosites detected, making it a protease almost equally powerful as, and complementary, to Trypsin.

Materials and Methods

In silico proteome coverage

Human proteins deposited in the Swissprot database (20,417 reviewed proteins, downloaded July 25th, 2019) were digested in silico using the specificity requirements listed in Table 1. Zero, one or two missed cleavages were allowed for each peptide, resulting in a database with all possible peptides formed by each of the nine listed proteases. Subsequently, these peptides were filtered on precursor m/z ($375 \leq m/z \leq 1500$), mass ($m \leq 10000$ Da) and only fully specific peptides were taken into account. All peptides passing these filters were mapped to the proteome to find the theoretical upper limit of proteome coverage possible.

Cell culture

HeLa cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum and 10 mM glutamine (Lonza) at 37°C/5%CO2. One hour prior to harvesting the medium was refreshed to stimulate phosphorylation. Cells were washed with ice cold PBS and cell pellets were collected by mild centrifugation (1,200 rpm) for three minutes and stored at -80°C until lysis.

Sample preparation

Cell pellets were lysed in a boiling lysis buffer containing 6 M guanidinium HCl (GuHCl), 5 mM tris(2-carboxyethyl)phosphine (TCEP), 10 mM chloroacetamide, 100 mM tris-HCl pH 8.5, supplemented with protease inhibitor (cOmplete mini EDTA-free, Roche). Pellets were boiled for 10 min at 99 °C, sonicated for 30 rounds of 5 seconds (Bioruptor Plus, Diagenode), and spun down at 20,000 x g for 15 min. Protein concentration was determined using Pierce™ BCA protein assay kit. Equal amounts of protein per condition were diluted to a final concentration of 2 M GuHCl, and pH was adjusted to pH = 2 and pH = 5.5 for EndoPro, or pH = 8.5 for Trypsin, using formic acid (FA) (Merck). Finally, proteins were digested with EndoPro (1:100, DSM) or Trypsin (1:100, Sigma Aldrich) for 1 hour or overnight at 37 °C. The resulting peptides were acidified to a final concentration of 1% FA, cleaned up using Sep-Pak cartridges (Waters), and dried in vacuo.

Phosphopeptide enrichment

Phosphorylated peptides were enriched using Fe(III)-NTA cartridges (Agilent technologies) in an automated fashion using the AssayMAP Bravo Platform (Agilent technologies)[27]. The cartridges were primed with 0.1% TFA in ACN and equilibrated with loading buffer (80% ACN/0.1% TFA). Samples were suspended in loading buffer and loaded onto the cartridge. The peptides bound to the cartridges were washed with loading buffer and the phosphorylated peptides were eluted with 1% ammonia directly into 10% formic acid. Samples were dried in vacuo and stored at -80 °C until LC-MS/MS analysis.

LC-MS/MS analysis

Peptide samples were resuspended in 20 mM Citric Acid with 2% FA and analyzed with an UHPLC 1290 system (Agilent technologies) coupled to an Orbitrap Fusion mass spectrometer (Thermo Fischer Scientific). Peptides were trapped (Dr Maisch Reprosil C18, 3 µm, 2 cm x 100 µm) and then separated on an analytical column (Agilent Poroshell EC-C18, 2.7 µm, 50 cm x 75 µm). All columns were made in-house. Trapping was performed for 5 min in solvent A (0.1% FA), followed by a gradient of the following: 0–8% solvent B (0.1% FA in 80% ACN) in 10 s, 8–32% in 100 min, 32–100% in 5 min, hold for 5 min, 100–0% in 1 min, and hold for 4 min. Flow was passively split to 300 nL/min.

The mass spectrometer was operated in data-dependent mode. Full scan MS spectra from m/z 375–1500 were acquired at a resolution of 60 000 after accumulation to a target value or 4e5 or a maximum injection time of 50 ms. The most intense precursor ions were selected for fragmentation for a duration of 3 s with a 24 s dynamic exclusion duration. Target peaks were isolated in a 1.6 Da isolation window and subjected to either higher-energy collision-induced dissociation (HCD), electron-transfer dissociation (ETD), or electron-transfer higher-energy collision-induced dissociation (EThcD)[28]. MS/MS spectra were acquired with a resolution of 30,000 using an AGC target of 1e5 ions with a maximum injection time of 125 ms. Charge state screening was enabled, and precursors with an unknown charge state or a charge state of 1+ were excluded. For the decision tree strategy, HCD and EThcD fragmentation were performed with normalized collision energies of 35% and 40% respectively. Fragmentation was done based on charge state. HCD was selected for peptide ions with charge states of 2+ and 3+; and for EThcD, charge states 4+ to 20+ were selected.

Data analysis

The resulting mass spectra were searched using Byonic (Protein Metrics Inc., v.3.3.11) in a fully non-specific or semi-specific search (C-terminal cleavage on Arg/Lys or Ala/Pro for Trypsin or EndoPro, respectively). The number of missed cleavages was not restricted. Mass tolerance was set at 10 ppm and 20 ppm for precursor and fragment ions, respectively. Carbamidomethylation was set as a fixed cysteine modification, oxidation of methionine, deamidation of asparagine and sodium adducts of aspartate, glutamate, serine and threonine were set as common modifications. The formation of pyro-glutamine from N-terminal glutamine or glutamate, loss of ammonia and acetylation were set as rare modifications. Overall, one common and one rare modification were allowed in the standard bottom-up workflow. For the phospho-enriched peptides, phosphorylation on serine or threonine was included as a common modification and in total 3 common and 1 rare modification were allowed.

Using Byonic Viewer (Protein Metrics Inc., v.3.3-421) the PSMs were filtered by a PEP 2D < 0.001 resulting in a 0.1% PSM level FDR. In the phospho-enriched dataset, we also asked that the delta mod. score (dmod) was larger than 20, to only include the more confident phosphosite localizations for each PSM. The

resulting PSMs from different fragmentation methods were combined prior to further data analysis.

Peptide and/or protein characteristics such as peptide length, charge, amino acid content and location of phosphosites on the peptide were determined using in-house R scripts (available upon request), Venn diagrams were made using both Venny (BioinfoGP v.2.1.0)[29] and BioVenn[30] and the bar graphs, boxplots and heatmaps were visualized using GraphPad Prism 8.0.1. IceLogo was generated as described in Colaert et al.[31].

In Figure 4D, the phosphorylation motifs were assigned to one of four categories: proline-directed phosphosites p(S/T)P, Acidophillic (D/E after the phosphosite), Basophillic (R/K before phosphosite) or other. Assignment was hierarchical, meaning that a phosphosite exhibiting both proline-directed and basophilic characteristic was only included in the proline-directed group.

Results

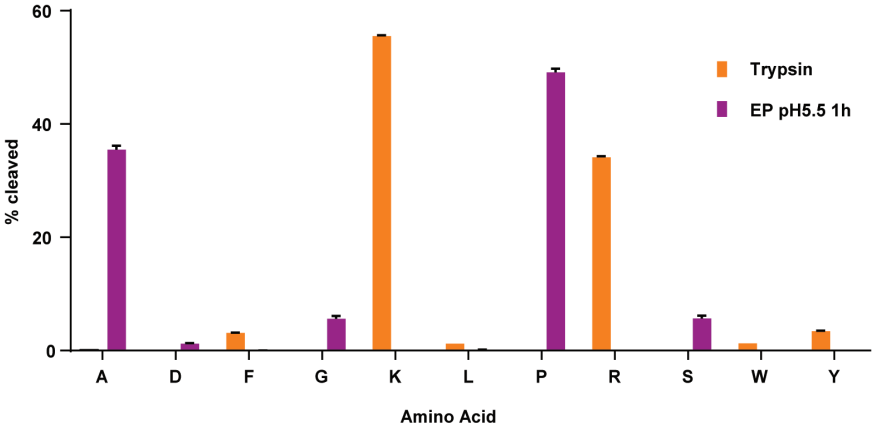
To assess the benefit of EndoPro in shotgun proteomics[25], we decided to evaluate and optimize the performance of this proline-directed protease on a complex HeLa cell lysate, first focusing on non-modified peptides. To characterize the specificity of EndoPro in a full proteome, we performed a quadruplicate digestion of a HeLa lysate with EndoPro for 1hr at pH = 5.5. For comparison, we also performed a standard Trypsin overnight digestion at pH = 8.5. These data were subjected to a non-specific search in Byonic and subsequently the environment of all cleaved sites was analyzed using an in-house R script. The distribution of amino acids following the residue cleaved by either EndoPro (purple) or Trypsin (orange) is shown in Figure 1A. EndoPro showed a strong specificity for cleavage C-terminal to proline (49.1%) and alanine (35.5%), resulting in an overall cleavage specificity close to that observed for Trypsin (84.6% Ala/Pro versus 89.6% Arg/Lys in our datasets). Inspection of the cleavage site environment of EndoPro (Figure 1B) revealed a disfavor for cleaving when the cleavage site is preceding a proline. In these cases, only the last proline is cleaved. In addition, positively charged residues appear disfavored in the P+2 position (Figure 1B).

Since EndoPro reached almost 85% specificity, we subsequently used less computationally heavy semi-specific database searches (allowing one side of the peptide to result from non-specific cleavages) which saves data analysis time and is inherently less error prone. Doing these two searches on the same dataset we observed that we still captured nearly all the peptides formed (97.6%).

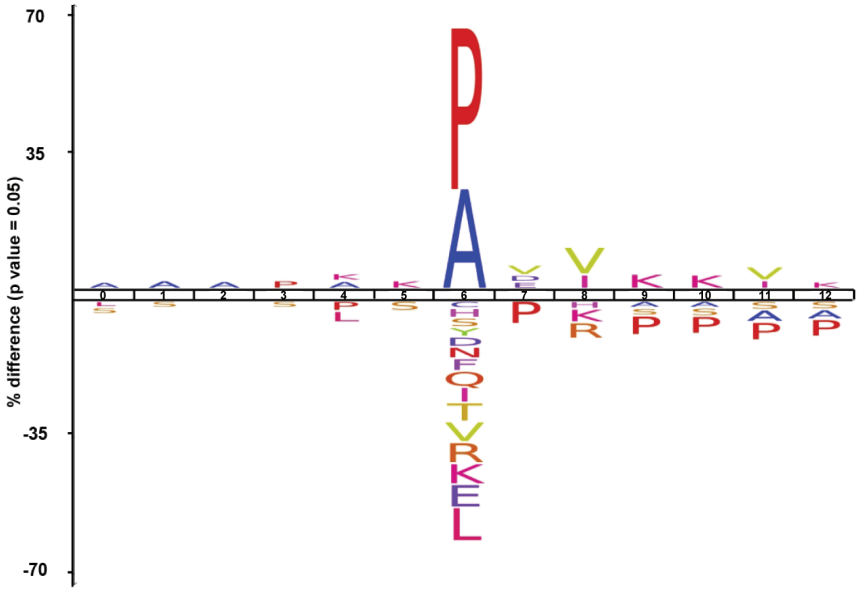
Performance evaluation of EndoPro at pH = 2 and pH = 5.5

Next, we set out to compare the performance of EndoPro and Trypsin. Thereby we took into account that EndoPro exhibits several maxima in its activity profile, with maxima at pH = 2 and 5.5, as also reported earlier[24]. Therefore, HeLa cell lysates were digested with either EndoPro at pH = 2 and pH = 5.5, and digested for 1 hour or over-night (ON), and additionally, for benchmarking, the same HeLa cell lysate sample was digested with Trypsin using conventional conditions (i.e. pH = 8.5, ON). The resulting peptides were analyzed by LC MS/MS on a Fusion Tribrid Mass spectrometer using in parallel ETD, EThcD and HCD as peptide fragmentation methods. Spectra were searched with Byonic. A global overview of the search outcomes is shown in Table S1. The different fragmentation methods resulted in a highly similar number of protein identifications (Table S1). It was therefore decided to pool all the data acquired with different fragmentation techniques to assess the performance of EndoPro across the different digestion conditions, independent of the used fragmentation method. An overview of the pooled datasets is shown in Table 2. In terms of unique proteins detected, an overlap of 35% was observed between the two proteases (Figure 1C). The four different EndoPro digestion conditions resulted in comparable identification rates, with slightly more PSMs and unique peptides in the experiments performed at pH = 5.5

A


60
40
20
0
% cleaved
Trypsin
EP pH5.5 1h
A
D
F
G
K
L
P
R
S
W
Y
Amino Acid

B


70
35
-35
-70
% difference (p value = 0.05)
0
1
2
3
4
5
6
7
8
9
10
11
12

C


EndoPro (4715)
2204
30%
2511
35%
2525
35%
Trypsin (5036)

Figure 1 – Characterization of EndoPro cleavage specificity. (A) Overview of amino acids after which was cleaved by EndoPro (n=4, purple) and Trypsin (n=4, orange), based on a nonspecific search, revealing a high specificity of 84.6% A/P and 89.6% R/K for EndoPro and Trypsin, respectively. Only amino acids with a cleavage frequency of 1% or higher were included. Data are represented as mean percentage of total cleavages per protease ± SEM. (B) An Icelogo showing the differences between the EndoPro cleavage site environment (17032 unique environments from nonspecific search) and the human proteome, illustrating a disfavor for R/K on the +2 position and a reluctance to cleave between proline residues. (C) Overlap of unique proteins identified by EndoPro or Trypsin using a semispecific search. Although the sizes of the identified proteomes are roughly equal, the overlap between the two is only 35%.

when compared to pH = 2. Under all four tested conditions roughly the same number of peptides (≈15000) and proteins (≈2600) could be identified (Table 2). As expected, the peptide identification rate achieved with Trypsin (67%) could not be reached with EndoPro (≈40%). Still the EndoPro ID rates of ≈40% are better than what has been reported for many other alternative enzymes (e.g. LysN, AspN, Chymotrypsin typically reach 20-30%)[10,24,25].

Table 2 – Characteristics of measured and analyzed EndoPro and Trypsin datasets.

				Byonic semi-specific search				
Protease	pH	Digestion time	Fragmentation	# MS2 scans	#PSMs 0.1 FDR	# unique peptides	# unique proteins	% identification
Trypsin	8.5	ON	ETD/EThcD/ HCD	163823	109682	35330	5036	67%
EndoPro	2	1 hr	ETD/EThcD/ HCD	152064	54251	13631	2633	36%
EndoPro	2	ON	ETD/EThcD/ HCD	151115	57722	15264	2439	38%
EndoPro	5.5	1 hr	ETD/EThcD/ HCD	155229	68902	18268	2810	44%
EndoPro	5.5	ON	ETD/EThcD/ HCD	155565	60318	17378	2621	39%
EndoPro cumulative				613973	241193	38004	4715	39%

Characteristics of EndoPro peptides generated at pH = 2 and pH = 5.5

For a more in-depth exploration of the type of peptides produced by EndoPro, we compared general peptide characteristics such as peptide length, mass and amino acid content and cleavage specificity as observed in the unique peptides identified from the EndoPro and, for comparison, tryptic digests, as depicted in Figure 2. In terms of peptide length and charge, the four explored different EndoPro cleavage conditions produced similar peptides. We observed a substantially broader peptide length distribution for EndoPro peptides than for Trypsin, revealing more peptides with a length of more than 20 amino acids and a tail towards peptides with a length of 50 or more amino acids (Figure 2A). This already indicates that EndoPro generates peptides with more missed cleavages than Trypsin. In terms of peptide charge, an average of about 33% of the unique EndoPro peptides carried four or more charges, compared to only 5% for the tryptic peptides (Figure 2B). This difference in charge distribution could not be explained by the increase in peptide length, as the average number

of amino acids to charge ratio of the different EndoPro conditions was lower than we found for Trypsin (5.75 and 6.22 amino acids per positive charge, respectively).

The identified unique peptide length and number of charges found after EndoPro digestions did not vary much with digestion time. In contrast, the amino acid content of the peptides as well as the cleavage specificity of EndoPro appeared to be sensitive to the digestion conditions. In total, four amino acids (alanine, aspartic acid, glutamic acid and proline) showed a substantial change in abundance when comparing EndoPro digests prepared at pH = 2 and pH = 5.5 (Figure 2C). With the increase in pH, the contribution of alanine and proline to the total amino acid content of the peptides decreased, whereas the contribution of the negatively charged aspartic acid and glutamic acid increased. Although clearly visible after 1hr of digestion, this effect is even more pronounced after ON digestion with EndoPro, where the proline content of the peptides at pH = 5.5 decreased to less than half of the value observed at pH = 2. The cleavage specificity of EndoPro also slightly decreased with increasing pH (Figure 2D). Interestingly, the location of Asp on the peptides also changed with pH (Figure 2E,F). This indicates that a different set of peptides is generated, depending on the digestion condition used.

The complete overview of amino acid content of the peptides generated by EndoPro under the four evaluated conditions and Trypsin is shown in Figure S1. For reference, the natural occurrence of each amino acid within the human proteome is indicated with a dashed line. Due to the Arg/Lys specific cleavage by Trypsin, these tryptic peptides clearly underrepresent the abundance of Arg/Lys in the human proteome. Peptides generated by EndoPro do not impose limits on the number of Arg/Lys residues and hence are richer in these positively charged residues, which is in agreement with the on average higher charges we observe for EndoPro peptides. In addition, at low pH these basic amino acids carry a positive charge, which may help to prevent aggregation and therefore aid protein solubility. Similarly, the observed increase in Asp/Glu content with pH may also be related to their charge, as the presence of negatively charged amino acids have been correlated to an increase in solubility[32]. At pH = 2, virtually none of the carbonic acid side chains will be negatively charged due to the excess in protons. At pH = 5.5, however, these amino acids would be predominantly negatively charged and essentially all would be charged at pH = 8.5. Therefore, this Asp/Glu rich subset of the proteome may have a better solubility over other proteins at increasing pH, which could explain why they are more abundantly represented on the peptide level.

As indicated above, we found the peptide dataset generated by EndoPro to be sensitive to both the cleavage time and pH (Figure 2D). This could be partly attributed to more subtle changes in specificity. After 1 hour digestion at pH = 2, 26% of the peptides were cleaved after alanine and 68% of the cleavages were proline specific. This decreased slightly to 25% alanine and 62% proline-specific cleavages following ON digestion. At pH = 5.5, however, the percentage

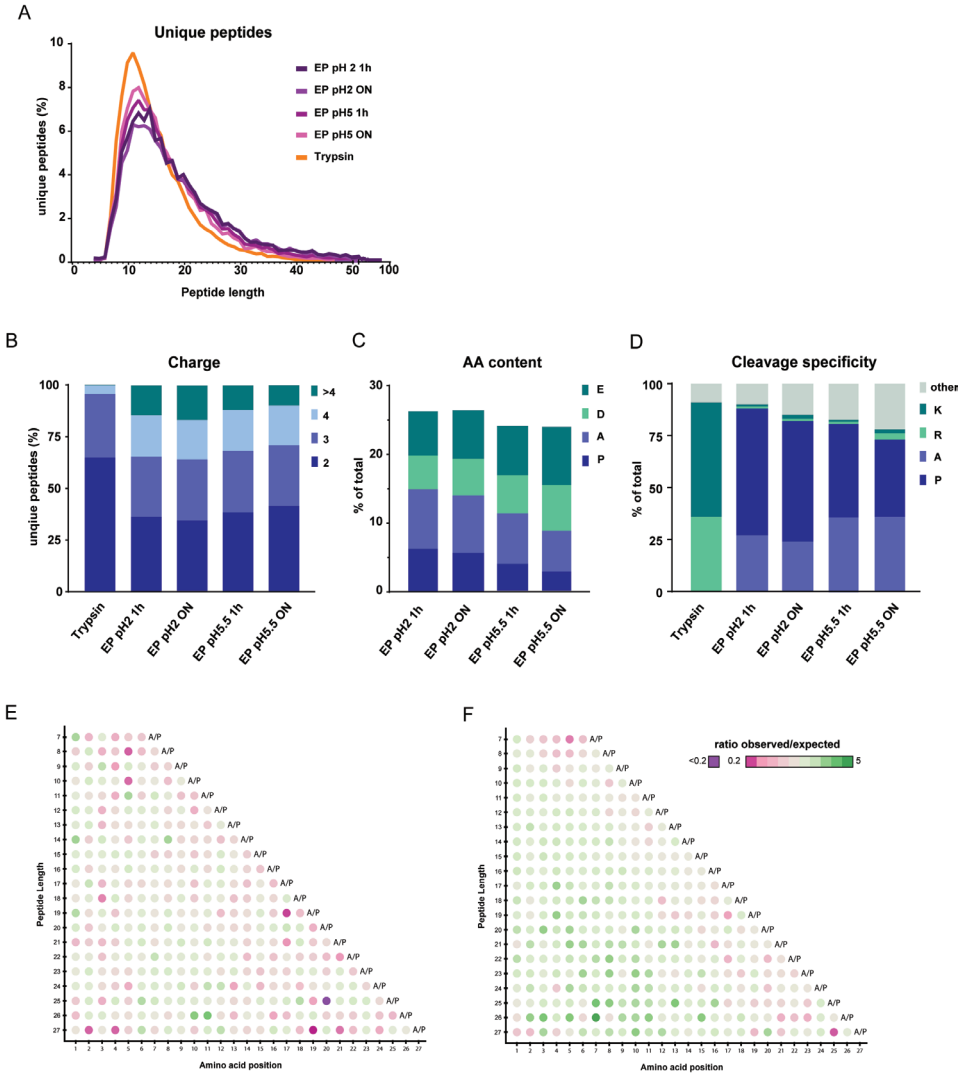


Figure 2 - Comparison of peptide characteristics in EndoPro and tryptic digests. (A) Peptide length distribution of identified unique peptides following digestion with Trypsin or EndoPro. All four EndoPro conditions probed here reveal a similar distribution, exhibiting a long tail towards peptides with more than 50 amino acids, which was not observed for tryptic peptides. (B) Charge distribution of all unique peptides identified following the different digestion conditions, where digestion with EndoPro results in more highly charged peptides (z ≥4). (C) Amino acid content of the peptides identified in the EndoPro digests under various digestion conditions. With increase in pH and digestion duration, negatively charged amino acids are more frequently observed and the A/P content of the peptides is reduced. (D) Cleavage specificity of the identified peptides. Digestion with EndoPro yields highly specific proline and alanine C-terminal peptides, especially at pH = 2, with a Pro/Ala specificity close to that of Trypsin for Arg/Lys. (E, F) Location of Asp on peptides digested ON with EndoPro at (E) pH = 2 and (F) pH = 5.5. At pH = 5.5 the negatively charged amino acid is disfavored at the C-terminus of the generated peptides. This was not observed for peptides produced at pH = 2, indicating that two distinct sets of peptides are formed at these pH values.

of cleavages C-terminal to alanine increased to 36% (1hr) and 37% (ON), with only 49% (1hr) and 39% (ON) proline specific cleavages. Furthermore, the percentage of non-specific cleavages (i.e. not C-terminal of Ala/Pro) observed after EndoPro digestion increased with an increase in pH from pH = 2 to pH = 5.5. Therefore, the specificity of EndoPro can be to some extent controlled via the pH in the digestion step. Notably, EndoPro reaches up to 87% specificity for Pro/Ala at pH = 2 and 1 hr digestion, thereby achieving a very high specificity, on par with Trypsin that reaches 91% specificity for Arg/Lys in our data. The ability of EndoPro to perform proteome-wide digestion with such high specificity could be advantageous in downstream data analysis, as specific searches are far less computationally demanding. Hence, we conclude that EndoPro may be used as a high performance protease for proteomics, as in many aspects its performance is comparable to that of Trypsin.

Performance of EndoPro versus Trypsin

Comparing the search input and output characteristics for all EndoPro and tryptic digests, we found that all digestion conditions generated a similar number of MS2 scans (Table S1), indicating that a similar number of peptides with suitable charge states were produced by EndoPro and Trypsin. However, we observed a lower conversion of MS/MS events to peptide identifications for EndoPro (around 40%) than for Trypsin (67%). Still, the 40% ID rate, which we obtained using Byonic, is well above what has been typically reported for other proteases (i.e. ~15-30%) than Trypsin[9,10,25]. To objectively compare the characteristics and performance of EndoPro and Trypsin the peptide and protein identification datasets should ideally be of similar size. Therefore, we decided to accumulate all non-redundant peptide and protein IDs obtained by EndoPro under the four tested digestion conditions, which resulted in a dataset in numbers comparable with that acquired following tryptic digestion (see Table 2).

Using these equally large datasets (around 5000 proteins and 35000 peptides each, see Table 2), we compared the overlap of unique proteins identified following digestion by either EndoPro or by Trypsin (Figure 3A). Of the 7240 unique proteins identified in total, only 35% were identified by both proteases, whereas 30% and 35% were uniquely identified in tryptic and EndoPro digests, respectively (Figure 3B,C). Typically, in our lab (and in line with many other labs) the overlap between proteome analyses on digests acquired under exactly identical digestion conditions is around 65% (Figure 3B,C), largely due to the undersampling problem which cannot be avoided in shotgun proteomics[33]. Hence, we consider this to be the maximum achievable protein overlap. The overlap in protein ID between the datasets obtained following digestion at pH = 2, comparing 1 hr and ON digestions, was 61%, slightly superior to the overlap between the datasets obtained following digestion at pH = 5.5, for either 1 hr or ON (49%). The overlap between the datasets acquired either at pH = 2 or pH = 5.5 was found to be only 44%. Even more strikingly, the overlap between peptides generated with EndoPro and Trypsin was even much lower, namely only 35% (Figure 3A). We conclude that this low overlap is not simply due to

the stochastic nature of shotgun mass spectrometry, as the increase in protein identifications when adding a replicate of the same protease is significantly smaller than when using another protease and 30% of the proteins reproducibly identified in EndoPro were not identified using Trypsin (Figure 3D).

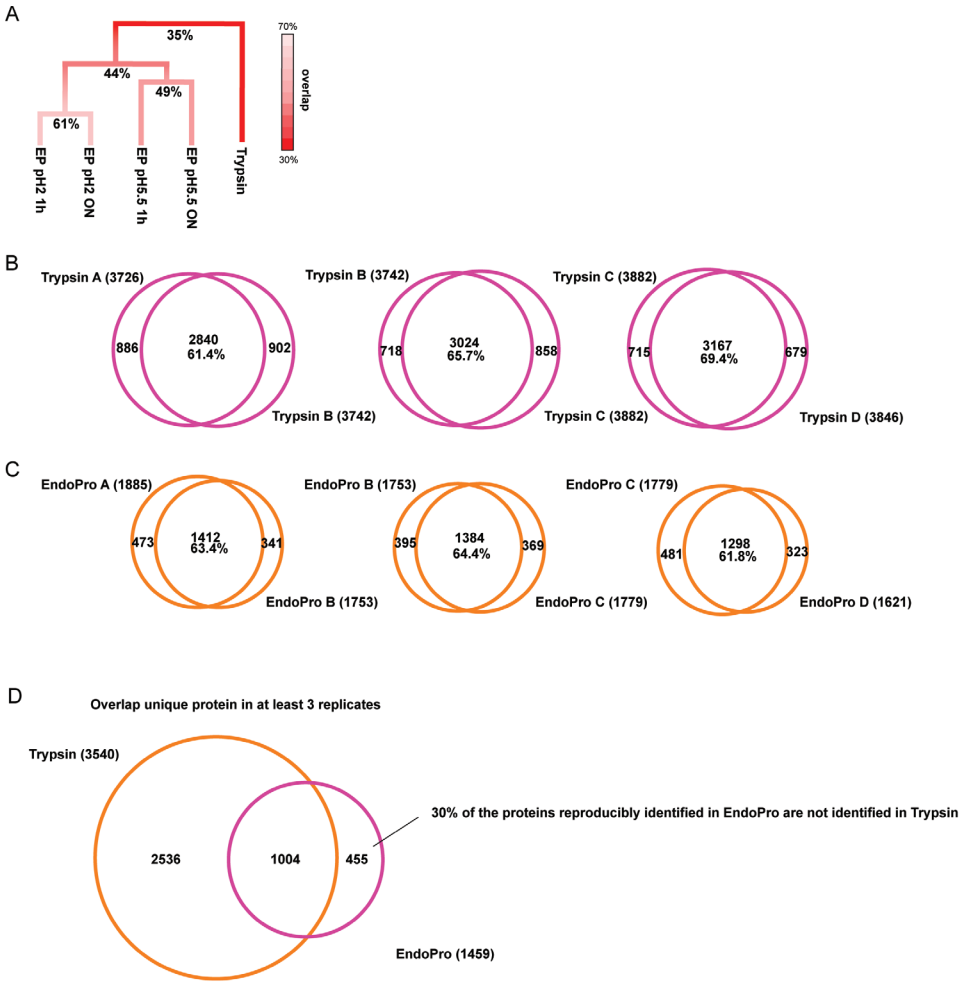


Figure 3 – Highly complementary protein identifications observed by using EndoPro or Trypsin. (A) Overview of the overlap in proteins identified by using the different proteases and varying digestion conditions as listed in Table 2, illustrating how complementarity increases when cleaving with EndoPro at different conditions. The smallest overlap, 35%, is observed between EndoPro and Trypsin. (B, C) Reproducibility of (B) Trypsin and (C) EndoPro technical replicate analyses, revealing a robust overlap of around 65%. (D) When comparing all unique protein groups identified in at least three out of four technical replicates, 30% of the proteins that are reproducibly identified using EndoPro are not identified in tryptic lysates.

Next, we set out to assess what kind of characteristics form the basis for the complementarity in proteome coverage we observed between EndoPro and Trypsin. To this end, we compared proteins for which one protease clearly outperformed the other. As a metric, we focused on proteins whose obtained sequence coverage with EndoPro was at least 50% higher than with Trypsin, or vice versa (Figure 4A). This data proved to be very consistent in all four biological replicates, as demonstrated in Figure S2. Although we identified many proteins with a sufficient sequence coverage in both EndoPro and tryptic digests, our data also revealed large clusters of proteins that remain seemingly undetectable by using Trypsin. These data nicely illustrate the increase in proteome depth that can be achieved when digesting with a protease other than Trypsin.

Since the digestions with EndoPro and Trypsin are performed at distinct pH values, the source of the low overlap could be due to differences in protein solubility and thus accessibility to the protease (i.e. different proteins precipitate at pH = 2, pH = 5.5 and pH = 8.5, removing them from the possible substrate pool), or on the proteases' substrate preferences. We considered various protein characteristics that might cause the complementarity between the two proteases (Figure 4B-E). Following expectations, EndoPro resulted in better sequence coverage for proteins that have a high Arginine and/or Lysine content (see Figure 4B), as these proteins likely give rise to very small and potentially ambiguous peptides when digested with Trypsin. With regard to the proline content, however, this trend is not observed (Figure 4C). No significant difference in proline content was found between Trypsin and EndoPro at pH = 5.5 and at pH = 2; EndoPro even outperformed Trypsin on proteins with a high proline content. This distinction might be caused by frequent occurrence of proline-rich regions. These Pro-Pro bonds are not cleaved by EndoPro, hence the protease likely produced less short, ambiguous peptides. In most cases, we only observed cleavage C-terminal to the last proline in a proline repeat. Following GO term analysis, no clear differences in protein function or localization were found between the proteins identified with EndoPro or Trypsin.

Subsequently, we evaluated whether the observed complementarity stems from the use of different proteases or is influenced significantly by the different digestion conditions, such as pH. Although the solubility of a protein is influenced by many factors, a key feature is its isoelectric point (pI), the pH where the protein carries no net charge. A comparison of the pI values of the identified proteins is shown in Figure 4D. For reference, we also included the distribution of pIs found in the total human proteome[34]. Despite the large pH difference between the five different conditions (i.e. four distinct EndoPro digestions and a Trypsin digestion), the pI distributions all have a median well below the median for the complete human proteome. Although some differences may be observed between the five conditions, it seems they differ more from the complete proteome than from each other. Hence, we conclude that solubility is not likely the cause of the increase in proteome depth that can

be achieved by utilizing EndoPro.

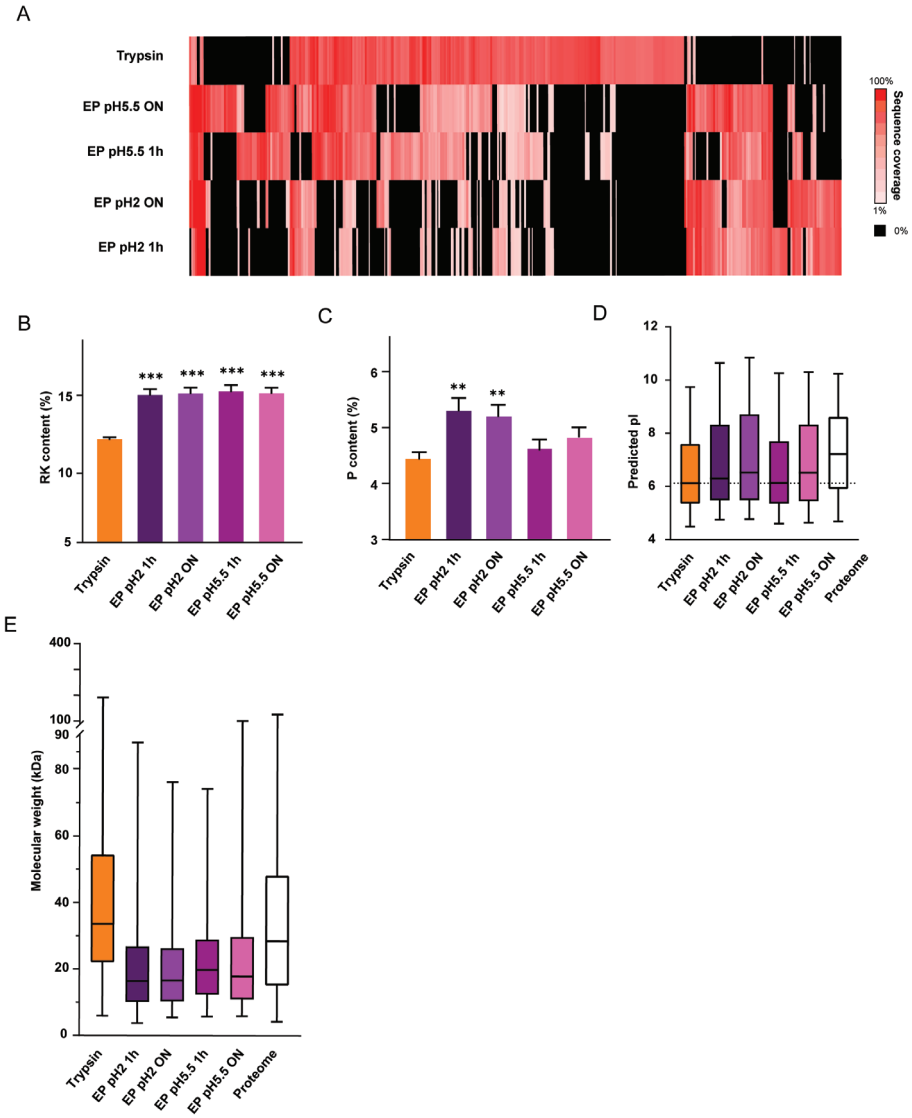


Figure 4 – Proteome Characteristics. (A) Comparison of the sequence coverage achieved by using Trypsin and EndoPro (the latter under 4 different digestion conditions) for in total 380 selected proteins. Only these 380 proteins showing at least 50% more sequence coverage in one of the datasets were considered in B-E. For clarity, proteins for which the two proteases performed comparably were not included. Black indicates no coverage of a protein in a certain condition. (B) Comparison of the arginine and/or lysine content, which is significantly higher in EndoPro peptides. (C-E) Comparison of the proline content (C), isoelectric point (D) and molecular weight (E) of proteins identified using EndoPro (at 4 different conditions) or Trypsin. Notably, as shown in (E) EndoPro favors smaller proteins, Trypsin shows a bias for larger proteins. Significance was determined using one-way ANOVA, with $\alpha = 0.05$. * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$, error bars represent SEM.

Finally, we evaluated whether there was a size bias within the subset of proteins for which one of the proteases outperformed the other (Figure 4E). When compared to the whole human proteome, Trypsin preferred slightly larger proteins, whereas EndoPro favored smaller substrates. Evaluation of protein function or localization yielded no clear preferences for either of the two proteases. Taken together, these data reveal that at the protein level, EndoPro and Trypsin perform comparable and give highly complementary results. The source of complementarity could be solubility based due to the large pH range spanned in these experiments, but this hypothesis is not supported by the distribution of pIs. Therefore, it is likely that enzyme specificity drives the observed complementarity. Interestingly, EndoPro digests also show clear differences based on the cleavage conditions used to generate them, making EndoPro a remarkably flexible proline-specific protease with great potential in bottom-up proteomic studies.

Phosphoproteomics with EndoPro

In addition to changes in its abundance, a proteins' function and/or activity can also be regulated by post-translational modifications (PTMs), such as phosphorylation. These phosphorylation events can be challenging to study due to their low stoichiometry compared to their non-modified counterparts and instability of the modification itself. The field of phosphoproteomics specializes in the analysis of this modification, usually employing enrichment of phosphorylated peptides prior to their analysis by LC-MS/MS. A common problem, however, is that many conventional proteases (e.g. Trypsin) have difficulties cleaving near a phosphorylated amino acid, leading to increased missed cleavages around phosphosites[17,35,36]. Using first several synthetic (phospho)peptides, however, we observed that EndoPro does not exhibit a significant decrease in cleavage rate when cleaving phosphorylated peptides when compared to their non-phosphorylated counterparts (data not shown). We hypothesized that this feature, combined with the high proline content present near phosphorylation sites, could make EndoPro an enzyme very well suitable for phosphoproteomics.

Table 3 – Search input and outcome characteristics for EndoPro and tryptic phospho-enriched digests.

				Byonic semi-specific search						
Protease	**pH**	**Digestion time**	**# MS2 scans**	**#PSMs FDR<0.1**	**#PSMs dmod > 20**	**phospho PSMs**	**% identification**	**% phos**	**total phos sites**	**unique phos sites**
Trypsin	8.5	ON	96641	51502	44933	35319	46%	79%	39905	8898
EndoPro	2	1 hr	87736	25285	20532	14918	23%	73%	15422	3275
EndoPro	2	ON	87415	25254	19895	16423	23%	83%	17489	3794
EndoPro	5.5	1 hr	90021	27805	23213	19406	26%	84%	20471	4326
EndoPro	5.5	ON	93374	26370	22658	17667	24%	78%	19070	4316
EndoPro cumulative				104714	86298	68414	24%	79%	72452	8486

To assess how EndoPro performs in phosphoproteomics, we enriched peptides generated by digestion with EndoPro at pH = 2 and pH = 5 for 1hr or ON using Fe(III)-NTA cartridges in an automated fashion using the AssayMAP Bravo Platform[27]. To benchmark the performance of EndoPro, phosphorylated tryptic peptides were enriched in parallel. For comparison, a general overview of the resulting datasets is shown in Table 3 and an extended overview of the contribution of each fragmentation technique is available in Table S2.

Since the main goal of looking beyond Trypsin as a protease in (phospho) proteomics is to increase our coverage of the phosphorylation sites present in the human proteome, we first set out to assess whether EndoPro is complementary to Trypsin in terms of phosphoprotein and unique phosphosite coverage. Using EndoPro, we identified 2937 unique phosphoproteins, which is comparable to the 3124 unique phosphoproteins identified using Trypsin, see Figure 5A. Interestingly, just 37% of the 4409 unique proteins identified in total were identified by both proteases. If we delve deeper into these shared phosphoproteins, it becomes evident that the two proteases mostly reveal different phosphosites on these shared proteins, see Figure 5B. On the 1652 proteins identified by both EndoPro and Trypsin, 9279 phosphosites were identified of which only 30% were found by both proteases. The remaining 6500 sites were identified by only one of the two enzymes; 3095 sites were uniquely identified by EndoPro and 3405 sites by Trypsin, therefore, the proteases appear extremely orthogonal and employing EndoPro in this setting yields a large increase in attainable information. To evaluate the coverage of phosphosites more thoroughly, we plotted the number of spectral counts we observed for each phosphosite, see Figure 5C (or Figure S3 for more extended heatmaps). This figure revealed that many phosphosites consistently identified with EndoPro (in at least 2 out of 3 biological replicates) were not found at all when digesting with Trypsin and vice versa, highlighting further the complementarity of the enzymes and the importance of extending phosphoproteomics analysis beyond the use of just a single protease[9].

Localization of phosphorylation and motif analysis

Since we expected EndoPro to cleave after prolines and these are extremely frequently occurring in mammalian phosphorylation motifs, we evaluated both the phosphorylated motifs present in our datasets and the location of the phosphorylation sites on the identified phosphopeptides. To assess the different types of kinase motifs present in the dataset, we isolated the environment of each phosphosite identified (seven amino acids up- and downstream of the phosphorylated amino acid) and assessed the relative contribution of known motifs to the EndoPro and tryptic datasets. For clarity, the motifs were classified in only four categories; proline-directed, acidophilic, basophilic or other (Figure 5D). Markedly, the contribution of proline-directed motifs is even larger for EndoPro digestions than we observe for Trypsin, most notably under the short digestion conditions (1hr). This observation is in line with the decrease in relative proline content observed at longer digestion times as depicted in Figure 2C. As expected, we see an increase in motifs containing

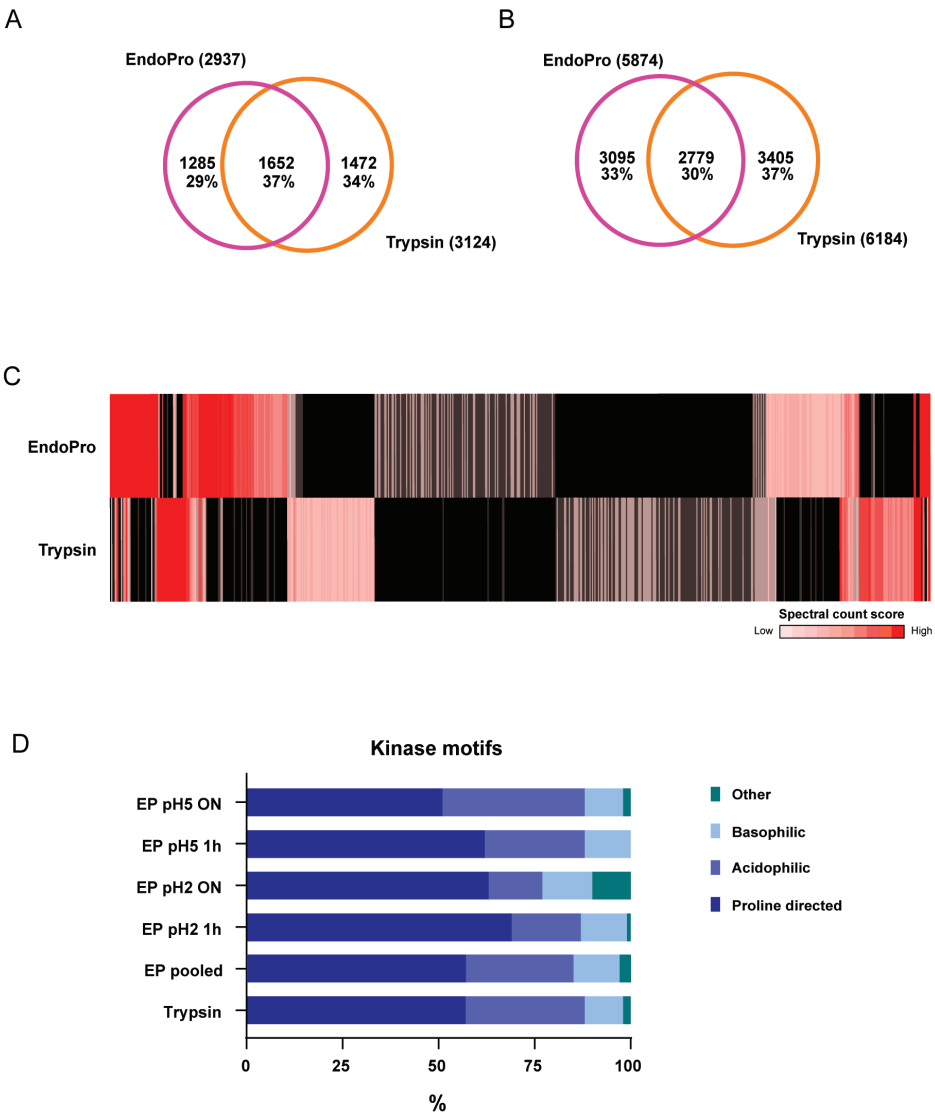


Figure 5 – EndoPro is highly complementary to Trypsin in the identification of site-specific phosphorylation events. (A) Comparison of identified unique phosphoproteins between EndoPro and Trypsin, revealing a 37% overlap. (B) Overlap in identified unique phosphosites on 1652 phosphoproteins identified by both proteases, indicating that on these shared phosphoproteins, only 30% of the phosphosites could be identified by both proteases. (C) Heatmap displaying phosphosite spectral count scores of 13,762 phosphosites from low (1) to high (>10), revealing that EndoPro is highly complementary to Trypsin in identification of phosphosites. Black indicated not identified. (D) Global kinase classification analysis of all identified phosphopeptides, dividing them into 4 categories: Proline-directed, acidophilic, basophilic or other. Although in all analyses the SP/TP motif encompasses over 50% of the detected sites, short digestion with EndoPro results in a further increase of this proline-directed motif to about 70%.

arginine and lysine after EndoPro digestion. Overall, our findings are in agreement with previous work from this lab, in which a thorough examination of multiple proteases for phosphoproteomics revealed that each protease exhibits a bias towards different classes of phosphorylation sites[17].

As EndoPro precisely cleaves after prolines, which are found in the most frequently occurring SerPro/ThrPro phosphorylation sites, and since it is well known that a phosphorylation close to an Arg/Lys hampers the cleavage activity of Trypsin, we queried whether the phosphorylation on these motifs would prevent cleavage of the following proline residue. To assess this, we evaluated the location of the phosphorylation on unique phosphopeptides. We computed the frequency of phosphorylations for each position of the phosphopeptide, with the exception of the last amino acid, as we expect this to be Ala/Pro and Arg/Lys for EndoPro and Trypsin, respectively. The frequency of the phosphorylation site was compared to the frequency expected if phosphorylations would have been randomly distributed across the amino acids of the phosphopeptide (Figure 6). The under- or overrepresentation of a phosphorylation location on the peptides are shown by a color gradient and extreme underrepresentation (at least 5 fold lower than expected) was indicated in purple. These "dot-plots" display several very interesting features.

For the EndoPro phosphopeptides, the highly preferred phosphorylation on the penultimate C-terminal amino acid is very evident, as is the disproportion for phosphorylation on the penultimate amino acid at the N-terminus (Figure 6A). Interestingly, EndoPro also seems to disfavor positively charged amino acids on this position (Figure 1B), which implies no charge is tolerated at this position in the substrate binding pocket. Repulsion of both charges would suggest steric hindrance to be the source of this lack of activity. For Trypsin-like proteases, the N+2 position is reported to be situated in a hydrophobic pocket prior to cleavage[37]. Based on our data, this might also be the case for EndoPro. For the tryptic phosphopeptides the dot-plot reveals a strong disfavor for phosphorylation on the ultimate and penultimate N-terminal amino acids and the penultimate C-terminal amino acid (Figure 6B), confirming that phosphorylation near the Arg/Lys hinders cleavage by Trypsin. Trypsin displays a preferential phosphorylation on the third N-terminal amino acid (likely representing the well-known RXXS/T basophilic kinase motif). These findings largely explain the increase in missed cleavages on phosphopeptides observed[36].

In contrast to what is observed with Trypsin, the activity of EndoPro seemed unaffected by a phosphorylation directly preceding the cleavage site, resulting in an overrepresentation of phosphorylation events on the second to last amino acid of the phosphopeptides, see Figure 5A. In total, of all detected singly phosphorylated EndoPro phosphopeptides, 19% had their phosphorylation on the C-terminal SP/TP. See Figure S4 for the phosphosite localization of all specific EndoPro digestion conditions employed.

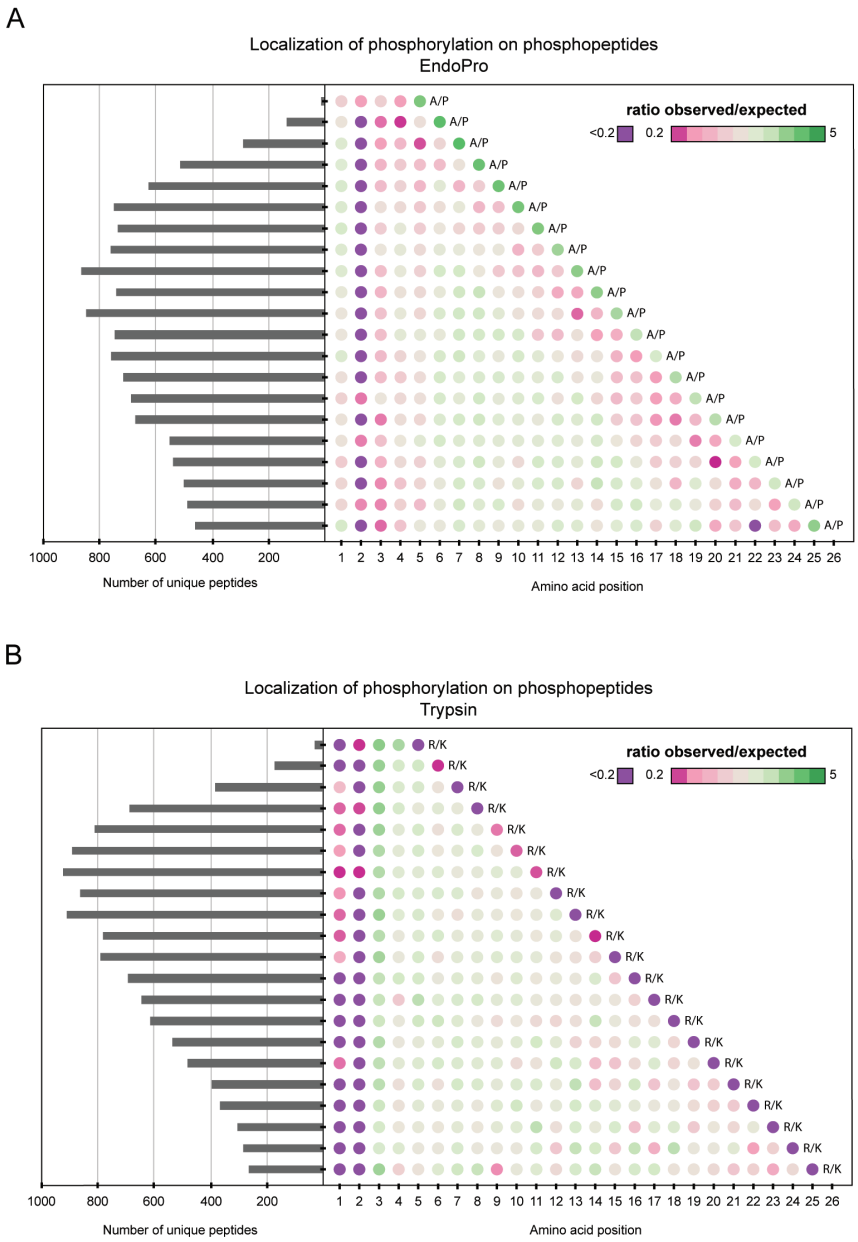


Figure 6 – Amino acid length and localization of phosphorylation sites on the identified phosphopeptides. (A) Localization of the phosphorylation on unique phosphopeptides from EndoPro, showing the highly preferred phosphorylation on the second to last amino acid on the peptide (i.e., Ser-Pro or Thr-Pro), and the disfavor for phosphorylation on the penultimate N-terminal amino acid on the EndoPro peptides. (B) Localization of phosphorylation on unique phosphopeptides following trypsin digestion at pH = 8.5, revealing a strong disfavor for phosphorylation on the ultimate and penultimate N-terminal amino acids on the peptides, and preferential phosphorylation on the third amino acid of the identified phosphopeptides.

Discussion

Although still not frequently used, the use of proline-directed proteases in a mass spectrometry-based proteomics setting has been explored previously[23–25]. Schräder *et al.* recognized the potential of proline-directed proteases in proteomics characterizing Neprosin, a protease originally from Nepenthes ventrata[25]. In their work, the ON digestion of a HeLa cell lysate at pH = 2.5 yielded 61% proline-specific cleavages for Neprosin, which is comparable to our findings for EndoPro (62% after ON digestion at pH = 2). Additionally, they nicely illustrated the potential of proline-specific proteases for the mapping of PTMs on a histone sample. Due to the high activity of EndoPro at low pH, the protease has also found applications in the food industry, where EndoPro was assessed for its ability to degrade gluten and the debittering of protein hydrolysates, as well as in structural studies based on hydrogen-deuterium exchange mass spectrometry, where a low pH is essential to reduce the rate of deuterium back-exchange[24,38–40]. Thus, proline-directed proteases are versatile proteases that can be used orthogonally to the more conventional proteases in various mass spectrometry-based studies.

Here, we have evaluated EndoPro for its use in bottom up (phospho) proteomics, with the aim to boost its performance by optimizing different digestion conditions, peptide fragmentation methods and scoring algorithms. We showed that the protease has a capacity to generate peptides from proteins comparable to Trypsin, evidenced by similar numbers of MS/MS events. When the proper digestion conditions are chosen, the cleavage specificity for alanine and proline is very high. Interestingly, EndoPro cleavage patterns appear influenced by the pH during the digestion, with a lower overall specificity observed using EndoPro at pH = 5.5 than at pH = 2. The mechanism underlying this pH-dependency was not studied thoroughly here, however, we did find that the overall Pro content of the proteins identified at pH = 5.5 was significantly lower than at pH = 2. Hence, it might be possible that fewer proline residues were available for cleavage, possibly due to the occurrence of a different pool of soluble proteins at pH = 5.5 when compared to pH = 2. This hypothesis is supported by the limited overlap (44%) in identified proteins between the two EndoPro conditions. We did not observe an effect on the length of the peptides identified, which also implies proline residues were not missed during digestion but likely not present as frequently.

Through this work we show that the performance of EndoPro as protease for proteomics applications is already very good, but its full potential is still not reached. As observed also with other proteases, EndoPro also suffers from the tryptic bias that has been created in the conventional proteomics pipelines, both in the peptide separation, fragmentation and scoring segments of the proteomics experiment. Despite that, our data shows that EndoPro is a protease very suitable for producing peptides for proteomics analysis. One should keep in mind, however, that the obtainable proteome coverage is rather distinct when the digestion is performed at pH = 2 or at pH = 5.5. In addition,

the proteome coverage generated with EndoPro is highly complementary to the coverage that can be reached using Trypsin. Finally, EndoPro provides one of the most complementary proteases for phosphoproteomics, delivering a large subset of phosphosites not easily covered by Trypsin. Furthermore, in contrast to Trypsin, cleavage by EndoPro is not hampered by the presence of a neighboring phosphorylation.

Identifying non-tryptic MS/MS spectra

One of the main concerns when using less conventional proteases in proteomics-type experiments is that the resulting datasets almost always give a lower peptide identification rate than the tryptic datasets. For ArgC, AspN, Chymotrypsin, GluC, LysC and LysN, average identification rates of 22%, 11%, 17%, 13%, 23% and 11% have been reported, compared to a 37% identification rate for Trypsin[10]. Similarly, Schräder *et al*. previously reported identification rates of 20%, 46% and 52% for Neprosin, LysC and Trypsin, respectively[25]. These findings are in agreement with our finding, where EndoPro identification rates are also about half of the Trypsin identification rate. The lower rates associated with non-tryptic digestions are probably not caused by a lack of good peptides, as the number of MS2 scans for these runs is similar. Hence, the number of peptides with suitable mass-to-charge ratio is expected to be similar for all digests.

This leaves several other sources likely responsible for the reduced identification rates. Firstly, the peptides produced by each of these proteases may have characteristics that make them less suitable for current fragmentation-based sequencing methods by mass spectrometry. They may for instance carry less positive charges, reducing the likelihood of observing charged fragment ions that can be used for database matching. For our dataset, however, this is not the case as the peptides generated by EndoPro have even more positive charges than the tryptic peptides. Secondly, the peptides' chemical composition may lead to fragmentations patterns or cleavages at positions that are unexpected for Trypsin. For instance, EndoPro peptides do not carry an arginine or lysine residue at their C-terminus, which likely leads to a less extended sequence informative y-ion series. Thirdly, database search and peptide scoring algorithms have mainly been optimized for tryptic peptides. Any fragmentation behavior not observed in tryptic peptides, therefore, is likely penalized by the conventional scoring algorithms, resulting in lower scores. Notably, using standard conditions with other search engines such as Mascot, Andromeda and Sequest gave us even lower identification rates than those reported here (by a factor 2, data not shown), evidently depending also on the fragmentation method employed.

Due to EndoPro's high preference to cleave C-terminal to proline, many of the peptides generated with this protease are expected to have a Pro residue at their C-terminus, making them very dissimilar to the typical tryptic peptides that carry a Lys or Arg at their C-terminus. Indeed, we observed a clear C-terminal

proline effect in their fragmentation spectra. During HCD fragmentation, more than 95% of the EndoPro MS2 spectra contained a very prominent y1 ion at 116.07 m/z, corresponding to the preferential gas-phase cleavage of the bond preceding the proline. Assuming that the presence of a 116.07 m/z ion is diagnostic for a peptide ending in C-terminal proline, we noticed in our LC-MS runs many more MS2 spectra likely originating from EndoPro peptides which sequence could not be assigned. This could possibly be improved by optimizing the MS parameters such as the collision energy, to maintain the diagnostic y1 ion while also allowing sufficient fragmentation in other parts of the peptide. In EThcD spectra, we observed significantly less proline y1 ion formation, only about 50% of the recorded MS2 spectra, which allows for the detection and assignment of other fragment ions and hence a better scoring of the PSMs. This is also reflected in the higher ID rate observed with the EThcD/HCD DT method.

Taken together, many factors contribute to a lower score for the EndoPro, illustrating a deeply rooted tryptic bias in proteomic workflows, resulting in lower PSMs for non-tryptic peptides. This argues especially for a better optimization of MS methods and search algorithms towards non-tryptic peptides.

Fragmenting with a C-terminal phosphorylation and phosphosite localization

Given the large proportion of phosphopeptides that carry their phosphorylation on the penultimate amino acid of the EndoPro peptide, we hypothesize that these phosphopeptides may have a negatively charged C-terminus. Again, this is in sharp contrast to tryptic phosphopeptides, which have a positively charged C-terminus and for the most part carry their phosphorylation somewhere in the middle of the peptide. Phosphorylation of the amino acid before the C-terminal proline seemed to reduce the proline-effect observed for the EndoPro peptides, resulting in a better fragmentation ion coverage than observed for non-phosphorylated peptides. In addition, since the EndoPro phosphopeptides predominantly carry their phosphorylation at the C-terminus, this affects the probability of having multiple potential phosphorylation sites directly adjacent to each other. When a phosphorylation site is directly preceding the Ala/Pro on the C-terminus, there cannot be a second potential phosphorylation site on that end of the peptide; hence the odds of having many potential sites side by side on a phosphopeptide are lower than when phosphorylations are located more towards the middle of a peptide. This could potentially boost phosphosite localization certainty, especially in peptides that harbor multiple putative phosphate acceptors, such as Ser, Thr, and Tyr. Since one of the major remaining issues in phosphoproteomics is the confident assignment of the exact site of phosphorylation, much computational effort has been invested in improving fragmentation methods and algorithms to boost confident site assignments. Knowledge about the natural occurrence of phosphorylation sites for each used protease, as depicted graphically in Figure 6, can be used to further

improve scoring algorithms and boost the confidence in site-localization.

Conclusion

Here, we evaluated EndoPro and show it is a versatile protease with a very high proline- and alanine-directed specificity. Its activity can be influenced by adjusting the pH of the digestion buffer, whereby it largely retains its specificity but seemingly samples a different part of the proteome. By benchmarking its performance against Trypsin, we observed that over 30% of all unique HeLa proteins were solely identified by EndoPro, as well as 5705 phosphosites that were not observed in the tryptic digests, illustrating EndoPro's high complementarity to Trypsin. This complementarity allows EndoPro to expand our coverage of the various proteomes and sheds light on previously dark, invisible stretches of (phospho)proteins. Since EndoPro clearly outperforms Trypsin on arginine and lysine rich proteins, we see potential for EndoPro in studying proteins involved in nucleotide and chromatin binding, which are often enriched in these positively charged amino acids[41]. In addition, the longer peptides generated by EndoPro and its ability to cleave close to modifications makes the enzyme an interesting candidate for middle-down approaches allowing for more combinatorial PTM information[42,43]. Compared to other alternative proteases, such as LysC, Chymotrypsin, ArgC, EndoPro performs better and is in our view one of the most complementary alternatives to Trypsin, due to its completely different activity profile and specificity. It is rather unique in effectively targeting proline residues in (phospho)proteomics, that are often causing complications for the other proteases.

Acknowledgements

We acknowledge support from the Netherlands Organization for Scientific Research (NWO) funding the large-scale proteomics facility Proteins@Work (project 184.032.201) and X-omics (project 184.034.019) embedded in the Netherlands Proteomics Centre. A.J.R.H. acknowledge further support by the NWO TOP-Punt Grant 718.015.003 and the and EU Horizon 2020 program INFRAIA project Epic-XS (Project 823839).

Conflicts of Interest

The authors declare the following competing financial interest(s): M.A. and M.O. are DSM employees. DSM sells An-PEP for food applications. MB is founder and employee of ProteinMetrics. ProteinMetrics develops and commercializes the Byonic software.

References

1 Cohen P (2001) The role of protein phosphorylation in human health and disease. Eur J Biochem 268, 5001–5010.
2 Hanash S (2003) Disease proteomics. Nature 422, 226–232.
3 Zhang Y, Fonslow BR, Shan B, Baek M-C & Yates JR (2013) Protein analysis by shotgun/bottom-up proteomics. Chem Rev 113, 2343–2394.
4 Mann M, Kulak NA, Nagaraj N & Cox J (2013) The coming age of complete, accurate, and ubiquitous proteomes. Mol Cell 49, 583–590.
5 Steen H & Mann M (2004) The abc's (and xyz's) of peptide sequencing. Nat Rev Mol Cell Biol 5, 699–711.
6 Vandermarliere E, Mueller M & Martens L (2013) Getting intimate with trypsin, the leading protease in proteomics. Mass Spectrom Rev 32, 453–465.
7 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.
8 Eichacker LA, Granvogl B, Mirus O, Muller BC, Miess C & Schleiff E (2004) Hiding behind hydrophobicity. J Biol Chem 279, 50915–50922.
9 Tsiatsiani L & Heck AJR (2015) Proteomics beyond trypsin. FEBS J 282, 2612–2626.
10 Giansanti P, Tsiatsiani L, Low TY & Heck AJR (2016) Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. Nat Protoc 11, 993–1006.
11 Biringer RG (2006) Enhanced sequence coverage of proteins in human cerebrospinal fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS. Brief Funct Genomic Proteomic 5, 144–153.
12 Gauci S, Helbig AO, Slijper M, Krijgsveld J, Heck AJR & Mohammed S (2009) Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. Anal Chem 81, 4493–4501.
13 Huesgen PF, Lange PF, Rogers LD, Solis N, Eckhard U, Kleifeld O, Goulas T, Gomis-Ruth FX & Overall CM (2015) LysargiNase mirrors trypsin for protein Cterminal and methylation-site identification. Nat Methods 12, 55–58.
14 Wu Z, Huang J, Huang J, Li Q & Zhang X (2018) LysC/Arg-C, a more specific and efficient digestion approach for proteomics studies. Anal Chem 90, 9700–9707.
15 Raijmakers R, Neerincx P, Mohammed S & Heck AJR (2010) Cleavage specificities of the brother and sister proteases Lys-C and Lys-N. Chem Commun 46, 8827.
16 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 MSbased proteomics. J Proteomics 75, 1454–1462.
17 Giansanti P, Aye TT, van den Toorn H, Peng M, van Breukelen B & Heck AJR (2015) An augmented multiple-protease-based human phosphopeptide Atlas. Cell Rep 11, 1834–1843.
18 Schimmel PR & Flory PJ (1968) Conformational energies and configurational statistics of copolypeptides containing L-proline. J Mol Biol 34, 105–120.
19 MacArthur MW & Thornton JM (1991) Influence of proline residues on protein conformation. J Mol Biol 218,

397–412.
20 Vaisar T & Urban J (1996) Probing the proline effect in CID of protonated peptides. J Mass Spectrom 31, 1185–1187.
21 Raulfs MDM, Breci L, Bernier M, Hamdy OM, Janiga A, Wysocki V & Poutsma JC (2014) Investigations of the mechanism of the "proline effect" in tandem mass spectrometry experiments: the "pipecolic acid effect". J Am Soc Mass Spectrom 25, 1705–1715.
22 Huo D, Qin T & Zu L (2019) Energetic switch of the proline effect in collision-induced dissociation of singly and doubly protonated peptide Ala-Ala-Arg-Pro-AlaAla. J Mass Spectrom 54, 55–65.
23 Sebela M, Rehulka P, K abrt J, Rehulkov a H, Ozdian T, Raus M, Franc V & Chmelık J (2009) Identification of N-glycosylation in prolyl endoprotease from
Aspergillus niger and evaluation of the enzyme for its possible application in proteomics. J Mass Spectrom 44, 1587–1595.
24 Tsiatsiani L, Akeroyd M, Olsthoorn M & Heck AJR (2017) Aspergillus niger prolyl endoprotease for hydrogen-deuterium exchange mass spectrometry and protein structural studies. Anal Chem 89, 7966–7973.
25 Schreader CU, Lee L, Rey M, Sarpe V, Man P, Sharma S, Zabrouskov V, Larsen B & Schriemer DC (2017) Neprosin, a selective prolyl endoprotease for bottom-up proteomics and histone mapping. Mol Cell Proteomics 16, 1162–1171.
26 Chen X, Wu D, Zhao Y, Wong BHC & Guo L (2011) Increasing phosphoproteome coverage and identification of phosphorylation motifs through combination of different HPLC fractionation methods. J Chromatogr B 879, 25–34.
27 Post H, Penning R, Fitzpatrick MA, Garrigues LB, Wu W, MacGillavry HD, Hoogenraad CC, Heck AJR & Altelaar AFM (2017) Robust, sensitive, and automated phosphopeptide enrichment optimized for low sample amounts applied to primary hippocampal neurons. J Proteome Res 16, 728–737.
28 Frese CK, Zhou H, Taus T, Altelaar AFM, Mechtler K, Heck AJR & Mohammed S (2013) Unambiguous phosphosite localization using electron-transfer/higherenergy collision dissociation (EThcD). J Proteome Res 12, 1520–1525.
29 Oliveros JC (2007) VENNY. An interactive tool for comparing lists with Venn Diagrams. https://bioinfogp. cnb.csic.es/tools/venny/index.html.
30 Hulsen T, de Vlieg J & Alkema W (2008) BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genom 9, 488.
31 Colaert N, Helsens K, Martens L, Vandekerckhove J & Gevaert K (2009) Improved visualization of protein consensus sequences by iceLogo. Nat Methods 6, 786– 787.
32 Trevino SR, Scholtz JM & Pace CN (2007) Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. J Mol Biol 366, 449–460.
33 Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham A-JL, Bunk DM, Kilpatrick LE, Billheimer DD, Blackman RK, Cardasis HL et al. (2010) Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. J Proteome Res 9, 761–

776.
34 Kozlowski LP (2017) Proteome- pI: proteome isoelectric point database. Nucleic Acids Res 45, D1112–D1116.
35 Schlosser A, Pipkorn R, Bossemeyer D & Lehmann WD (2001) Analysis of protein phosphorylation by a combination of elastase digestion and neutral loss tandem mass spectrometry. Anal Chem 73, 170–176.
36 Molina H, Horn DM, Tang N, Mathivanan S & Pandey A (2007) Global proteomic profiling of phosphopeptides using electron transfer dissociation
tandem mass spectrometry. Proc Natl Acad Sci USA 104, 2199–2204.
37 Huber R & Bode W (1978) Structural basis of the activation and action of trypsin. Acc Chem Res 11, 114–122.
38 Mitea C, Havenaar R, Drijfhout JW, Edens L, Dekking L & Koning F (2008) Efficient degradation of gluten by a prolyl endoprotease in a gastrointestinal model: implications for coeliac disease. Gut 57, 25–32.
39 Konig J, Holster S, Bruins MJ & Brummer RJ (2017) Randomized clinical trial: effective gluten degradation by Aspergillus niger-derived enzyme in a complex meal setting. Sci Rep 7, 13100.
40 Edens L, Dekker P, van der Hoeven R, Deen F, de Roos A & Floris R (2005) Extracellular prolyl endoprotease from Aspergillus niger and its use in the debittering of protein hydrolysates. J Agric Food Chem 53, 7950–7957.
41 Chandana T & Venkatesh YP (2016) Occurrence, functions and biological significance of arginine-rich proteins. Curr Protein Pept Sci 17, 507–516.
42 Pandeswari PB & Sabareesh V (2019) Middle-down approach: a choice to sequence and characterize proteins/ proteomes by mass spectrometry. RSC Adv 9, 313–344.
43 Cristobal A, Marino F, Post H, van den Toorn HWP, Mohammed S & Heck AJR (2017) Toward an optimized workflow for middle-down proteomics. Anal Chem 89, 3318–3325.

Supplemental Information

Figure S1. Amino acid content of PSMs in different conditions.

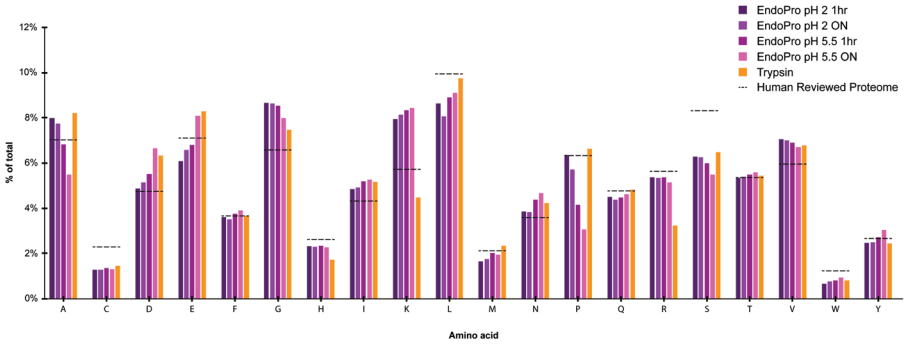


Figure S1 – Amino acid content of PSMs following different digestion conditions. Complete overview of amino acid content of the peptides generated by EndoPro under the four evaluated experimental conditions and in orange that for Trypsin. The dashed line indicated the natural occurrence of that amino acid in the human proteome. The data shows that the occurence of a particular amino acid in peptides produced can strongly depend on the digestion conditions and protease employed.

Figure S2. Extended heatmap of proteome dataset.

Figure S2 – Extended heatmap op proteome dataset. All identified protein groups with a sequence coverage difference of 50% or higher of Trypsin (orange) and EndoPro pH = 5.5 1h (purple) were subjected to hierarchical clustering (589 protein groups). For both experimental conditions, the four technical replicates cluster nicely together.

Figure S3. Extended heatmap of phosphoproteomics data and assessment of reproducibility.

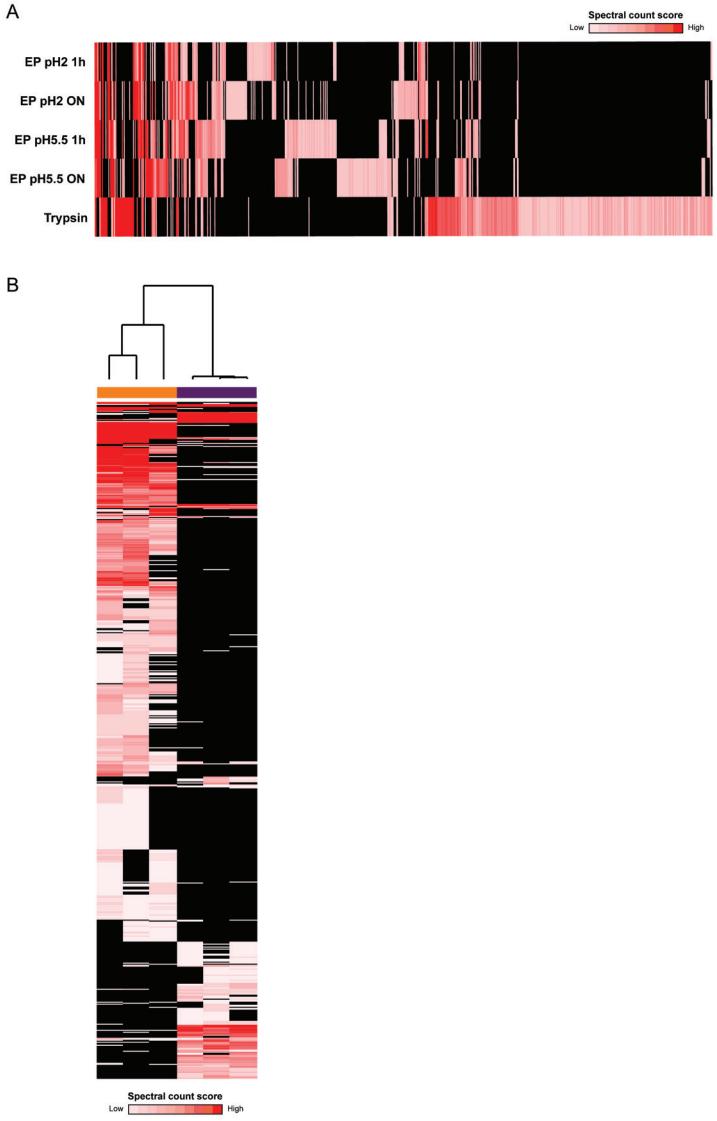


Figure S3 – Extended heatmap of the phosphoproteomics data and assessment of reproducibility. (A) Heatmap showing the coverage of the 13,762 unique phosphosites across the four EndoPro conditions and Trypsin. (B) Phosphosites identified in minimal 2 out of 3 biological replicates by at least one protease (1628 sites) in Trypsin (orange) and EndoPro pH = 5.5 1h (purple) were subjected to hierarchical clustering. For both experimental conditions, the three technical replicates cluster nicely together.

Figure S4. Phosphosite localization as extracted from the data for all different EndoPro digestion conditions.

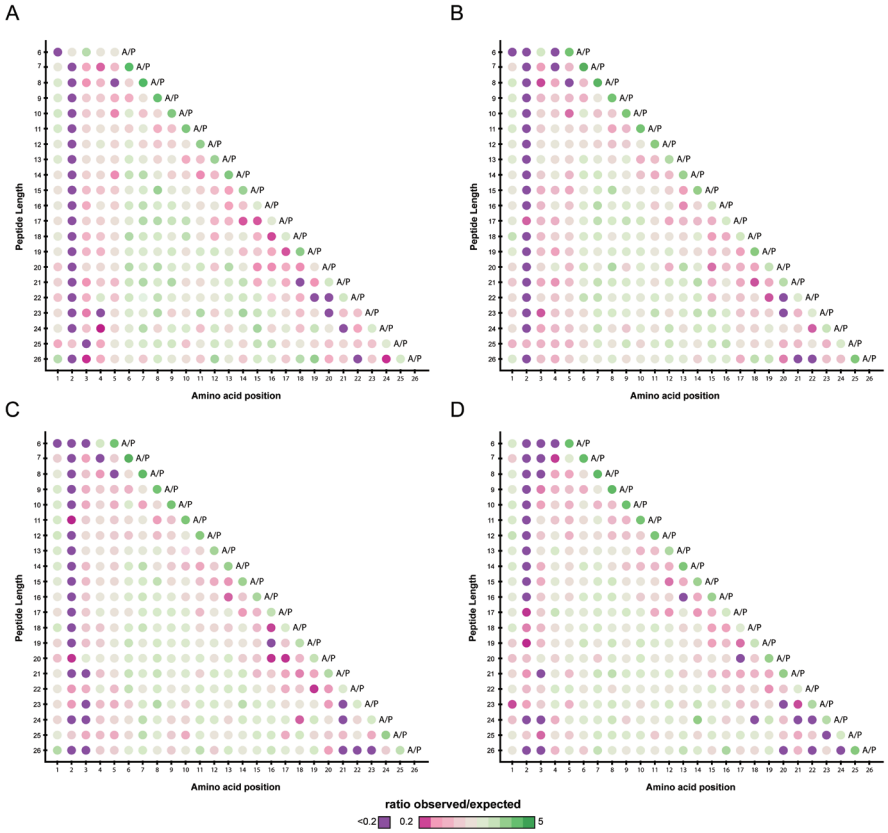


Figure S4 – Phosphosite localization as extracted from the data for all different EndoPro digestion conditions: (A) pH = 2 1hr, (B) pH = 2 ON, (C) pH = 5.5 1hr and (D) pH = 5.5 ON.

Table S1 - Search input and outcome characteristics for EndoPro and tryptic digests.

Protease	pH	Digestion time	Fragmentation	# MS2 scans	# PSMs FDR<0.1	# Unique Peptides	# Unique Proteins	ID %
EndoPro	2	1 hr	ETD	38422	18336	9707	1974	48%
EndoPro	2	1 hr	EThcD/HCD (DT)	53723	19628	9221	1725	37%
EndoPro	2	1 hr	HCD	59919	16287	8250	1565	27%
EndoPro	2	ON	ETD	37925	17120	9491	1738	45%
EndoPro	2	ON	EThcD/HCD (DT)	53328	22350	10779	1620	42%
EndoPro	2	ON	HCD	59862	18252	9446	1517	30%
EndoPro	5.5	1 hr	ETD	38707	21721	12318	2053	56%
EndoPro	5.5	1 hr	EThcD/HCD (DT)	54822	25385	13034	1899	46%
EndoPro	5.5	1 hr	HCD	61700	21796	12078	1824	35%
EndoPro	5.5	ON	ETD	37405	18723	11238	1741	50%
EndoPro	5.5	ON	EThcD/HCD (DT)	55484	22184	12973	1875	40%
EndoPro	5.5	ON	HCD	62676	19411	11983	1743	31%
Trypsin	8.5	ON	ETD	34688	22998	16373	2377	66%
Trypsin	8.5	ON	EThcD/HCD (DT)	63689	43204	28352	3798	68%
Trypsin	8.5	ON	HCD	65446	43480	28726	3969	66%

Table S2 - Search input and outcome characteristics for EndoPro and tryptic phosphopeptides.

Protease	pH	Digestion time	Fragmentation	# MS2 scans	# PSMs FDR<0.1	# PSMs dmod > 20	ID %	# Phospho PSMs	% Phos PSMs
EndoPro	2	1 hr	ETD	22132	8839	7239	33%	5006	69%
EndoPro	2	1 hr	EThcD/HCD (DT)	29600	10081	8370	28%	5790	69%
EndoPro	2	1 hr	HCD	36004	6365	4923	14%	4122	84%
EndoPro	2	ON	ETD	21475	8297	6515	39%	5272	81%
EndoPro	2	ON	EThcD/HCD (DT)	28696	10641	8626	37%	6931	80%
EndoPro	2	ON	HCD	37244	6316	4754	17%	4220	89%
EndoPro	5.5	1 hr	ETD	21608	10110	8419	39%	6918	82%
EndoPro	5.5	1 hr	EThcD/HCD (DT)	29732	10767	9154	31%	7522	82%
EndoPro	5.5	1 hr	HCD	38681	6928	5640	15%	4966	88%
EndoPro	5.5	ON	ETD	22658	9237	7992	41%	6094	76%
EndoPro	5.5	ON	EThcD/HCD (DT)	31262	10527	9197	34%	6880	75%
EndoPro	5.5	ON	HCD	39454	6606	5469	17%	4693	86%
Trypsin	8.5	ON	ETD	22287	11530	10279	52%	7552	73%
Trypsin	8.5	ON	EThcD/HCD (DT)	33421	19453	17158	58%	13198	77%
Trypsin	8.5	ON	HCD	40933	20519	17496	50%	14569	83%

Chapter 4

Shaping the human proteome: occurrence and influence of proline *cis*/*trans* isomerization

Saar A.M. van der Laarse and Albert J.R. Heck

Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

a preliminary study

Abstract

The actions of a protein are determined by a range of factors, including their three-dimensional shape. This shape is largely determined by a proteins' amino acid sequence, the order in which its amino acids are connected. Chemically, each of these peptidyl bonds can exist in two isomers, *cis* and *trans*, both with a distinct effect on the overall protein shape. Here, we investigate protein structures listed in the ProteinDataBank (PDB) to evaluate the abundance of *cis* peptidyl bonds in the human proteome. Across all amino acid combinations, we found that 0.35% was present in a *cis* conformation. Most of these *cis* bonds occur preceding proline residues, which are hard to study due to a lack of isomer specific proteases that cleave at proline. Therefore, we turned our attention to evaluating the isomer-specificity of EndoPro, a protease that specifically cleaves C-terminal to prolines. We report that EndoPro does exhibit isomer-specificity and preferentially cleaves *trans*-proline residues. Hence, the protease might be a valuable tool in studying the role of isomer specificity in biochemical reactions and interactions.

Introduction

Proteins are a key component of the cell, involved in all essential biochemical processes required in any given organism. Humans express over 20,000 different proteins, each one comprised of a linear chain of amino acids which is folded to produce the secondary, tertiary and quaternary structures. The actions of a protein are determined by a range of factors, such as their three-dimensional shape, the combination of amino acids used and potential post-translational modifications (PTMs) present on the proteins surface which can influence a proteins activity. In humans, primarily 20 different amino acids are used, each with its own biochemical properties. These amino acids are linked together in a condensation reaction, producing the peptidyl bond.

Chemically, these peptidyl bonds can exist in two isomers, *cis* and *trans*, each with a different effect on the overall protein shape (Figure 1). However, due to differences in energy between the two isomers, the lowest energy isomer, i.e. the *trans*-isomer, is biologically preferred and results in peptide chains that are largely linear. An interesting exception to this rule is proline, a cyclic amino acid that carries a secondary amine in a ring-type side chain, resulting in a smaller energy difference between the two conformations and hence a relatively more stable *cis*-isomer. Due to the rotation across the amide bond, *cis*-prolines drastically affect the higher-order structure of a protein, influencing its function and/or activity[1,2]. Across the human proteome, an estimated 5.2% of all X_{AA}-Pro bonds populates the *cis*-isomer[3].

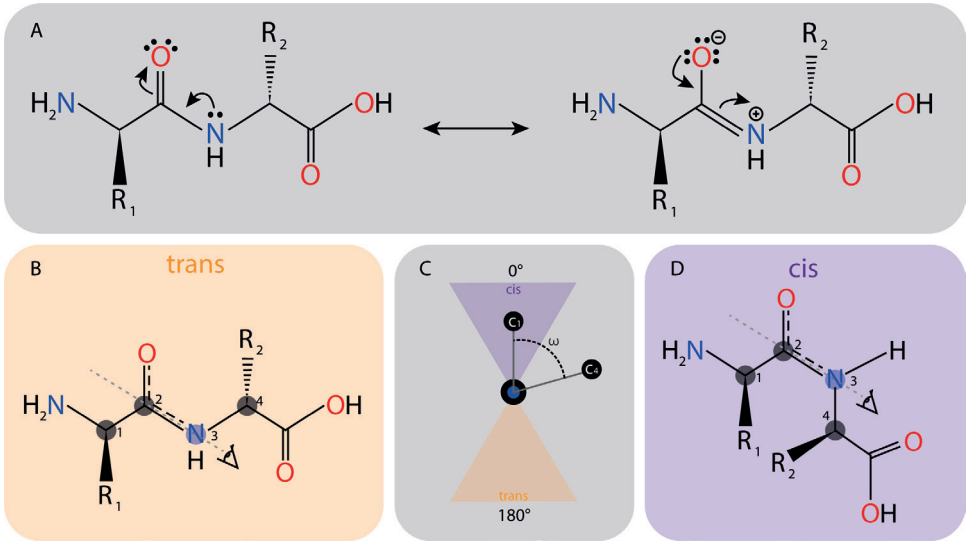


Figure 1 - *Cis*/*trans* isomerization across a peptidyl bond. A) Delocalization of the nitrogen lone pair across the carbonyl bond gives rise to a partial double bond character, restricting the rotational freedom of the C2-N3 bond. *Cis* (D) and *trans* (C) conformations are defined by the angle between C1 and C4, denoted by omega. C) Angles of 0 ± 15 degrees are termed *cis*, those of 180 ± 15 degrees are termed *trans* peptidyl bonds.

In peptides, the *cis*/*trans* equilibrium can be influenced by amino acid interactions, temperature, pH and solvent; however, the equilibrium in proteins is harder to alter as the general protein fold can increase the thermodynamic advantage of a certain isomer over the other. Of course, the two isomers are not locked and *cis*/*trans* isomerization can occur in both proteins and peptides, albeit that these reactions are generally very slow. The *cis*-to-*trans* isomerization has been determined for short oligopeptides, averaging at one isomerization every six to twenty minutes[4]. In proteins, the rate of exchange is hard to determine due to their increased length and structural complexity. Hence, *cis*/*trans* isomerization is a very slow reaction on the biological timescale.

To increase the rate of isomerization, the reaction can be catalyzed by a wide-range of enzymes called peptidyl-prolyl isomerases (PPIases). Most PPIases are involved in catalyzing the *cis*/*trans* isomerization reaction directly following protein translation and are conveniently located near the ER[5,6]. The enzyme Pin1 is an interesting exception, believed to be responsible for the conformational change often following phosphorylation by proline-directed kinases. Due to the large conformational difference between *cis*- and *trans* isomers, it has been postulated that all kinases are isomer specific, implicating a crucial role for proline conformation and PPIases in the regulation of protein-protein interactions and PTMs[7,8]. For a few proteins – not exclusively kinases – this strong isomer specificity has been demonstrated; Mitogen-activated protein kinase 1, Cyclin-dependent kinase 2, protein phosphatase 2A and Chymotrypsin were reported as *trans*-specific proteins and Ssu72, the RNA polymerase 2 phosphatase, has been reported to be *cis*-specific[9–14].

Usually, the involvement of *cis*/*trans* isomerization or the isomer specificity of a certain enzyme is shown using partial degradation or a spectrophotometric assay, which both depend on the isomer-specific cleavage of the *trans* isomer by a *trans*-specific protease, subtilisin or V8 protease [11,15–19]. Interestingly though, none of these proteases cleave C- or N-terminally of proline residues, hence their isomer-specificity depends on the proximity of their cleavage site to the proline that undergoes *cis*/*trans* isomerization. If their cleavage site is not affected by the conformational change (e.g. because the two sites are separated in space), their isomer specificity cannot be observed.

To circumvent this spatial dependance of an isomer-specific protease, we focused our attention on a prolyl endopeptidase purified from Aspergillus Niger (Endopro, also known as An-PEP), which cleaves directly C-terminal to proline. EndoPro is commercially available and has found applications in food industry as well as in proteomics science[20,21]. In addition, the protease may be beneficial for use in patients suffering from gluten sensitivity, as it can efficiently degrade these proline-rich peptides[22–27]. Here, we describe how we evaluated the current status of proline isomer occurrence in human protein structures, as deposited in the PDB, and evaluate EndoPro for its use to determine the population of cis-prolines and to assess conformational change.

Materials & Methods

PDB structures were downloaded January 8th of 2020 and using an in-house script, information such as organism of origin, structure resolution, the presence of mutations and the amino acid pairs present in cis conformation were isolated from each file. All redundant or non-human structures were discarded. To determine the *cis* ratio, occurrence of amino acid combinations was determined using the amino acid sequences included in the PDB structures (i.e. pieces of protein not expressed for structure determination were not included in the analysis). Resulting data was plotted using R.

To test protease isomer specificity, several short, synthetic peptides where selected based on availability in the lab and likelihood of undergoing proline-directed phosphorylation *in vivo*. Peptide mixtures were generated as follows. Peptides were thawed and zip-tipped (u-C18, Millipore) from water (1 mM) to 100% ACN. Tips were wetted with 7 uL ACN, washed with 3x 7 uL H2O with 1% FA, 7 uL acidified peptide solution was loaded and extraction was performed in 7 uL ACN. Assuming perfect recovery, peptide concentration in ACN was 1mM. Vials were spun (15000 rpm, 10 min) to remove any undissolved material, no precipitate was visible. Peptides were mixed 1:1 and diluted with ACN to 100 uM stock solution. Six peptide mixes were prepared, three were kept in ACN (referred to as not re-equilibrated) and three were diluted 9-fold in 100 mM Ammonium Acetate buffer (pH 7.5) (referred to as re-equilibrated). Mixes were stored at 4°C for at least 24 hrs prior to analysis to ensure equilibrium was reached.

For the cleavage assay, we distinguish re-equilibrated and not re-equilibrated conditions. For the re-equilibrated peptides, the mix previously 9-fold diluted in Ammonium Acetate buffer was used. 9uL of the pepide mix was supplemented with 1 uL 100nM EndoPro solution (DSM) to initiate cleavage. The not re-equilibrated condition was produced similarly. 100uM peptide stock mix in ACN was diluted 10-fold using a mixture of 8uL Ammonium Acetate buffer and 1 uL 100 nM EndoPro solution, directly initiating cleavage. Note that final concentrations of peptides, protease and ACN are identical between the two cleavage conditions. For each timepoint, 0.5 uL was removed from the reaction mix and diluted 5-fold in ACN to deactivate the protease. Timepoints were kept on ice prior to analysis.

MS analysis was performed on a QE-UHMR mass spectrometer (Thermo Fisher Scientific) using in house gold-coated borosilicate capillaries. Standard mass spectrometer settings were used, including a capillary voltage of 1.3 kV in positive ion mode, noise level parameter 4.64, AGC off, 10 ms ion injection time, FT resolution set at 17500, S-lens RF level 200, S-lens voltage 25V, skimmer voltage 15V, Inject flatapole offset 8V, Bent flatapole DC 6V, Gate lens voltage 4V. No HCD energy was applied. Instrument calibration was performed using cesium iodide clusters, mass range was set at 500 to 1500 m/z . Extended mass range was not enabled. Spectra were viewed in Xcalibur Freestyle software

(Thermo Fisher Scientific) and converted to csv files. Using an in-house R script, the median intensities of all ions of interest were extracted for each timepoint. Intensities were normalized to the total intensity for a peptide precursor and its cleaved product ions per time point. EndoPro activity was analyzed using Prism (GraphPad), one phase exponential association was fitted to the data.

NMR spectra were recorded at DSM, enabled by Michiel Akeroyd and Peter Lankhorst, on a Ascend 600 MHz and 700 MHz spectrometer (Bruker), equipped with a 5 mm cryo probe at several probe-temperatures.

Results

First, we set out to evaluate the prevalence of *cis* isomers in protein structures. To do so, we downloaded all protein data bank protein structures and compiled a database of proteins, which included data on their organism of origin, structure resolution, the presence of mutations and the amino acid pairs present in *cis* conformation. We focused on human proteins and removed all redundant or mutated protein structures, which resulted in 3203 unique protein structures corresponding to more than 900000 amino acid pairs.

For each amino acid combination, we calculated the percentage of *cis* bonds observed, see Figure 2A and Supplemental Tables 1 and 2. The area of the dot represents the number of times a combination of amino acids was observed, and the color indicates the amount of *cis* bonds found among these observations. Overall, we found that 25% of the amino acid combinations were never found in *cis*, while others, especially pairs containing a C-terminal proline have a high percentage of *cis* bonds. See Figure 2B for combinations with a C-terminal proline. Tyrosine followed by a proline was found 1493 times, almost 15% of those observations were in *cis* conformation. This might be surprising as all other Tyrosine pairs are far less likely to be *cis* (Figure 2C). In the context of phosphorylation, the conformation of serine containing amino acid pairs is quite interesting, see Figure 2D. Serine followed by any amino acid other than proline is likely to be *trans*, however, the combination with proline is *cis* in more than 1 out of 20 observations and this might have direct implications for proline-directed phosphorylation. Across all amino acid combinations, 0.35% was present in a *cis* conformation.

Since *cis*-peptidyl bonds are quite abundant, present an estimated two times in the average protein, the isomer specificity of biology might have far more influence on the regulation of proteins then it is currently given credit for.

Making the significant observation that *cis*-Prolines are quite frequently observed in the deposited human protein structures we next went on to explore further the Proline directed protease EndoPro/ANPep, and asked the question whether this enzyme would prefer cleavage around a *cis* or *trans* conformation. To assess if EndoPro preferably cleaves *cis* or *trans* prolines, we sought to

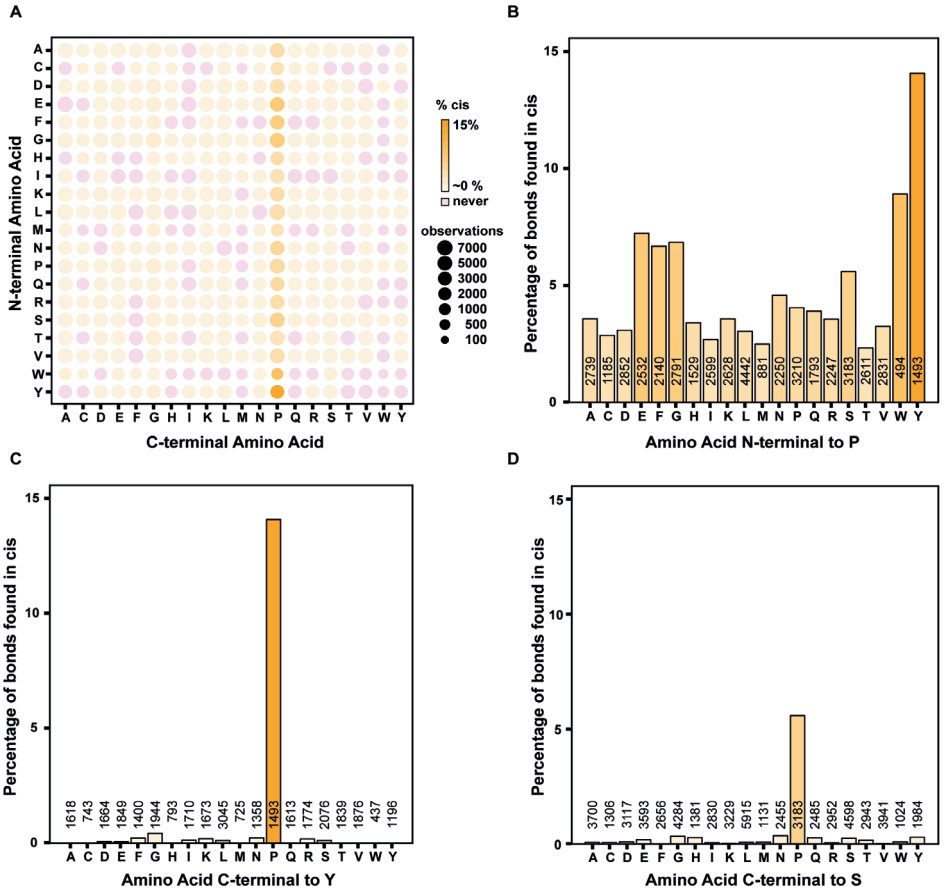


Figure 2 – Frequency of *cis*/*trans* peptidyl bonds in the human structural proteome. A) Overview of the isomerization state of all 400 possible amino acid combination. Dot size indicates the number of observations made for each amino acid pair, color indicates the percentage of these observations being *cis*. B) Bar graph showing the abundance of the *cis* conformation for each of the 20 amino acids followed by a proline residue. C) Bar graph showing the abundance of *cis* for amino acid pairs with an N-terminal tyrosine. D) Bar graph showing the abundance of cis for amino acid pairs with an N-terminal Serine. The number of observations for each combination is shown in the base of the corresponding bar.

influence the position of *cis*/*trans* equilibrium for several proline-containing peptides without losing control of peptide concentration or temperature, as changes in these conditions will affect reaction kinetics of the protease. In addition, high concentrations of salts were also not desirable as those are expected to interfere with MS analysis.

To this end, we developed an assay where peptides were initially incubated in acetonitrile and diluted tenfold into aqueous buffer prior to cleavage by EndoPro. We distinguished two conditions here: one where the peptide was given time to re-equilibrate to its aqueous conditions and one where cleavage

was initiated immediately upon dilution. It is important to note that the conditions during cleavage by Endopro were the same between these two experimental conditions, the only difference between the two conditions is whether the peptides have had time to re-equilibrate their *cis*/*trans* equilibrium or not, see Figure 3. During the cleavage reaction aliquots were taken and the protease was inactivated by a fivefold dilution in ACN, which is sufficient to inactivate the protease (data not shown).

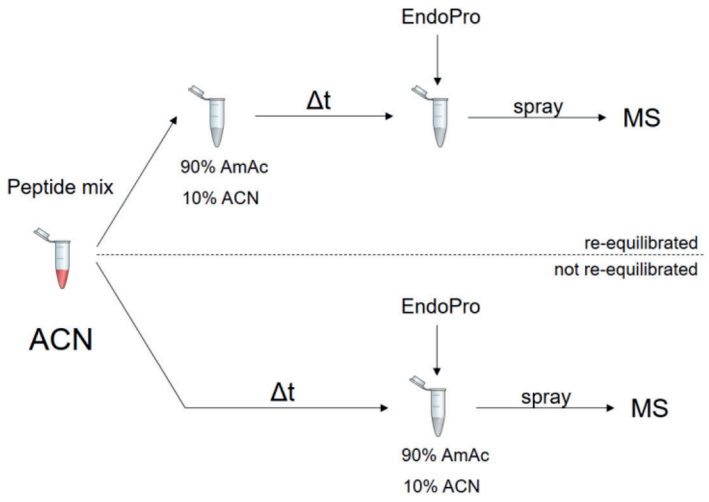
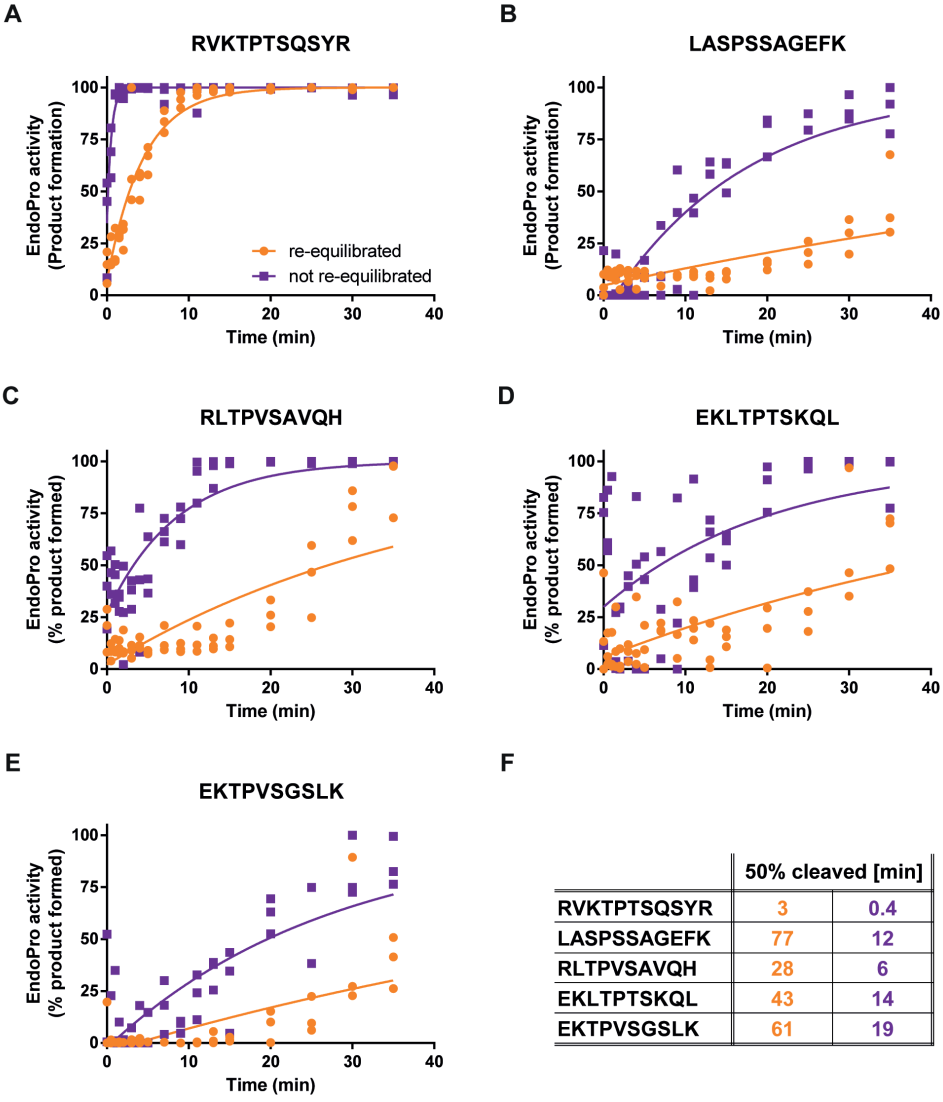


Figure 3 – Schematic representation of the EndoPro cleavage assay. Peptides were incubated in acetonitrile to shift their *cis*/*trans* equilibrium and diluted directly into the cleavage reaction buffer (not re-equilibrated) or diluted in aqueous buffer and after re-equilibration to this aqueous system, EndoPro was added to start cleavage of the peptides (re-equilibrated).

We chose five short synthetic peptides which were available in the lab and we suspected were subjected to proline-directed phosphorylation *in vivo* containing each a single proline (see Supplemental Table 3) and subjected them to our assay. We plotted the conversion to cleavage products for each peptide, see Figure 4. For each condition, we fitted a one phase exponential curve, as described in the materials and methods section. Re-equilibrated peptides (shown in orange) had time to shift their *cis*/*trans* equilibrium back to as it supposed to be in aqueous environments prior to the addition of protease, while not re-equilibrated peptides (shown in purple) were exposed to the protease directly upon dilution into aqueous buffer. For peptide RVKTPTSQSYR, the exponential curve fit was quite good under both assay conditions ($R^2 > 0.86$), revealing a decrease in enzyme activity of around 8-fold upon re-equilibration to aqueous solvent. Although the curve fit for the other peptides was not as good, they each showed a similar trend, with longer cleavage times after re-equilibration to aqueous solvent. Hence, EndoPro reveals a clear preference for peptides not re-equilibrated to aqueous conditions, which likely stems from a higher cleavage efficiency towards *trans*-proline isomers.



	50% cleaved [min]	
RVKTPTSQSYR	3	0.4
LASPSSAGEFK	77	12
RLTPVSAVQH	28	6
EKLTPTSKQL	43	14
EKTPVSGSLK	61	19

Figure 4 – Cleavage of a range of synthetic peptides by EndoPro. A-E) For each peptide, formation of the cleavage product is plotted in time. The two assay conditions, re-equilibrated to the aqueous buffer system and not re-equilibrated, are shown in orange and purple, respectively. One phase exponential association was fitted to these data, indicated by the solid lines. For RVKTPTSQSYR, the resulting fit was decent under both assay conditions ($R^2 > 0.86$), revealing a decrease in enzyme activity of 8.3-fold upon re-equilibration in aqueous buffer. The other peptides did reveal similar trends and EndoPro digestion of all synthetic peptides was found to be faster without allowing re-equilibration to aqueous environments. F) Based on one phase exponential fit, the time required to convert 50% of the peptides to cleavage fragments was estimated.

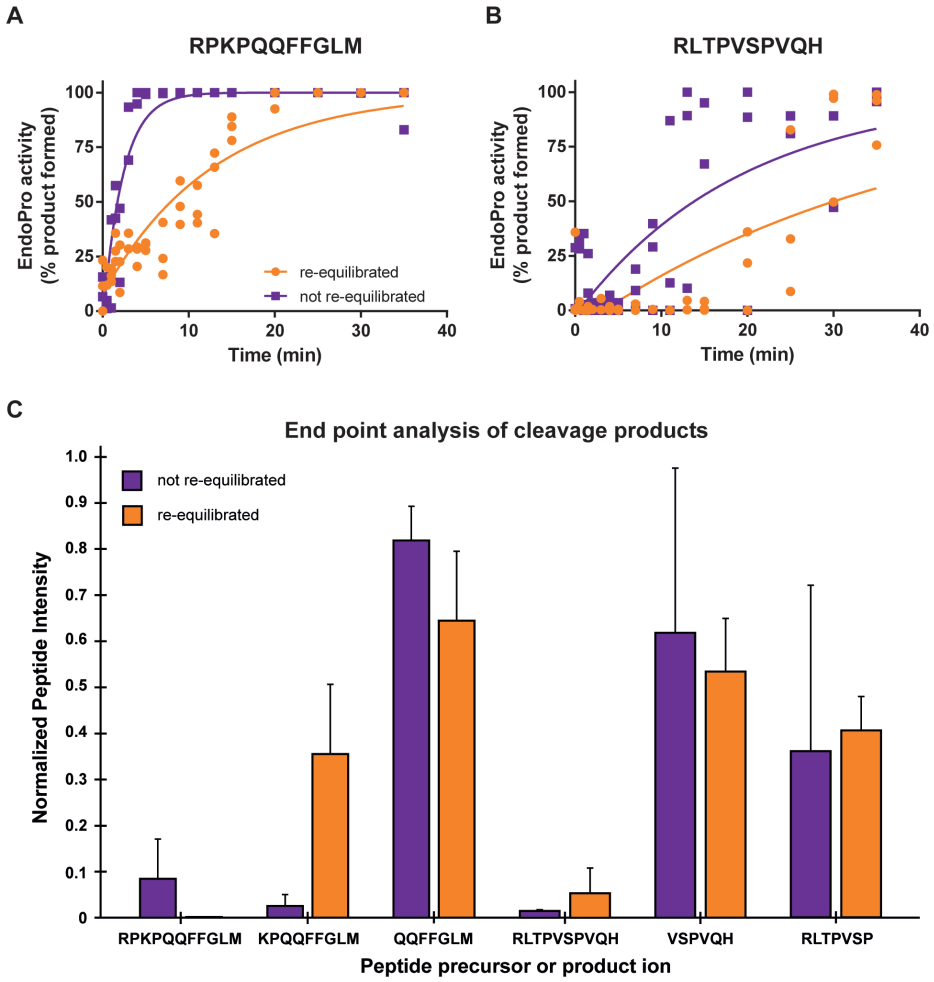


Figure 5 – EndoPro cleavage specificity in peptides containing two Pro residues; RPKPQQFFGLM and RLTPVSPVQH. A,B) The formation of cleavage products for RPKPQQFFGLM (A) and RLTPVSPVQH (B) were plotted over time. The two assay conditions, re-equilibrated to the aqueous buffer system and not re-equilibrated, are shown in orange and purple, respectively. One phase exponential association was fitted to these data, indicated by the solid lines. For RPKPQQFFGLM, the resulting fit was decent in both assay conditions ($R^2 > 0.87$), revealing a decrease in enzyme activity of 5.6-fold upon re-equilibration in aqueous buffer. C) Analysis of the final cleavage products revealed a shift in preferred cleavage site for RPKPQQFFGLM. Upon incubation in acetonitrile cleavage at the first proline appears less favorable, however, these changes were hardly significant (two-sided t-test, $\alpha = 0.05$).

Next, we investigated two peptides with two potential EndoPro cleavage sites, namely RPKPQQFFGLM and RLTPVSPVQH. Again, curve fitting of RPKPQQFFGLM ($R^2 > 0.87$) revealed a decrease in enzyme activity (5.6-fold) upon re-equilibration to aqueous solvent. RLTPVSPVQH followed the same trend, with reduced enzyme activity upon re-equilibration to aqueous buffer (Figure 5A and B). We observed no clear change in the preferential cleavage site in RLTPVSPVQH upon incubation in acetonitrile, however, products of RPKPQQFFGLM did reveal a shift in EndoPro specificity, see Figure 5C. In aqueous buffer, 36% of the cleavages had occurred C-terminal to the first proline (i.e. forming KPQQFFGLM as the hydrolysis product), whereas the not re-equilibrated peptide revealed only 3% of cleavages at this position. This would suggest that incubation in acetonitrile increases the unfavorable *cis* conformation of the first proline residue.

Finally, we set out to determine the position of *cis*/*trans* equilibrium for a few of the studied peptides in acetonitrile and deuterated water, see Figure 6. Unfortunately, due to a limited solubility of these peptides in pure acetonitrile, NMR spectra had to be recorded in 50/50 v/v ACN/D_2O mixtures, making it impossible to directly compare equilibrium positions of the peptides with the cleavage conditions used in our MS assay. Nonetheless, for all peptides accept RPKPQQFFGLM, addition of organic solvent increased the *trans*-isomer of the proline undergoing cleavage by EndoPro. For RPKPQQFFGLM, however, NMR data on the first proline shows a shift towards the *cis*-isomer upon the addition of ACN, which is in line with the decrease in cleavage by EndoPro observed at this position. Taken together, these data suggest that EndoPro favors the cleavage of *trans*-proline peptidyl bonds.

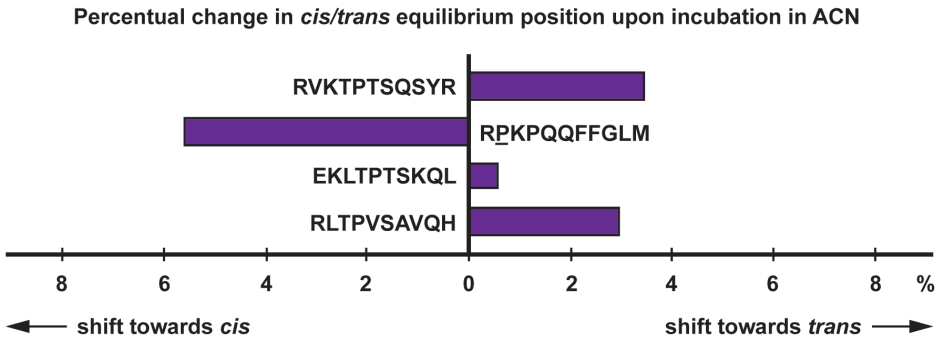


Figure 6 – Change in *cis*/*trans* equilibrium upon incubation in acetonitrile. Percentual change of the *trans* isomer was calculated from NMR measurements of individual peptides in D_2O and 50/50 D_2O/ACN, except for RVKTPTSQSYR, which was analyzed in D_2O and 90/10 ACN/D_2O. Measurements were performed at 300K. For RPKPQQFFGLM, the change for the first proline was plotted.

Discussion

Here we describe that cis/trans isomerization is present in human proteins with an average of 1 *cis* every 286 peptidyl bonds in the PDB structures of human proteins. Most of the *cis* bonds we identified occur N-terminal to a proline residue, but we also observe non-proline *cis* bonds. Our analysis extend on, but is also in agreement with, previous publications[28,29]. Some older studies show decent agreement with our data, but as less proteins structure could then be assessed they were limited in the number of observations for each peptide pair. For example, Steward *et al.* found no *cis* W-P bonds in their PDB analysis, but their data set only contained 16 observations for this peptide pair, where we observed these *cis* W-P bonds more than 400 times[29]. The substantial discrepancy in sample size is a likely explanation for the differences observed.

The true abundance of *cis* bonds in the human proteome might be significantly higher, as our analysis is likely biased towards favoring *trans* due to at least three reasons. Firstly, the PDB structures often do not contain coordinates for small, flexible loops occurring in proteins leaving a gap in the protein. For such a gap no information on any *cis* bonds is available, yet we do count these amino acid pairs in our normalization set because we count all peptide pairs between the start and end of the sequence used for structure determination. Secondly, only proteins with a PDB structure reported can be considered, which means we have a biased subset of less flexible proteins in our analysis. Finally, we counted all instances of *cis* bonds between peptide pairs, meaning an omega rotation of around 0 degrees (see Figure 1) and normalize these to the number of observations for each corresponding amino acid pair. Consequently, all other, potentially unusual omega rotations are showing up as *trans*. Therefore, we expect our analysis to be a lower indication of the actual abundance of *cis* peptidyl bonds across the human proteome.

If we look across said human proteome, the average protein length is 560 amino acids, which means we should expect about on average 2 *cis* peptidyl bonds in each human protein. Thus, *cis*/*trans* isomerization is expected to affect all human proteins of decent size and is likely to be very abundant in proline-rich members of the human proteome. We found that peptidyl bonds preceding a proline were in *cis*-conformation in 4.5% of the observations, with aromatic residues showing a higher preference for *cis* when followed by a proline than any of the other amino acids. This stronger tendency towards *cis* for aromatic amino acids preceding prolines has been attributed to their electron-donating properties[30]. Serine and threonine followed by a proline are *cis* in about 5.6% and 2.3% of the occurrences, respectively. Phosphorylation of these residues is reported to slow down *cis*/*trans* isomerization significantly and was postulated to fix the equilibrium in one of the two conformations[8,31]. This interplay between phosphorylation and *cis*/*trans* isomerization could mean that proline-directed phosphorylation is a far more complex form of protein regulation, with a potential of four subsets of conformations for a given site in a protein compared to the two subsets considered in classical phosphorylation events.

One of the most direct ways to investigate *cis*/*trans* isomerization is by NMR and many early works focusing on *cis*/*trans* isomerization and its kinetics have been done using NMR. Here, we tried to use NMR to assess the direction of change induced by incubation in ACN, however, substantial differences between NMR and mass spectrometry led to discrepancies, making it impossible to directly compare these two sets of data. As our cleavage assays were optimized for mass spectrometric detection of peptides and products, they rely on a volatile buffer at pH 7.5 for digestion and use only low concentrations of peptide which are pre-incubated in 100% ACN. These conditions, however, were not compatible with NMR, as the technique is less sensitive. It required higher concentrations of the peptides, which were impossible to achieve in 100% ACN due to limited peptide solubility. Hence, NMR measurements shown here were performed at much higher peptide concentrations and at lower concentration of organic solvent. Previous studies have indicated that trace amounts of water can drastically influence the conformational equilibrium [32]. Therefore, the shifts observed for these peptides likely do not encompass the shifts achieved in the assay.

Regardless of the uncertainty in the amount of equilibrium shift, these experiments show that the *cis*/*trans* equilibrium of the investigated peptides is affected by incubation in organic solvent. Furthermore, our EndoPro digestion assay reveals a clear memory effect induced by this incubation in acetonitrile. Since *cis*/*trans* isomerization is the rate limiting step in conformational changes and folding[2,33], it is likely the source of the differences in peptide cleavage kinetics we observed here. Since pH is known to affect the position *cis*/*trans* equilibrium, this isomer-specificity might also explain the double activity optimum previously reported for EndoPro[34].

Conclusion

Overall, we observe a clear perturbation of the *cis*/*trans* equilibrium upon incubation of peptides in acetonitrile and a large effect of re-equilibration of these peptides on the cleavage kinetics of EndoPro. Hence, EndoPro likely does prefer to cleave C-terminal to *trans* proline, which would be in line with reports for other isomer-specific proteases. Since EndoPro directly cleaves C-terminal to proline, it would have a clear advantage over other currently reported isomer-specific proteases. However, its utility in partial digestion experiments on intact proteins remains to be confirmed to further investigate EndoPro as a tool for studying isomer-specificity in biology.

Acknowledgements

The authors would like to thank Henk van der Toorn for the collection of all the pdb files.

References

1 Fort KL, Silveira JA, Pierson NA, Servage KA, Clemmer DE & Russell DH (2014) From Solution to the Gas Phase: Factors That Influence Kinetic Trapping of Substance P in the Gas Phase. J Phys Chem B 118, 14336–14344.

2 Wedemeyer WJ, Welker E & Scheraga HA (2002) Proline Cis–Trans Isomerization and Protein Folding †. Biochemistry 41, 14637–14644.

3 MacArthur MW & Thornton JM (1991) Influence of proline residues on protein conformation. J Mol Biol 218, 397–412.

4 Grathwohl C & Wüthrich K (1981) Nmr studies of the rates of proline cis - trans isomerization in oligopeptides. Biopolymers 20, 2623–2633.

5 Gothel SF & Marahiel M a (1999) Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts. Cell Mol Life Sci 55, 423–436.

6 Schmidpeter PAM & Schmid FX (2015) Prolyl Isomerization and Its Catalysis in Protein Folding and Protein Function. J Mol Biol 427, 1609–1631.

7 Lu KP (2004) Pinning down cell signaling, cancer and Alzheimer's disease. Trends Biochem Sci 29, 200–2009.

8 Lu KP, Finn G, Lee TH & Nicholson LK (2007) Prolyl cis-trans isomerization as a molecular timer. Nat Chem Biol 3, 619–629.

9 Zhou XZ, Kops O, Werner A, Lu P-J, Shen M, Stoller G, Kullertz G, Stark M, Fischer G & Lu KP (2000) Pin1-dependent Prolyl Isomerization Regulates Dephosphorylation of Cdc25C and Tau Proteins. Mol Cell 6, 873–883.

10 Kofron JL, Kuzmic P, Kishore V, Colón-Bonilla E & Rich DH (1991) Determination of kinetic constants for peptidyl prolyl cis-trans isomerases by an improved spectrophotometric assay. Biochemistry 30, 6127–6134.

11 Fischer G, Bang H & Mech C (1984) [Determination of enzymatic catalysis for the cis-trans-isomerization of peptide binding in proline-containing peptides]. Biomed Biochim Acta 43, 1101–11.

12 Weiwad M, Küllertz G, Schutkowski M & Fischer G (2000) Evidence that the substrate backbone conformation is critical to phosphorylation by p42 MAP kinase. FEBS Lett 478, 39–42.

13 Werner-Allen JW, Lee C-J, Liu P, Nicely NI, Wang S, Greenleaf AL & Zhou P (2011) cis -Proline-mediated Ser(P) 5 Dephosphorylation by the RNA Polymerase II C-terminal Domain Phosphatase Ssu72. J Biol Chem 286, 5717–5726.

14 Brown NR, Noble MEM, Endicott JA & Johnson LN (1999) The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. Nat Cell Biol 1, 438–443.

15 Stukenberg PT & Kirschner MW (2001) Pin1 Acts Catalytically to Promote a Conformational Change in Cdc25. Mol Cell 7, 1071–1083.

16 Jalouli M, Dery M-AC, Lafleur VN, Lamalice L, Zhou XZ, Lu KP & Richard DE (2014) The prolyl isomerase Pin1 regulates hypoxia-inducible transcription factor (HIF) activity. Cell Signal 26, 1649–1656.

17 Zacchi P, Gotissa M, Uchida T, Salvagno C, Avollo F, Volinina S, Ronal Z, Blandino G, Schneider C & Del Sal G (2002) The prolyl isomerase Pin1 reveals a mechanism to control p53 functions after genotoxic insults. Nature 419, 853–857.

18 Steger M, Murina O, Hühn D, Ferretti LP, Walser R, Hänggi K, Lafranchi L, Neugebauer C, Paliwal S, Janscak P, Gerrits B, Del Sal G, Zerbe O & Sartori AA (2013) Prolyl Isomerase PIN1 Regulates DNA Double-Strand Break Repair by Counteracting DNA End Resection. Mol Cell 50, 333–343.
19 Caporale A, Mascanzoni F, Farina B, Sturlese M, Di Sorbo G, Fattorusso R, Ruvo M & Doti N (2016) FRET-Protease-Coupled Peptidyl-Prolyl cis-trans Isomerase Assay. J Biomol Screen 21, 701–712.
20 Akeroyd M, van Zandycke S, den Hartog J, Mutsaers J, Edens L, van den Berg M & Christis C (2016) AN-PEP, Proline-Specific Endopeptidase, Degrades All Known Immunostimulatory Gluten Peptides in Beer Made from Barley Malt. J Am Soc Brew Chem 74, 91–99.
21 van der Laarse SAM, van Gelder CAGH, Bern M, Akeroyd M, Olsthoorn MMA & Heck AJR (2020) Targeting proline in (phospho)proteomics. FEBS J 287, 2979–2997.
22 König J, Holster S, Bruins MJ & Brummer RJ (2017) Randomized clinical trial: Effective gluten degradation by Aspergillus niger-derived enzyme in a complex meal setting. Sci Rep 7, 13100.
23 Salden BN, Monserrat V, Troost FJ, Bruins MJ, Edens L, Bartholomé R, Haenen GR, Winkens B, Koning F & Masclee AA (2015) Randomised clinical study: Aspergillus niger -derived enzyme digests gluten in the stomach of healthy volunteers. Aliment Pharmacol Ther 42, 273–285.
24 Janssen G, Christis C, Kooy-Winkelaar Y, Edens L, Smith D, van Veelen P & Koning F (2015) Ineffective Degradation of Immunogenic Gluten Epitopes by Currently Available Digestive Enzyme Supplements. PLoS One 10, e0128065.
25 Montserrat V, Bruins MJ, Edens L & Koning F (2015) Influence of dietary components on Aspergillus niger prolyl endoprotease mediated gluten degradation. Food Chem 174, 440–445.
26 Tack GJ, Van de Water JMW, Bruins MJ, Kooy-Winkelaar Y, Van Bergen J, Bonnet P, Vreugdenhil ACE, Korponay-Szabo I, Edens L, von Blomberg BME, Schreurs MWJ, Mulder CJ & Koning F (2013) Consumption of gluten with gluten-degrading enzyme by celiac patients: A pilot-study. World J Gastroenterol 19, 5837.
27 Mitea C, Havenaar R, Drijfhout JW, Edens L, Dekking L & Koning F (2007) Efficient degradation of gluten by a prolyl endoprotease in a gastrointestinal model: implications for coeliac disease. Gut 57, 25–32.
28 Jabs A, Weiss MS & Hilgenfeld R (1999) Non-proline Cis peptide bonds in proteins. J Mol Biol 286, 291–304.
29 Steward D, Sakar A & Wampler J (1990) Occurrence and Role of Cis Peptide bonds in Protein Structures. J Mol Biol 214, 253–260.
30 Thomas KM, Naduthambi D & Zondlo NJ (2006) Electronic Control of Amide cis−trans Isomerism via the Aromatic−Prolyl Interaction. J Am Chem Soc 128, 2216–2217.
31 Hamelberg D, Shen T & McCammon JA (2005) Phosphorylation Effects on cis/trans Isomerization and the Backbone conformation of Serine-Proline Motifs: Accelerated MD Analysis. JACS 127, 1969–1974.
32 Kunz C, Jahreis G, Günther R, Berger S, Fischer G & Hofmann H-J (2012) Influence of lithium cations on prolyl peptide bonds. J Pept Sci 18, 400–404.
33 Schultz DA, Baldwin RL & Schmid

FX (1992) Cis proline mutants of ribonuclease A. II. Elimination of the slow-folding forms by mutation. Protein Sci 1, 917–924.
34 Tsiatsiani L, Akeroyd M, Olsthoorn M & Heck AJR (2017) Aspergillus niger Prolyl Endoprotease for Hydrogen–Deuterium Exchange Mass Spectrometry and Protein Structural Studies. Anal Chem 89, 7966–7973.
35 Hornbeck P V, Zhang B, Murray B, Kornhauser JM, Latham V & Skrzypek E (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 43, D512–D520.
36 The Uniprot Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43, D204–D212.

Supplemental Information

Supplemental Table 1 - Occurence of different amino acid combinations across the human proteome.

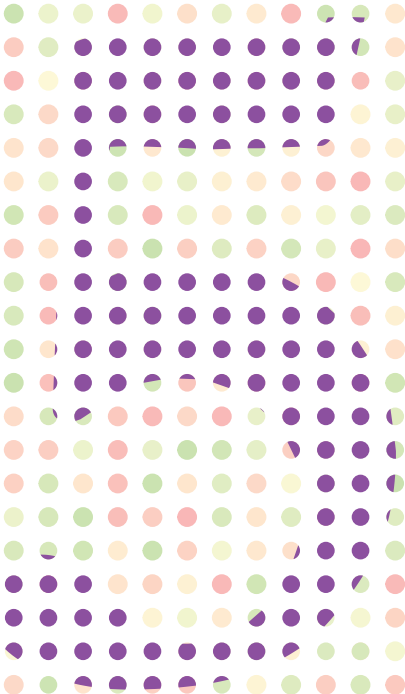
C-terminal Amino Acid	N-terminal Amino Acid																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4564	1112	2717	4308	2046	3642	1155	2694	3567	5666	1618	1969	2979	2611	2889	3700	3054	4027	745	1618
C	1170	408	1091	1223	795	1170	656	1048	1078	1761	380	813	882	754	994	1306	1112	1422	284	743
D	2882	1075	2467	3752	2118	3166	900	2363	3271	4697	1203	1589	2749	2019	2680	3117	2440	3318	709	1664
E	3976	1124	3124	5349	2275	3778	1182	2688	4229	6012	1715	2175	4059	2889	3126	3593	2943	3907	1036	1849
F	2419	697	2753	2448	1463	2546	1143	1926	2076	3153	744	1766	1879	1434	1911	2656	2165	2570	454	1400
G	3883	1299	3237	3733	2593	4011	1524	2619	3331	5204	1208	2509	3419	2371	2861	4284	3270	3545	786	1944
H	1380	607	1242	1438	1059	1477	629	1389	1307	2444	454	961	1091	1012	1143	1381	1184	1578	349	793
I	2831	947	3041	3314	1761	2981	1346	2517	3150	3633	855	2359	1908	2091	2528	2830	2442	3119	636	1710
K	3198	1127	2668	4334	1921	3751	1093	2995	3755	5773	1361	2174	2478	2259	2797	3229	2419	3731	862	1673
L	5838	1842	4970	5625	3896	5046	2370	4484	4944	8568	1837	3776	3895	4048	4412	5915	4751	5777	1153	3045
M	1235	323	1041	1468	704	1340	493	998	1220	1832	438	848	748	851	1001	1131	966	1272	394	725
N	1932	732	1827	2883	1512	2310	937	2059	2354	3476	848	1590	1926	1566	2040	2455	1712	2317	708	1358
P	2739	1185	2852	2532	2140	2791	1529	2599	2628	4442	881	2250	3210	1793	2247	3183	2611	2831	494	1493
Q	2307	950	1741	2348	1925	2418	1018	2253	2089	4432	708	1412	1942	1770	1832	2485	1912	2440	532	1316
R	2925	1091	2143	3003	2024	3050	1393	2357	2482	5144	871	1702	2251	1962	2569	2952	2093	2860	632	1774
S	3801	1390	3082	3246	2856	4205	1598	3333	3021	5936	1219	2275	3384	2155	2657	4598	3015	3881	993	2076
T	2981	1034	2257	2747	2217	3584	1106	2712	2600	4648	1037	1853	2197	1663	2217	2943	2248	4036	731	1839
V	4287	1180	3369	4387	2283	3610	1456	2805	3746	5386	1459	2391	3152	2601	3068	3941	3791	4116	703	1876
W	726	315	707	842	486	775	404	604	711	1086	266	641	776	664	650	1024	737	865	181	437
Y	1656	626	1815	2026	1470	2015	935	1522	1947	2783	602	1453	1452	1308	1575	1984	1761	1950	506	1196

Supplemental Table 2 - Ratio of cis-bonds over total occurences

N-terminal Amino Acid	C-terminal Amino Acid																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.04%	0.17%	0.14%	0.08%	0.04%	0.41%	0.07%	0.00%	0.09%	0.03%	0.24%	0.16%	3.58%	0.17%	0.14%	0.18%	0.03%	0.05%	0.00%	0.06%
C	0.00%	0.49%	0.28%	0.00%	0.14%	0.08%	0.33%	0.00%	0.00%	0.11%	0.00%	0.14%	2.87%	0.21%	0.09%	0.00%	0.00%	0.00%	0.00%	0.32%
D	0.07%	0.18%	0.28%	0.16%	0.11%	0.43%	0.08%	0.00%	0.15%	0.08%	0.29%	0.33%	3.09%	0.34%	0.09%	0.26%	0.04%	0.00%	0.14%	0.00%
E	0.00%	0.00%	0.19%	0.09%	0.12%	0.54%	0.21%	0.00%	0.16%	0.02%	0.07%	0.21%	7.23%	0.04%	0.10%	0.12%	0.07%	0.07%	0.00%	0.15%
F	0.05%	0.13%	0.19%	0.04%	0.07%	0.39%	0.00%	0.00%	0.05%	0.03%	0.00%	0.00%	6.68%	0.00%	0.00%	0.14%	0.09%	0.09%	0.00%	0.07%
G	0.33%	0.26%	0.32%	0.29%	0.20%	0.97%	0.27%	0.10%	0.29%	0.14%	0.07%	0.48%	6.84%	0.21%	0.33%	0.36%	0.14%	0.25%	0.00%	0.15%
H	0.00%	0.46%	0.22%	0.00%	0.00%	0.33%	0.32%	0.07%	0.18%	0.04%	0.20%	0.00%	3.40%	0.10%	0.07%	0.19%	0.09%	0.00%	0.00%	0.00%
I	0.07%	0.00%	0.04%	0.00%	0.00%	0.19%	0.00%	0.00%	0.03%	0.02%	0.20%	0.15%	2.69%	0.00%	0.00%	0.00%	0.04%	0.11%	0.00%	0.00%
K	0.14%	0.09%	0.24%	0.05%	0.10%	0.42%	0.08%	0.03%	0.11%	0.04%	0.00%	0.08%	3.58%	0.19%	0.16%	0.03%	0.04%	0.05%	0.14%	0.15%
L	0.04%	0.06%	0.13%	0.10%	0.00%	0.21%	0.00%	0.00%	0.09%	0.01%	0.05%	0.00%	3.04%	0.02%	0.02%	0.05%	0.09%	0.02%	0.18%	0.04%
M	0.12%	0.00%	0.00%	0.06%	0.00%	0.50%	0.00%	0.00%	0.07%	0.05%	0.00%	0.35%	2.50%	0.00%	0.00%	0.33%	0.00%	0.14%	0.00%	0.00%
N	0.15%	0.37%	0.00%	0.23%	0.06%	0.16%	0.21%	0.04%	0.14%	0.00%	0.00%	0.13%	4.58%	0.14%	0.18%	0.04%	0.00%	0.08%	0.00%	0.14%
P	0.20%	0.11%	0.07%	0.27%	0.05%	0.67%	0.18%	0.00%	0.32%	0.08%	0.00%	0.10%	4.05%	0.31%	0.22%	0.12%	0.32%	0.16%	0.13%	0.28%
Q	0.08%	0.00%	0.25%	0.07%	0.21%	0.17%	0.10%	0.00%	0.09%	0.07%	0.00%	0.19%	3.90%	0.17%	0.10%	0.14%	0.18%	0.08%	0.00%	0.00%
R	0.03%	0.10%	0.37%	0.19%	0.00%	0.38%	0.09%	0.08%	0.07%	0.05%	0.20%	0.15%	3.56%	0.05%	0.04%	0.15%	0.05%	0.00%	0.00%	0.00%
S	0.08%	0.08%	0.10%	0.19%	0.00%	0.35%	0.29%	0.07%	0.03%	0.08%	0.09%	0.37%	5.59%	0.28%	0.07%	0.26%	0.17%	0.03%	0.10%	0.30%
T	0.10%	0.00%	0.20%	0.10%	0.00%	0.40%	0.25%	0.00%	0.17%	0.04%	0.00%	0.12%	2.34%	0.00%	0.10%	0.10%	0.00%	0.03%	0.00%	0.06%
V	0.05%	0.07%	0.09%	0.05%	0.00%	0.20%	0.06%	0.03%	0.03%	0.10%	0.08%	0.17%	3.25%	0.04%	0.07%	0.03%	0.02%	0.05%	0.00%	0.05%
W	0.54%	0.35%	0.00%	0.29%	0.22%	0.13%	0.00%	0.00%	0.00%	0.00%	0.00%	0.14%	8.91%	0.00%	0.00%	0.30%	0.00%	0.00%	0.00%	0.00%
Y	0.00%	0.00%	0.06%	0.05%	0.21%	0.41%	0.00%	0.12%	0.18%	0.10%	0.00%	0.22%	14.07%	0.00%	0.17%	0.10%	0.00%	0.00%	0.00%	0.00%

Supplemental Table 3 - Synthetic peptides used in the EndoPro digestion assays. Modifications as listed in the phosphosite and/or Uniprot database [35,36].

Peptide Sequence	Corresponding Protein	Uniprot ID	AA index	Modifications	Remarks
RVKTPTSQSYR	Zinc Finger Protein 281	Q9Y2X9	885 - 895	K887-ub T888-p S891-gl	
LASPSSAGEFK	Spindle apparatus protein lin-5	P45970	742 - 752	S744-p	
RLTPVSPVQH RLTPVSAVQH	DNA-binding protein RFX7	Q2KHR2	1023 - 1032	T1025-p S1028-p	Second proline was substituted by alanine in initial experiments
EKLTPTSKQL	WD repeat-containing protein 75	Q8IWA0	675 - 684	K676-ub T678-p T680-p K682-ub	
EKTPVSGSLK	THO complex subunit 2	Q8NI27	1383 - 1392	T1385-p S1388-p S1390-p	
RPKPQQFFGLM	TKN1_HUMAN	P20366	57 - 68	K60-ac	Not included in initial experiments



Chapter 5

Probing the phospho-dependant Pin1 interactome

Saar A.M. van der Laarse and Albert J.R. Heck

Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences,
Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

a preliminary study

Introduction

As we have seen in the previous chapters, all amino acids exhibit some extent of double bond character around their peptidyl bonds to other amino acids. Because of this double bond behavior, the carbon-nitrogen bond is restricted in its rotational freedom and the peptidyl bond can exist in two isoforms: *cis* or *trans*. In Chapter 4, we established that this double bond character is especially strong N-terminal to proline residues, due to their more basic, tertiary nitrogen atom. Therefore, peptidyl-proline bonds are found in the *cis* conformation more frequently than any other amino acid combination.

In human biology, the *cis*/*trans* isomerization around prolines is often the rate limiting step in protein folding[1,2]. Uncatalyzed isomerization is extremely slow on a biological timescale, with a rate constant in the order of 0.002 s^{-1}[3]. To overcome this, the human proteome contains a whole arsenal of peptidyl-prolyl isomerases (PPIases). These proteins catalyze the *cis*/*trans* isomerization of peptidyl-prolyl bonds present in their substrates. Generally, PPIases are divided into three families: cyclophilins, FP-504 binding proteins (FPBPs) and parvulins, all with a similar PPIase-domain[4–8].

In addition to protein folding, prolines play an important role in protein phosphorylation. Human kinases tend to target (i.e. cyclin-dependent kinases and mitogen-associated protein kinases) or disfavor (i.e. calmodulin-dependent protein kinase) prolines in their phosphorylation motifs[9–11]. Interestingly, these two proline-centered effects can also be combined to enhance the sophistication of the regulation of phosphorylated proteins. For example, studies on kinase and phosphatase specificity have revealed that these enzymes can have a clear preference for either the *cis* or the *trans* proline isomer of their substrate proteins[12–14]. Furthermore, phosphorylation of the serine or threonine residue preceding proline was shown to drastically reduce the rate of *cis*/*trans* isomerization, effectively locking the proline in the isomeric state it was phosphorylated in[15].

Once the residue preceding proline is phosphorylated, most of the conventional PPIases can no longer act on that particular proline site. This is where the protein Pin1 comes into play. Pin1 is a small, 18kDa protein that contains a WW-domain and a PPIase domain, which are connected by a flexible linker (Figure 1). Relying on its WW-domain, Pin1 has been shown to preferentially bind Ser-Pro and Thr-Pro motifs in substrate proteins in a phosphorylation-dependent manner[16]. Simply stated: once the residue preceding proline is phosphorylated, regular PPIases can no longer act, but the resulting motif is specifically targeted by Pin1 and this interaction can catalyze a conformational change in the substrate[17,18].

Due to this unique function, Pin1 has been studied quite extensively. Using KOs, it was observed that Pin1 is essential in some organisms, however, not all organisms fully depend on the presence of Pin1, indicating some form of

redundancy[19–21]. Pin1 is regulated by phosphorylation and both inhibitory and activating effects have been reported, depending on the residue that is modified[22,23]. For example, phosphorylation on serine-16 and serine-71 have been reported to inhibit Pin1, likely because it prevents the phosphorylation-specific binding of substrates to the WW-domain and PPIase domain, respectively. In contrast, phosphorylation of Pin1 serine-65 was reported to extend protein life time by preventing ubiquitination-driven proteasomal degradation[24]. The protein has been connected to many pathologies, including multiple neurodegenerative disorders and cancers[25–29].

Despite all of these works, the exact mechanism of action for Pin1 is not yet fully understood. Mutational studies have been performed to evaluate the importance

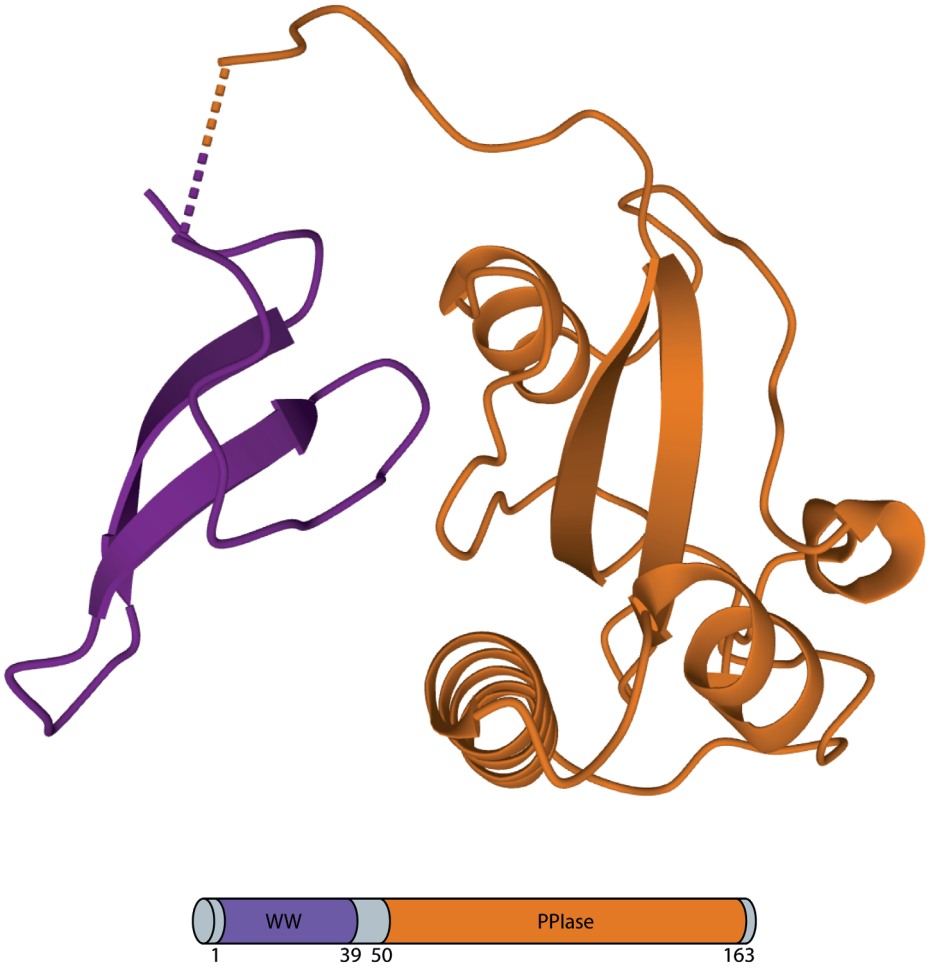


Figure 1 - The Structural model of Pin1, highlighting its WW-domain that binds Ser-Pro and Thr-Pro in a phosphorylation-dependent manner in purple and a PPIase domain responsible for catalyzing its substrates *cis*/*trans* isomerization in orange[22,26–28]. Figure adapted from [28].

of certain amino acids in the protein for its PPIase activity, however, none of the mutations resulted in a fully deactivated Pin1 protein[30–32]. Adding to the complexity: small levels of PPIase activity were demonstrated simply by generating an appropriately sized cavity within unrelated proteins[33]. Hence, PPIase activity might be the result of binding and releasing, essentially reshuffling the peptidyl-prolyl conformations of target proteins, rather than a chemical reaction occurring in a PPIase active site[34].

In addition to the uncertainty surrounding the PPIase domain and its mechanism, the relation between Pin1's WW-domain and its PPIase domain is also unclear. Several studies have determined that binding events on one domain induce structural changes in the other, indicating allosteric modulation between the two domains across the flexible linker[35–37]. To date, four main hypotheses have been reported surrounding the mechanism of action for the two domains of Pin1, which have been nicely summarized in a recent review[38].

In short, the WW-domain may bind substrates on phosphorylated serine/threonine proline motifs, allowing the PPIase domain to act on a second proline or occupy and isomerize the proline initially bound to the WW-domain[39,40]. A variation on this is the simultaneous binding mechanism, where the WW domain and PPIase domain bind two separate serine/threonine proline sites on the same protein at the same time. Alternatively, the WW domain may bind an active proline-directed kinase, ensuring that the PPIase domain of Pin1 is close to any newly phosphorylated serine/threonine proline motifs on the enzymes substrates[41]. Finally, the WW domain may bind any *trans* serine/threonine proline motifs on the substrate to stabilize the *trans* isomer. Should the substrate be in the *cis* isoform, the PPIase domain catalyzes the *cis*-to-*trans* isomerization prior to stabilization through binding of the WW domain occurs[42–44]. Which of these mechanisms is correct, or whether several of these may occur in parallel, remains to be elucidated.

In this study, we set out to develop an affinity-based pull-down to evaluate Pin1 interactors. In future work, the approach could also be used to assess substrates of the two Pin1 domains separately, or investigate PTM-dependent interactors based on phosphorylations on various sites or other protein modifications.

Materials and Methods

Expression of Pin1
E. Coli BL21DE3 transformed with his-tagged Pin1 were used to inoculate 5 mL LB with ampicillin (100 ug/mL). The bacteria were cultured ON at 37 degrees Celsius, while shaking at 220 rpm. The following morning, cultures were diluted 10x and OD600 was monitored. At OD600 = 0.960, 400 µL of culture was diluted into 400 mL of auto induction media and 400 mL of LB with ampicillin and left shaking at 37 degrees Celsius, 220 rpm ON. Bacteria were pelleted by centrifugation and pellets were stored at -80 degrees Celsius until use.

Preparation of beads
A pellet from 30 mL bacterial culture with and without His-Pin1 expression was thawed on ice and resuspended in 10 mL E. Coli lysis buffer consisting of 50 mM HEPES, pH 7.5, 50 mM NaCl and 1mM TCEP with EDTA free protease inhibitors (Complete Mini, Roche). Suspensions were sonicated on ice (cycle 0.6, amplitude 70%, 5 rounds of 1 min with breaks) and subsequently spun down (14000 rpm, 4 degrees Celsius, 30 min) to remove cell debris. A 30 µL aliquot was kept on ice for analysis of bead loading.

Next, 250 µL 50% Ni sepharose beads slurry (Merck) was washed twice with 1 mL milliQ water followed by three washes with 1mL E. Coli lysis buffer each. Vials were spun down between wash steps (500 x g, 3 min, RT) and supernatant was removed. After washing, a 50% slurry was prepared in E. Coli lysis buffer and 200 µL of 50% beads slurry was added to E. Coli lysate with Pin1 expression, whereas 50 µL beads slurry were added to lysate without Pin1 expression. Rotated head over head at 4 degrees Celsius ON to allow binding of Pin1 to nickel resin.

The next morning, beads were spun down (500 x g, 5 min, 4 degrees Celsius) and supernatant was removed (30 µL aliquot was kept for analysis). Beads were washed twice with 1 mL E. Coli lysis buffer, once with 1 mL high salt (50mM HEPES, pH = 7.5, 300 mM NaCl and 1 mM TCEP, Complete Mini, EDTA free protease inhibitors) and twice with 1 mL E. Coli lysis buffer.

A 50% slurry of beads was made using E. Coli lysis buffer and the Pin1 containing beads were split in two conditions, either subjected to phosphorylation or not. To each vial, 2 uL PKAc (~40 uM) was added. To phosphorylate Pin1, 2 µl MgATP (100 mM) was added, for the unphosphorylated conditions, 2 µL milliQ water was used instead. Vials were incubated for 6 hrs at 30 degrees Celsius, while shaking at 1000 rpm. Thereafter, beads were pelleted (500 x g, 4 degrees C, 3 min) and supernatant was removed. Next, beads were washed three times with 1mL binding buffer (50 mM Trizma preset pH 7.4, 50 mM NaCl, Complete Mini EDTA free protease inhibitors, PHOSstop phosphatase inhibitors [Roche]) to remove any residual kinase or ATP. Beads were diluted to a 25 % slurry for the pull-down experiment.

Assessment of bead loading
10.5 µL SDS buffer (1.5% SDS in milliQ water) was added to 6 µL of each 25% bead slurry and heated to 95 degrees Celsius for 10 min, 1000 rpm. Spin down and transfer supernatant to a new tube. Add 10 µL 4x loading buffer (Bio-Rad, info) supplemented with SDS to each lysate or flowthrough sample vial (30 µL) and 5 µL to boiled bead samples (~15 µL). Samples were heated (96 degrees C, 5 min, 300 rpm) and spun down with a tabletop centrifuge. For the lysate samples 25 uL was loaded on the gel, for protein samples from beads all 20 µL was loaded. SDS-Page was performed using the Bio-Rad gel system, XT sample buffer 4x and XT MOPS running buffer and Criterion XT gels (4-12% Bis-Tris) in combination with Imperial Protein Stain (Thermo Fisher Scientific).

Pull-down experiment
HEK293T cell pellet (obtained from D. Fasci) was diluted 5x in PBS lysis buffer (Phosphate buffered saline, Complete mini EDTA free protease inhibitors[Roche], 0.5% TX100, Phosphatase inhibitors [Roche]) and left on ice for approximately 20 minutes to lyse. Lysate was split across Eppendorf tubes and spun down (14000 rpm, 4 degrees Celsius, 40 min). Supernatant was pooled and protein concentration was estimated using a BCA kit (info). Lysate was diluted with PBS lysis buffer to 1.5 mg/mL protein concentration. Four Eppendorf vials were filled with 1.2 mL HEK293T lysate each and 6 uL of 25% bead slurry was added. Beads were either not incubated in E. Coli background (no BG), incubated in E. Coli background without Pin1 expression, incubated in E. Coli with Pin1 but not phosphorylated or incubated with E. Coli lysate containing Pin1 and phosphorylated. Lysates with beads were left to rotate head over head at 4 degrees Celsius ON.

Beads were spun down (500 x g, 4 degrees Celsius, 3 min) and supernatant was removed. Beads were washed with 3x 1 mL binding buffer. Next, 20 µL elution 1 buffer (8M Urea in 50 mM Ammonium Bicarbonate, pH = 8) was added to each vial and they were left rotating head over head at RT for 30 min. Vials were spun down (500 x g, 4 degrees Celsius, 3 min), supernatant was transferred to a new vial and beads were resuspended in another 20 µL elution 1 buffer, left to rotate at RT for 30 minutes and were spun down again (500 x g, 4 degrees Celsius, 3 min). Supernatant was stored in a separate vial.
Next, beads were resuspended in 20 µL elution 2 buffer (8M Urea in 50 mM Ammonium Bicarbonate, pH = 8, 250 mM imidazole, freshly prepared and kept in the dark) and rotated head over head at RT for 30 minutes. Vials were spun down (500 x g, 4 degrees Celsius, 3 min) and supernatant was transferred to a new vial, supplemented with 20 µL elution 1 buffer to reduce imidazole concentration.

Reduction, alkylation and digestion
A 20 mM DTT stock was freshly prepared by dissolving 2 mg DTT in 650 uL 50 mM Ammonium Bicarbonate buffer and 10 µL was added to each vial and incubated at 60 degrees Celsius, 300 rpm for 1 hr. Next, a 40 mM IAA stock was freshly prepared by dissolving 2 mg IAA in 270 µL 50 mM Ammonium

bicarbonate buffer. Subsequently, 10 µL was added to each vial and incubated for 30 min in the dark (RT). Finally, another 10 µL of 20 mM DTT stock was added to each vial to quench the IAA.

LysC stock was thawed and diluted in 50 mM Ammonium Bicarbonate to 0.2 µg/µL. 3 µL of the protease was added to each vial and incubated at 37 degrees C for 4 hrs. Next, 87 µL 50 mM Ammonium Bicarbonate buffer was added to reduce the urea concentration to 2M and 0.6 ug Trypsin in 50 mM ammonium carbonate buffer was added to each vial. Samples were incubated at 37 degrees C ON.

Peptide cleanup
Stage tips were prepared using a C18 paper plug and 30 uL ReproSil (C18, 3 µM particles) 25% slurry, tips were spun at 500 x g for 10 min to pack (RT). Columns were washed with 50 µL ACN, 2x 50 µL B (80% ACN, 0.6% Acetic Acid) and 2x 50 µL A (99.4% water, 0.6% Acetic Acid). Samples were acidified to pH 2 using FA and loaded on the tips. Peptides were washed with 2x 100 µL A and finally eluted in 50 µL B. Eluent solvents were removed in speedvac (SPD, Thermo Savant) and vials were stored at -20 degrees Celsius prior to analysis.

LCMS analysis
Upon analysis, vials were thawed on ice and resuspended in 20 µL 10% FA and spun down (14000 rpm, 10 min, 4 degrees Celsius). 1 µL sample was diluted in 49 µL 10% FA and 2 µL was injected. LC-MS/MS analysis was performed using an Agilent 1290 Infinity System (Agilent Technologies) in combination with a Q Exactive Biopharma (Thermo Fisher Scientific). Nanoflow RP separation was carried out using a 2 cm trap column (100 µM ID) packed with ReproSil C18 material (3 µM particles), followed by a 50 cm analytical column (75 µM ID, packed with PoroShell 120 EC-C18, 2.7 µM particles) and a 120-minute gradient. Mobile phases consisted of 0.1% FA in water (A) and 0.1% FA and 80% ACN (B). MS analysis was performed using standard settings: 60k resolution for full MS, 3e6 AGC target, 375 – 1600 m/z scan range. Top 12 MS/MS were performed relying on a 1e5 Intensity threshold, 1.4 m/z isolation window and NCE = 27 with a 30k resolution. Unknown, singly and excessively (5-8) charged species were excluded. Dynamic exclusion of 16 seconds was used.

Results and discussion

To set up a method to identify proteins interacting with Pin1 and distinguish proteins acting on Pin1 (i.e. Pin1 being their substrate) from protein substrates of Pin1, we performed pull-down experiments with Pin1 or negatively regulated Pin1 protein coupled to nickel beads via a histidine tag, followed by MS-based protein identification. We compared Pin1 with serine-16 phosphorylated Pin1, carrying an inhibitory phosphorylation which prevents substrates from binding to the Pin1 WW domain. Unloaded beads were included to control for nonspecific binding to the nickel beads. An interesting aspect of this approach is that the bait protein can be characterized after the pull-down, so in theory it would be possible to compare the effect of different PTMs on the bait protein and its interactors. See Figure 2 for an illustration of the method and Figure 3 for a schematic overview of the pull-down experiment.

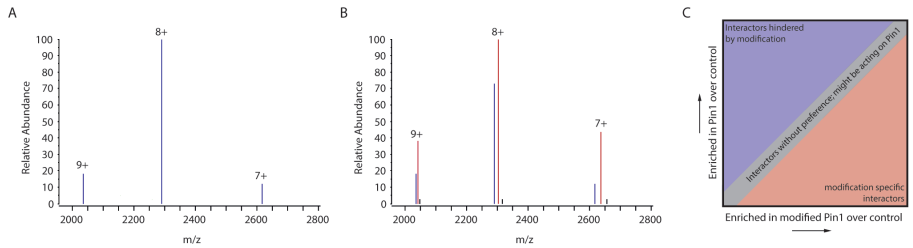


Figure 2 - Principle of pull-down method. A) Illustration of Pin1 native MS spectrum after elution from beads. B) Illustration of singly phosphorylated Pin1 native MS spectrum. C) Expected distribution of various interactors. Proteins may prefer interaction with unmodified Pin1 (blue area) or they may prefer interaction with the singly phosphorylated proteoform (red area). Finally, proteins may not have a preference, indicating that they are likely interacting with a different area of Pin1 or perhaps acting on Pin1 instead of being a substrate of Pin1.

Proteins in the input lysate and the main elution steps were separated using SDS-PAGE and Coomassie stained for a visual comparison of the different conditions. Despite multiple washing steps, unfolding of proteins on beads containing Pin1 shows the release of a significant portion of Pin1 protein (Figure 4). As the histidine tag interaction with nickel should remain intact under denaturing conditions, this might indicate that Pin1 has affinity for binding other Pin1 proteins. If that is the case, many of the available substrate binding sites may have been occupied due to the high concentration of Pin1 in the E. Coli lysate.

Dissociation of the bound Pin1 fraction using imidazole not only released a large quantity of bound Pin1 but also showed residual binding of proteins, including NONO (Q15233) and Transcriptional repressor protein YY1 (P25490), to the beads. No clear differences between Pin1-bound beads and phosphorylated Pin1-bound beads could be observed, however, indicating that these other proteins might lead to some binding events unspecific to Pin1 but they are unlikely to lead to differences between these two conditions. Note

that the background of proteins binding to empty beads is high and similar with or without E. Coli background incubation (i.e. likely stems from the HEK293T incubation step). Since the mass spectrometer has both speed and resolving power, the presence of such a high background was not expected to seriously hinder the identification of interactors, however, it might reduce our sensitivity. Therefore, further steps to reduce nonspecific interactions might be worth evaluating.

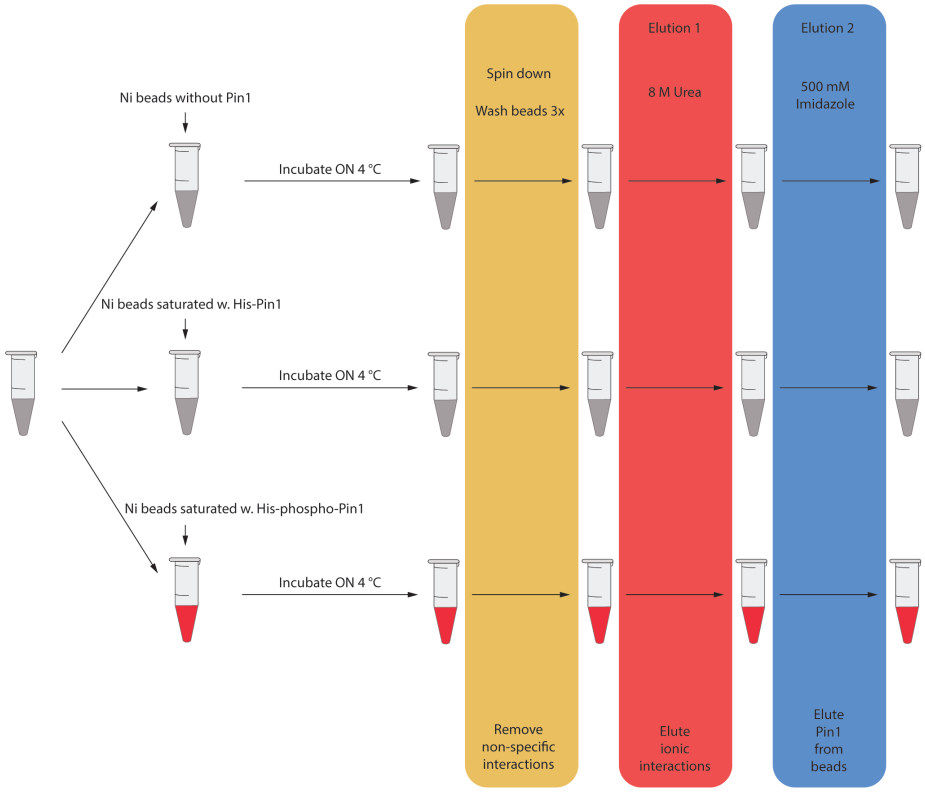


Figure 3 - Schematic overview of the Pin1 pulldown experiment. A lysate was prepared from HEK293T cells and three types of beads, namely unloaded, Pin1 or phospho-Pin1, were incubated with lysate ON while rotating head over head at 4 degrees Celsius. Specific and nonspecific interactors where eluted using a series of washing steps, see materials and methods for more information.

After MS analysis, the MS spectra were searched using MaxQuant[45], and the correlation between the different samples was assessed. Triplicates of the same condition correlated well, indicating that the used method is quite repeatable (Figure 5). Much lower correlation was found between empty beads and Pin1 loaded beads, which is expected as the empty beads likely carry mainly proteins not specifically interacting with Pin1 or its phosphorylated counterpart. Subsequently, the identified proteins and their MS2 spectral counts were analyzed using the CRAPome database[46]. In this database, data of many different affinity-based mass spectrometric experiments is collected,

allowing users to interrogate their data more strictly by incorporating negative controls from other studies as well as their own, thereby reducing the potential of false positive identifications. Since nickel-beads were not incorporated in the CRAPome database, we incorporated several studies with pulldowns in HEK293T cells and sepharose based beads, to include nonspecific binders to the bead material itself in the appropriate background.

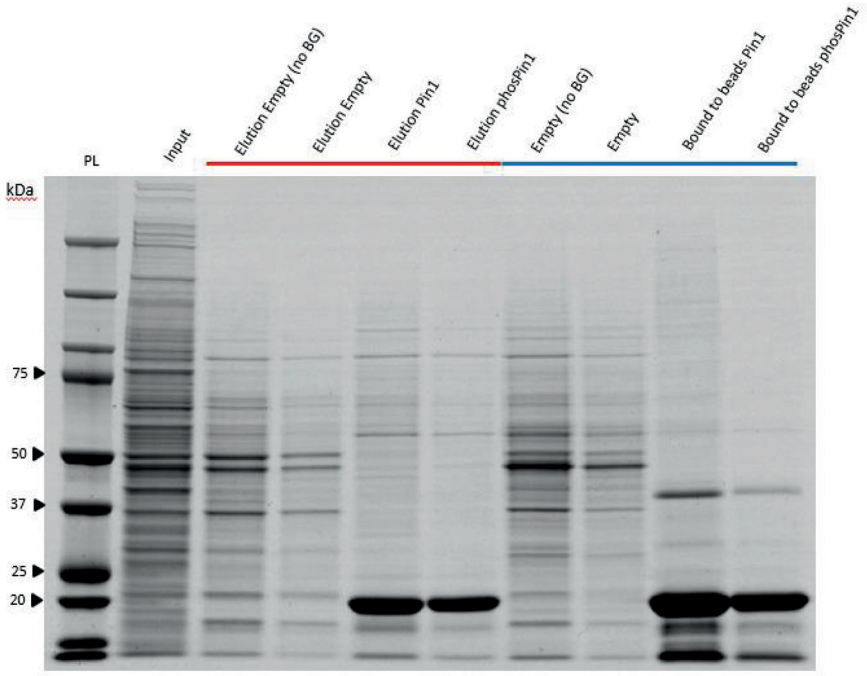


Figure 4 - Analysis of elution steps. Input lysate of HEK293T cells as well as aliquots from different bead conditions following elution 1 (red) and elution 2 (blue) were compared. PL indicates protein ladder, no BG refers to empty beads not pre-incubated in E. Coli lysate prior to incubation in HEK293T lysate.

Analysis using the CRAPome database yields three different metrics, namely a fold change based on an arithmetic mean (indicated as fold change A, FC_A), a geometric mean (indicated as fold change B, FC_B) and a SAINT score (the probability of an interaction being a true, specific interactor). Since FC_B is more stringent and also includes negative controls from other studies, we decided to proceed with a cutoff of at least 2 in FC_B over empty beads. Using this cutoff, all remaining proteins have SAINT scores of over 0.5, with the majority scoring over 0.8, indicating that these are probable true interactors (Supplemental Figure 1).

An overview of the CRAPome analysis for Pin1 over empty beads and serine-16 phosphorylated Pin1 over empty beads based on FC_B revealed a large number

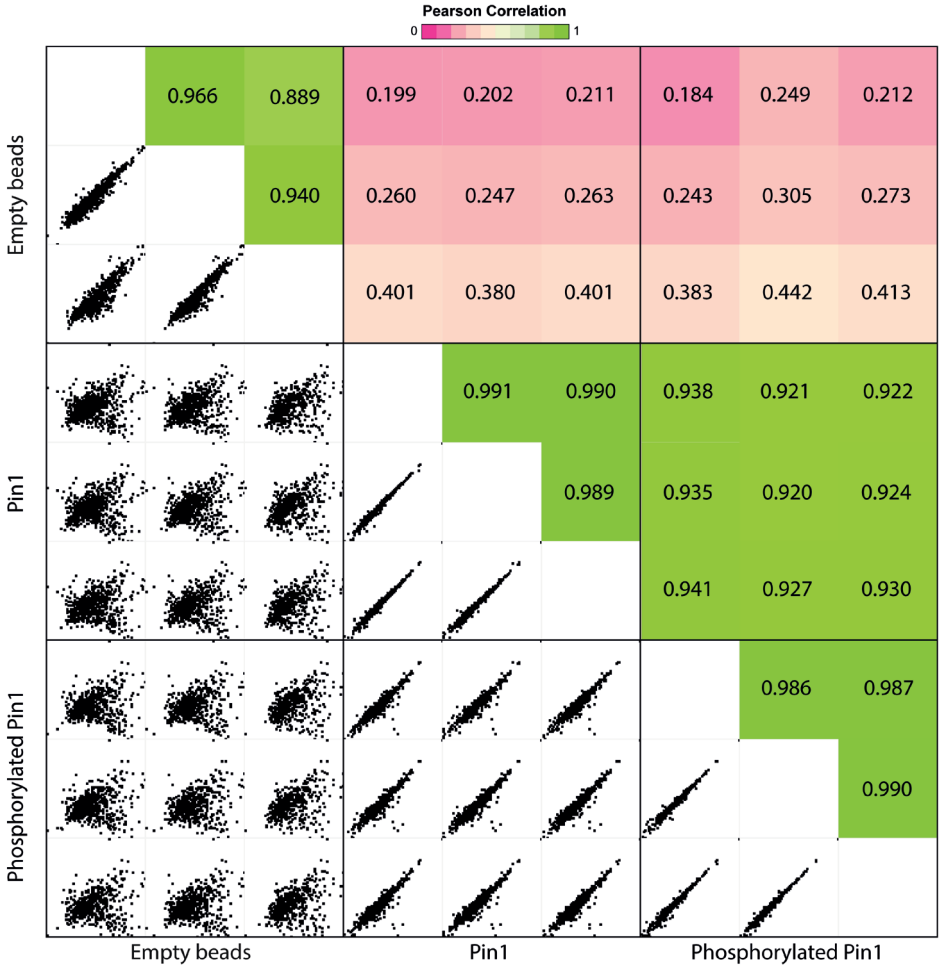


Figure 5 - Assessment of the correlation between quantifications of all identified proteins in the LC-MS runs following the distinctive pull-down. For each replicate, the Pierson correlation was calculated based on LFQ intensity between other replicates and conditions. Correlation within replicates is highest and reduces slightly when comparing Pin1 with phosphorylated Pin1 eluted proteins. As expected, the correlation between empty beads and Pin1 or phosphorylated Pin1 is lowest.

of previously reported interactors sitting below the $FC_B \geq 2$ cutoff value (Figure 6, shown in red). This could indicate that these proteins are simply not true Pin1 interactors under these conditions; they might require different modifications, cellular context or a different conformation to be able to bind Pin1. At least 45 of the known Pin1 interactors were identified with a pull-down experiment using only the WW-domain[47], so the presence of the PPIase domain on the linker could influence WW-domain conformation and hence binding at the WW-domain in the full-length Pin1 protein, resulting in a lower binding affinity. Interaction between the two domains has been observed in previous studies[35,36].

Naturally, different or extended washing steps might reduce background binding to beads leading to a higher sensitivity and thereby increase the number of proteins identified as ≥ 2-fold change. Alternatively, the limit used here to identify true interactors might be too stringent. The inclusion of negative controls from other studies resulted in the exclusion of 364 proteins, including 28 previously reported interactors, which did reveal a $FC_A \geq 2$ over the negative controls included in this work.

Next, we focused on proteins that are enriched using beads with Pin1 (n=44), phosphorylated Pin1 (n=26) or in both (n=125), showing a FC_B of at least 2 over beads not loaded with Pin1. For these selected proteins, the spectral counts were normalized to the amount of Pin1 bound to the nickel beads (identified from the corresponding imidazole elution; elution 2) and the expected number of peptides per protein, to ensure that the normalized spectral counts are not biased towards larger proteins. The top interactors for Pin1-bound and phosphorylated Pin1-bound beads are listed in Table 1 and 2, respectively.

Most of the interactors in Table 1, including SRSF2, show a higher normalized spectral counts score for Pin1 than phosphorylated Pin1. Since phosphorylation of Pin1 occurred mainly at inhibitory phosphorylation sites serine 16/18/19 (30%) and serine 71 (70%), on the WW and PPIase domains, that would indicate that these proteins are likely substrates of Pin1. In addition, most have multiple SP/TP sites, which would fit well with several of the hypothesized mechanisms of action for Pin1.

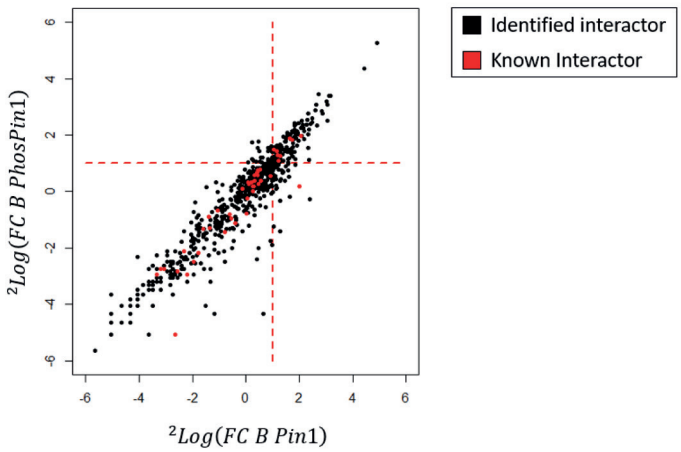


Figure 6 - Comparison of fold change B observed in Pin1 over phosphorylated Pin1 beads. Each dot represents a unique protein, red dots represent Pin1 interactors already listed in the BioGRID database. Dotted lines indicate a FCB equal to 2. The WW-domain binding substrates of Pin1 are expected to have a stronger interaction with the unphosphorylated Pin1 beads compared to the phosphorylated Pin1 beads. In total, 1978 proteins were identified and included in this figure (76 known interactors in red, 1902 other proteins shown in black).

For the phosphorylated Pin1 specific proteins in Table 2 the distinction is less clear, most proteins do not show a clear preference for phosphorylated over unphosphorylated Pin1. This could reflect the phosphorylation degree of phosphorylated Pin1, which should be verified by the intact protein analysis of phosphorylated Pin1 eluted from the beads. Alternatively, it could also indicate that these proteins interact with other parts of the Pin1 protein then its WW-domain and regulate Pin1 rather than being Pin1 substrates. However, to validate whether this type of pull-down can also be used to enrich proteins acting on Pin1, it would be very insightful to spike some known Pin1 modulators into the proteomes and evaluate their enrichment.

In general, many of the enriched proteins hypothesized to bind Pin1 in this study seem to be interacting with RNA. For example, we identified splicing factors, RNA binding proteins, helicases and RNA-related transport proteins. It would be interesting to validate if these proteins are all actual Pin1 substrates, as that might imply that these RNA-related roles are under more tight control than other biological processes by relying on the combination of phosphorylation and conformation change in which Pin1 is a critical player.

Conclusion

In summary, we have developed a pull-down method based on affinity chromatography which enables the investigation of Pin1 interactors. Due to the reversible bead binding, it can be expanded towards an intact protein analysis approach to characterize the modifications present in detail and quantify the amount of modified protein on the beads. Therefore, the method might be suitable for explorations of PTM-dependent interactions on His-tagged proteins.

Table 1 - Top Pin1 interactors. Spectral counts (MS/MS spectra) assigned as peptides originating from an interactor, normalized to Pin1 load on the beads and protein length. Fold change B of Pin1 over empty beads. Entries given in red represent previously reported interactors. Phosphorylation sites listed as reported in phosphosite database.

Interactor	Spectral counts Pin1	Spectral counts PhosPin1	Fold change B over empty	SAINT score	pS/TP Sites	Protein name
SRSF2	83.1	0	2.54	1.00	$pS^{26}P$, $pS^{129}P$, $pS^{191}P$, $pS^{208}P$, $pS^{212}P$	Serine/arginine-rich splicing factor 2
ESS2	13.7	3.1	2.76	0.98	$pT^{3}P$, $pT^{104}P$, $pT^{110}P$, $pS^{292}P$, $pT^{333}P$, $pT^{339}P$, $pT^{386}P$, $pS^{391}P$, $pS^{395}P$, $pT^{419}P$, $pS^{421}P$, $pT^{430}P$, $pT^{437}P$, $pT^{450}P$	Splicing factor ESS-2 homolog
BCKDK	12.9	9.4	3.98	1.00	No sites in phosphosite db, TP/SP in sequence	[3-methyl-2-oxobutanoate dehydrogenase [lipoamide]] kinase, mitochondrial
PCYT1A	12.2	0	4.52	1.00	$pS^{315}P$, $pS^{319}P$, $pS^{323}P$, $pS^{329}P$, $pS^{331}P$, $pS^{343}P$, $pS^{347}P$	Choline-phosphate cytidylyltransferase A
RAE1	10.0	4.0	2.65	0.98	$pS^{209}P$	mRNA export factor
DDX6	9.3	7.0	2.45	0.99	No sites in phosphosite db, SP/TP in sequence	Probable ATP-dependent RNA helicase DDX6
RBM12	7.9	0.5	5.29	1.00	$pS^{352}P$, pS^{375P}, $pS^{424}P$	RNA binding protein 12
C2orf49	7.0	6.3	2.04	0.96	$pS^{150}P$, $pS^{182}P$, $pS^{189}P$, $pS^{193}P$, $pT^{198}P$	Ashwin
RBM26	6.9	1.9	2.43	0.97	$pS^{127}P$, $pS^{518}P$, $pS^{616}P$, $pS^{795}P$	RNA binding protein 26
NKAP	6.9	1.2	2.47	0.98	$pS^{9}P$, $pS^{149}P$, $pT^{161}P$	NF-kappa-B activating protein

Table 2 - Top interactors binding to phosphorylated Pin1. Spectral counts (MS/MS spectra) assigned as peptides originating from an interactor, normalized to Pin1 load on the beads and protein length. Fold change B for proteins calculated between phosphorylated Pin1 over empty beads. Red indicates a previously reported interactor. Phosphorylation sites listed as reported in phosphosite database.

Interactor	Spectral counts Pin1	Spectral counts PhosPin1	Fold change over empty	SAINT score	pS/TP Sites	Protein name
NONO	210.5	156.8	2.27	0.67	$pT^{15}P$, $pT^{428}P$, $pT^{450}P$	Non-POU domain-containing octamer-binding protein
C20orf27	126.6	148.0	8.97	1.00	$pT^{147}P$ (found in mice)	UPF0687 protein C20orf27
NUDT21	71.9	64.2	2.30	0.98	No sites in phosphosite/sequence	Cleavage and polyadenylation specificity factor subunit 5
SERPINH1	59.5	53.5	2.33	0.98	$pS^{37}P$, $pS^{70}P$	Serpin H1
SNW1	54.8	55.3	2.75	0.99	$pT^{180}P$, $pS^{224}P$, $pS^{232}P$	SNW domain-containing protein 1
SF1	49.8	47.9	3.16	0.99	$pT^{8}P$, $pS^{80}P$, $pS^{82}P$, $pT^{232}P$, $pT^{338}P$	Splicing factor 1
MIF	49.7	33.8	2.90	0.99	$pS^{91}P$	Macrophage migration inhibitory factor
DDX42	43.9	44.0	30.17	1.00	$pS^{96}P$, $pT^{228}P$, $T^{603}P$, $pS^{751}P$, $pS^{754}P$, $pS^{831}P$	ATP-dependent RNA helicase DDX42
KHDRBS1	42.4	21.9	2.47	0.97	$pT^{33}P$, $pS^{58}P$, $pT^{84}P$, $pT^{317}P$	KH domain-containing, RNA-binding, signal transduction-associated protein 1
SAMHD1	41.1	27.9	21.74	1.00	$pS^{18}P$, $pT^{21}P$, $pS^{33}P$, $pT^{138}P$, $pS^{278}P$, $pT^{592}P$	Deoxynucleoside triphosphate triphosphohydrolase SAMHD1
CTNND1	40.0	23.1	5.02	1.00	$pS^{47}P$, $pT^{310}P$, $pS^{651}P$, $pT^{906}P$	Catenin delta-1
CPSF7	39.8	46.4	3.95	1.00	$pT^{203}P$, $pS^{427}P$	Cleavage and polyadenylation specificity factor subunit 7
BAIAP2L1	39.1	41.8	8.13	1.00	$pT^{248}P$, $pT^{252}P$, $pT^{257}P$, $pS^{281}P$, $pS^{295}P$, $pT^{412}P$, $pS^{414}P$, $pT^{416}P$	Brain-specific angiogenesis inhibitor 1-associated protein 2-like protein 1

References

1 Kim PS & Baldwin RL (1982) Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding. Annu Rev Biochem 51, 459–489.

2 Wedemeyer WJ, Welker E & Scheraga HA (2002) Proline Cis–Trans Isomerization and Protein Folding †. Biochemistry 41, 14637–14644.

3 Grathwohl C & Wüthrich K (1981) Nmr studies of the rates of proline cis - trans isomerization in oligopeptides. Biopolymers 20, 2623–2633.

4 Fischer G, Tradler T & Zarnt T (1998) The mode of action of peptidyl prolyl cis/trans isomerases in vivo: binding vs. catalysis. FEBS Lett 426, 17–20.

5 Maruyama T & Furutani M (2000) Archaeal Peptidyl Prolyl Cis-Trans Isomerases (PPIases). Front Biosci 5, 821–236.

6 Hanes SD (2014) Prolyl isomerases in gene transcription. Biochim Biophys Acta 1850, 2017–2034.

7 Gothel SF & Marahiel M a (1999) Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts. Cell Mol Life Sci 55, 423–436.

8 He Z, Li L & Luan S (2004) Immunophilins and parvulins. Superfamily of peptidyl prolyl isomerases in Arabidopsis. Plant Physiol 134, 1248–1267.

9 Ubersax JA & Ferrel Jr JE (2007) Mechanisms of Specificity in Protein Phosphorylation. Nat Rev 8, 530–541.

10 Pinna LA & Ruzzene M (1996) How do protein kinases recognize their substrates? Biochim Biophys Acta 1314, 191–225.

11 Zhu G, Fujii K, Belkina N, Liu Y, James M, Herrero J & Shaw S (2005) Exceptional Disfavor for Proline at the P+1 Position among AGC and CAMK kinases Establishes Reciprocal Specificity between Them and the Proline-directed Kinases. J Biol Chem 280, 10743–10748.

12 Werner-Allen JW, Lee C-J, Liu P, Nicely NI, Wang S, Greenleaf AL & Zhou P (2011) cis -Proline-mediated Ser(P) 5 Dephosphorylation by the RNA Polymerase II C-terminal Domain Phosphatase Ssu72. J Biol Chem 286, 5717–5726.

13 Schutkowski M, Bernhardt A, Zhou XZ, Shen M, Reimer U, Rahfeld JU, Lu KP & Fischer G (1998) Role of phosphorylation in determining the backbone dynamics of the serine/ threonine-proline motif and Pin1 substrate recognition. Biochemistry 37, 5566–75.

14 Weiwad M, Küllertz G, Schutkowski M & Fischer G (2000) Evidence that the substrate backbone conformation is critical to phosphorylation by p42 MAP kinase. FEBS Lett 478, 39–42.

15 Lu KP, Finn G, Lee TH & Nicholson LK (2007) Prolyl cis-trans isomerization as a molecular timer. Nat Chem Biol 3, 619–629.

16 Yaffe MB, Schutkowski M, Shen M, Zhou XZ, Stukenberg PT, Rahfeld JU, Xu J, Kuang J, Kirschner MW, Fischer G, Cantley LC & Lu KP (1997) Sequence-Specific and Phosphorylation-Dependent Proline Isomerization: A Potential Mitotic Regulatory Mechanism. Science (80-) 278, 1957–1960.

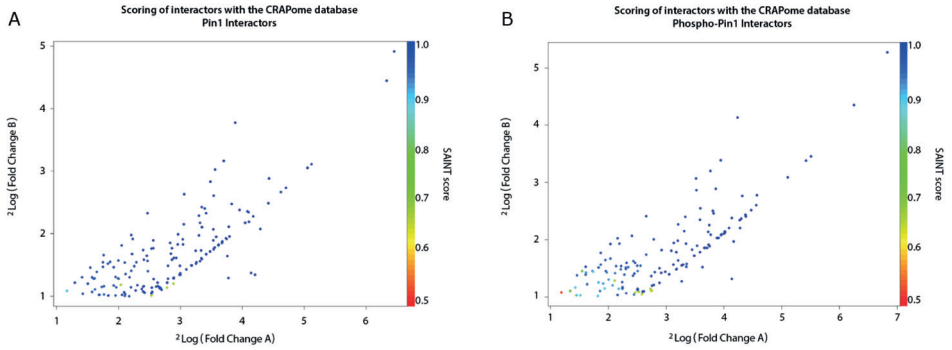
17 Stukenberg PT & Kirschner MW (2001) Pin1 Acts Catalytically to Promote a Conformational Change in Cdc25. Mol Cell 7, 1071–1083.

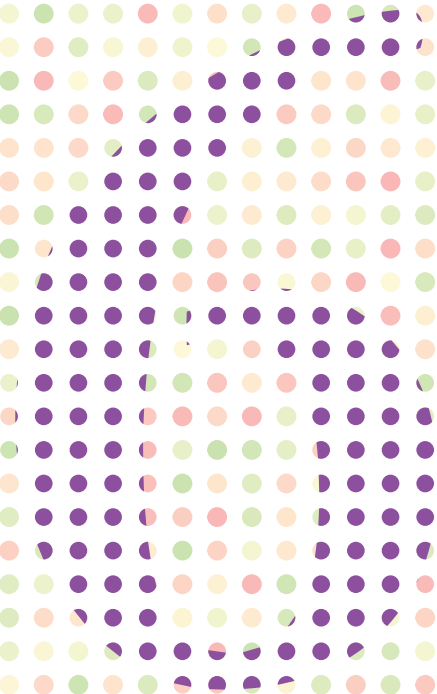
18 Zhou W, Yang Q, Low CB, Karthik BC, Wang Y, Ryo A, Yao SQ, Yang D & Liou YC (2009) Pin1 catalyzes conformational changes of Thr-187 in p27Kip1 and mediates its stability

through a polyubiquitination process. J Biol Chem 284, 23980–23988.
19 Hanes SD, Shank PR & Bostian KA (1989) SEQUENCE AND MUTATIONAL ANALYSIS OF ESS1, A GENE ESSENTIAL FOR GROWTH IN SACCHAROMYCES-CEREVISIAE. YEAST 5, 55–72.
20 Lu KP, Hanes SD & Hunter T (1996) A human peptidyl-prolyl isomerase essential for regulation of mitosis. Nature 380, 544–547.
21 Matena A, Rehic E, Hönig D, Kamba B & Bayer P (2018) Structure and function of the human parvulins Pin1 and Par14/17. Biol Chem 399, 101–125.
22 Galat A & Riviere S (1998) Peptidyl-prolyl cis/trans isomerases: protein profile Oxford University Press, Oxford.
23 Ranganathan R, Lu KP, Hunter T & Noel JP (1997) Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. Cell 89, 875–886.
24 Eckerdt F, Yuan J, Saxena K, Martin B, Kappel S, Lindenau C, Kramer A, Naumann S, Daum S, Fischer G, Dikic I, Kaufmann M & Strebhardt K (2005) Polo-like kinase 1-mediated phosphorylation stabilizes Pin1 by inhibiting its ubiquitination in human cells. J Biol Chem 280, 36575–36583.
25 Wulf GM, Ryo A, Wulf GG, Lee SW, Niu T, Petkova V & Lu KP (2001) Pin1 is overexpressed in breast cancer and cooperates with Ras signaling in increasing the transcriptional activity of c-Jun towards cyclin D1. EMBO J 20, 3459–3472.
26 Rustighi A, Zannini A, Tiberi L, Sommaggio R, Piazza S, Sorrentino G, Nuzzo S, Tuscano A, Eterno V & Benvenuti F (2014) Prolyl-isomerase Pin1 controls normal and cancer stem cells of the breast. EMBO Mol Med 6, 99–119.
27 Lu P-J, Wulf G, Zhou XZ, Davies P & Lu KP (1999) The prolyl isomerase Pin1 restores the function of Alzheimer-associated phosphorylated Tau protein. Nature 399, 784–788.
28 Lu Z & Hunter T (2014) Prolyl isomerase Pin1 in cancer. Cell Res 24, 1033–1049.
29 Driver JA, Zhou XZ & Lu KP (2014) Regulation of protein conformation by Pin1 offers novel disease mechanisms and therapeutic approaches in Alzheimer's disease. Discov Med 17, 93–9.
30 Bailey M, Shilton B, Brandl C & Litchfield D (2008) The Dual Histidine Motif in the Active Site of Pin1 Has a Structural Rather than Catalytic Role. Biochemistry 47, 11481–11489.
31 Wang J-Z, Xi L, Lin T, Wang Y, Zhu G-F & Du L-F (2015) The structural and functional role of the three tryptophan residues in Pin1. J Photochem Photobiol B Biol 146, 58–67.
32 Born A, Henen MA & Vögeli B (2019) Activity and Affinity of Pin1 Variants. Molecules 25, 36.
33 Ikura T, Kinoshita K & Ito N (2008) A cavity with an appropriate size is the basis of the PPIase activity. Protein Eng Des Sel 21, 83–89.
34 Tork Ladani S, Souffrant MG, Barman A & Hamelberg D (2015) Computational perspective and evaluation of plausible catalytic mechanisms of peptidyl-prolyl cis–trans isomerases. Biochim Biophys Acta - Gen Subj 1850, 1994–2004.
35 Namanja AT, Peng T, Zintsmaster JS, Elson AC, Shakour MG & Peng JW (2007) Substrate recognition reduces side-chain flexibility for conserved hydrophobic residues in human Pin1. Structure 15, 313–327.

36 Guo J, Pang X & Zhou H-X (2015) Two Pathways Mediate Interdomain Allosteric Regulation in Pin1. Structure 23, 237–247.
37 Wang J, Tochio N, Kawasaki R, Tamari Y, Xu N, Uewaki J, Utsunomiya-Tate N & Tate S (2015) Allosteric Breakage of the Hydrogen Bond within the Dual-Histidine Motif in the Active Site of Human Pin1 PPIase. Biochemistry 54, 5242–5253.
38 Lee YM & Liou Y-C (2018) Gears-In-Motion: The Interplay of WW and PPIase Domains in Pin1. Front Oncol 8.
39 Lu KP, Liou Y-C & Zhou XZ (2002) Pinning down proline-directed phosphorylation signaling. Trends Cell Biol 12, 164–172.
40 Zhou XZ, Lu P-J, Wulf G & Lu KP (1999) Phosphorylation-dependent prolyl isomerization: a novel signaling regulatory mechanism. Cell Mol Life Sci 56, 788–806.
41 Jacobs DM, Saxena K, Vogtherr M, Bernardo P, Pons M & Fiebig KM (2003) Peptide binding induces large-scale changes in inter-domain mobility in human Pin1. J Biol Chem 278, 26174–26182.
42 Verdecia MA, Bowman ME, Lu KP, Hunter T & Noel JP (2000) Structural basis for phosphoserine-proline recognition by group IV WW domains. Nat Struct Biol 7, 639–43.
43 Wintjens R, Wieruszeski J-M, Drobecq H, Rousselot-Pailley P, Buée L, Lippens G & Landrieu I (2001) 1H NMR Study on the Binding of Pin1 Trp-Trp Domain with Phosphothreonine Peptides. J Biol Chem 276, 25150–25156.
44 Namanja AT, Wang XJ, Xu B, Mercedes-Camacho AY, Wilson KA, Etzkorn FA & Peng JW (2011) Stereospecific gating of functional motions in Pin1. Proc Natl Acad Sci 108, 12289–12294.
45 Cox J & Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26, 1367–1372.
46 Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis NA, Li T, Miteva Y V, Hauri S, Sardiu ME, Low TY, Halim VA, Bagshaw RD, Hubner NC, Al-Hakim A, Bouchard A, Faubert D, Fermin D, Dunham WH, Goudreault M, Lin Z-Y, Badillo BG, Pawson T, Durocher D, Coulombe B, Aebersold R, Superti-Furga G, Colinge J, Heck AJR, Choi H, Gstaiger M, Mohammed S, Cristea IM, Bennett KL, Washburn MP, Raught B, Ewing RM, Gingras A-C & Nesvizhskii AI (2013) The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. Nat Methods 10, 730–736.
47 Ingham RJ, Colwill K, Howard C, Dettwiler S, Lim CSH, Yu J, Hersi K, Raaijmakers J, Gish G, Mbamalu G, Taylor L, Yeung B, Vassilovski G, Amin M, Chen F, Matskova L, Winberg G, Ernberg I, Linding R, O'Donnell P, Starostine A, Keller W, Metalnikov P, Stark C & Pawson T (2005) WW Domains Provide a Platform for the Assembly of Multiprotein Networks. Mol Cell Biol 25, 7092–7106.

Supplemental Figure 1 – Comparison of fold change A, B and SAINT score obtained after CRAPome analysis. A) Analysis of interactors with Pin1 loaded beads over empty beads. B) Analysis of interactors with phosphorylated Pin1 loaded beads over empty beads. Saint scores were shown using a colour gradient.





Chapter 6

Closing opinion
Nederlandse samenvatting

A final short personal perspective about some caveats of mainstream proteomics

Proteomics is quite literally the study of all proteins collectively[1]. Of course, we are never able to see all proteins collectively. Therefore, every Pro in Proteomics should be aware of the choices they make in their workflow, and how this might possibly affect the outcome. It is crucial that we understand that proteins are complex, folded polymers of a mixture of many amino acids, each with their own chemical properties. Proteins can be large, they can be small, they can have many sulfur bridges or none at all, they can have a defined structure, or not, or partly. They can carry no post-translational modifications, or many of them and everything in between. They can have a clear fold, or not, or a multitude of different conformations. They can be soluble in a certain buffer, but likely not in others. When they are modified, they can be activated by a modification, or inhibited, or their specificity can be changed by the modification, or their cellular localization. Multiple modifications can also affect each other, by promoting or preventing the placements of others or by altering the final effect of the modifications on the protein. To briefly summarize, proteins are complex. As scientists that work with them, it is crucial that we keep an open mind.

When studying these proteins, we often like to get an overview of as many proteins as we can get some sort of evidence for, using 'shotgun' proteomics[2]. In a traditional 'shotgun' proteomics experiment, a proteome is typically reduced and alkylated to release sulfur bridges and then fully digested by trypsin. In the next step, most of the salts, nucleotides and other small molecules present in the cells or tissue under investigation are removed by solid phase extraction (SPE), generating a complex mixture of peptides ready for separation and analysis by LC-MS/MS[3]. At this stage, we have already made multiple choices that will strongly influence the sample that we load onto the LC-MS system.

Between the buffer we choose to lyse our cells - including the possible presence of detergents, the ionic strength and the pH - and the lysis technique we employ, we will end up with a subset of all the proteins that were expressed in our sample. Simply put, we can only continue with proteins that are free and soluble in the buffer during the lysis steps. Next, we reduce and alkylate. At the risk of coming across as a pessimist, this can go wrong in many, many ways. You simply cannot alkylate what you have not properly reduced, and you can easily alkylate amino acids other than cysteines[4]. Both will have similar effects on your overall protein IDs: you will end up with less. Unfortunately, in the realm of proteomics less ≠ more, it is just disappointing.

The next step in the typical workflow is to digest the alkylated proteins into peptides. To do so, most proteomics researchers have a largely monogamous relationship with trypsin. We like trypsin, because it is fairly specific (see Chapter 3) and it produces peptides that have a positive charge at either end,

which is very helpful during peptide fragmentation in the mass spectrometer[5]. In addition, the post-digestion proteomics workflow is optimized for trypsin, so it tends to give us the 'best' results. That is, if you value the number of proteins you can identify.

Like any boyfriend, trypsin is not perfect. It doesn't cleave absolutely specifically, and its ability to cleave is strongly affected by the presence of prolines and PTMs near its cleavage site[6]. To make matters worse, arginine and lysine are both quite abundant, leading to a lot of rather short peptides[7]. The theoretical average peptide length for the trypsinated human proteome is nine amino acids. Shorter peptides can be a problem, because the number of human proteins containing a peptide with a certain amino acid combination increases with decreasing peptide size, see Table 1. Therefore, all peptides with a length shorter than seven amino acids form a big risk of wasted resources. Since these short peptides are likely present in multiple proteins, most data-processing workflows don't even consider them.

Following digestion, SPE is a common way to get rid of residual protease, salts and small molecules present in the buffer and in the cells or tissues that were digested. Obviously, all the peptides we produced have their own hydrophobicity and will have their own retention on the SPE material chosen. We must consider the loading of the column and the mobile phases we decide on. In practice, we like to run laboratory-standard methods on sample prep and on the mass spectrometers, because those enable us to compare data from different studies and we often use conventional search settings to avoid discussions and re-inventing the wheel time after time. Over the years, the search engines matching our MS2 spectra to peptides have learned to expect and score based on tryptic behavior, maximizing the number of peptide IDs we can achieve.

Table 1 – In silico polypeptide lengths and their unique combinations. Evaluation of the amino acid combinations of various lengths present in the human proteome (20386 proteins, Swissprot, July 2021). Total number of combinations compared to those ending in R/K (i.e. tryptic combinations).

Peptide length	Number of combinations	% unique combinations	Number of tryptic combinations	% unique tryptic combinations
8	11220299	87.1	1277278	86.9
7	11240677	81.9	1279781	81.9
6	11261055	54.0	1282360	53.7
5	11281433	5.6	1284998	4.6
4	11301813	0 (758)	1287625	0 (29)

Basically, we have fully optimized our entire workflow to accommodate trypsin, which has enabled us to maximize IDs in the past, but which might not be the best way forward. Trypsin will obviously remain a very powerful tool in our proteomic toolbox, but for certain applications we will probably gain more

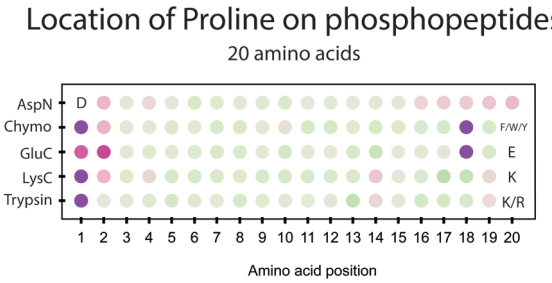


Figure 1 - Relative occurrence of proline on peptides generated with various proteases. Data from publication by Giansanti et al [12].

traction by broadening our scope and optimizing for other proteases[8–11]. Of course, researchers have made steps in this area, but most studies are met with comments regarding poor performance of 'unconventional' proteases as compared to trypsin. In our EndoPro work, however, we observed similar MS2 events for trypsin and EndoPro (Chapter 3, table 2). This is indicative of similar numbers of peptide-like species being generated by the proteases. We simply annotate less of these MS2 spectra, because spectral quality isn't ideal (i.e. proline effect in case of EndoPro) or perhaps because our fragmentation or scoring methods are not optimized for non-tryptic behavior.

When studying the interplay between different PTMs, the generation of longer peptides is beneficial as these longer peptides give more information on co-occurring PTMs without the need for extensive purification of individual proteins as is required for native MS experiments[13]. Therefore, proteases specific for less abundant amino acids might be helpful, or those recognizing motifs of two or three specific amino acids. An interesting course of action might be to attempt to modify trypsin by site directed mutagenesis and expand its specificity from R/K towards a multi-amino acid motif. This way, we can keep our 'ideal' peptide characteristic of N- and C-terminal positive charges, while increasing the average peptide length. However, care must be taken not to limit the proteases activity[14]. For example, engineering a docking site for specific amino acids a bit further from the active site might allow us to generate a (D/E)XXX(R/K) type of specificity while keeping the tryptic catalytic site intact.

Since most proteases have distinct cleavage specificities, the use of different proteases results in different visible subsets of the proteome[9,15]. This is exemplified by the suppression of R/K in tryptic digests, which is not observed in peptides generated with EndoPro (Chapter 3, Figure 4B) but more generally by the different compositions of phosphopeptides generated by different proteases[12,15]. For example, tryptic peptides show phosphosite placement that is distinct from EndoPro peptides (Chapter 3, Figure 6) and if we look at a previously published comparison between five different proteases, we can observe clear differences in the toleration of proline near cleavage sites of these

proteases (Figure 1).

To summarize, tryptic peptides and phosphosites have been extensively studied[8]. For those of us interested in areas of the proteome that are hard to see using trypsin, other proteases might be a good solution. More generally, I expect the field of proteomics to gain more new insights by broadening its scopes and investing time and resources in the optimization of different workflows. We cannot expect these other workflows to instantly outperform our golden standard, but they should certainly enable us to see beyond the tryptic proteome we have been studying to date.

References

1 Aslam B, Basit M, Nisar MA, Khurshid M & Rasool MH (2017) Proteomics: Technologies and Their Applications. Journal of Chromatographic Science 55, 182–196.

2 Zhang Y, Fonslow BR, Shan B, Baek M-C & Yates JR (2013) Protein analysis by shotgun/bottom-up proteomics. Chemical reviews 113, 2343–94.

3 Wither MJ, Hansen KC & Reisz JA (2016) Mass Spectrometry-Based Bottom-Up Proteomics: Sample Preparation, LC-MS/MS Analysis, and Database Query Strategies. Current Protocols in Protein Science 86.

4 Müller T & Winter D (2017) Systematic Evaluation of Protein Reduction and Alkylation Reveals Massive Unspecific Side Effects by Iodine-containing Reagents. Molecular & Cellular Proteomics 16, 1173–1187.

5 Saveliev S, Bratz M, Zubarev R, Szapacs M, Budamgunta H & Urh M (2013) Trypsin/Lys-C protease mix for enhanced protein mass spectrometry analysis. Nature Methods 10, i–ii.

6 Bubis JA, Gorshkov V, Gorshkov M V. & Kjeldsen F (2020) PhosphoShield: Improving Trypsin Digestion of Phosphoproteins by Shielding the Negatively Charged Phosphate Moiety. Journal of the American Society for Mass Spectrometry 31, 2053–2060.

7 Hildonen S, Halvorsen TG & Reubsaet L (2014) Why less is more when generating tryptic peptides in bottom-up proteomics. PROTEOMICS 14, 2031–2041.

8 Tsiatsiani L & Heck AJR (2015) Proteomics beyond trypsin. FEBS Journal 282, 2612–2626.

9 Giansanti P, Tsiatsiani L, Low TY & Heck AJR (2016) Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. Nature Protocols 11, 993–1006.

10 Meyer JG (2014) In Silico Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage. ISRN Computational Biology 2014, 1–7.

11 Meyer B, Papasotiriou DG & Karas M (2011) 100% protein sequence coverage: a modern form of surrealism in proteomics. Amino Acids 41, 291–310.

12 Giansanti P, Aye TT, van den Toorn H, Peng M, van Breukelen B & Heck AJR (2015) An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas. Cell Reports 11, 1834–1843.

13 Cristobal A, Marino F, Post H, van den Toorn HWP, Mohammed S & Heck AJR (2017) Toward an Optimized Workflow for Middle-Down Proteomics. Analytical Chemistry 89, 3318–3325.

14 Craik C, Largman C, Fletcher T, Roczniak S, Barr P, Fletterick R & Rutter W (1985) Redesigning trypsin: alteration of substrate specificity. Science 228, 291–297.

15 van der Laarse SAM, van Gelder CAGH, Bern M, Akeroyd M, Olsthoorn MMA & Heck AJR (2020) Targeting proline in (phospho)proteomics. The FEBS Journal 287, 2979–2997.

The Pro in Proteomics – Nederlandse samenvatting

Deze thesis beschrijft een studie rondom enkele aspecten van proline, een van de twintig aminozuren die gezamenlijk de bouwstenen vormen voor onze eiwitten. In hoofdstuk één bekijken we de kenmerken van de verschillende aminozuren en de algemene opbouw van eiwitten. Ook worden enkele aspecten van massa spectrometrie kort uitgelegd.

Proline is een uniek aminozuur, omdat de zijgroep een ring vormt terug naar de amine. Dit amine is daardoor secundair, of wanneer proline in een aminozuurketen aanwezig is zelfs tertiair, wat er voor zorgt dat deze amine basischer is dan die van andere aminozuren. Proline wordt selectief herkent door verschillende type eiwitten die modificaties aanbrengen op andere eiwitten, de zogenaamde schrijvers. De aangebrachte modificaties kunnen de eigenschappen van het gemodificeerde eiwit beïnvloeden, waardoor bijvoorbeeld de vorm, locatie of activiteit van het eiwit kunnen veranderen. Het komt ook voor dat modificaties een eiwit aanmerken voor afbraak. Meerdere modificaties kunnen elkaar ook beïnvloeden. Dit kan in zowel positieve als negatieve zin gebeuren en op verschillende niveaus. In hoofdstuk twee onderscheiden we verschillende vormen van deze wisselwerking en bespreken we het algemene mechanisme waarmee verschillende modificaties geplaatst en verwijderd kunnen worden. Daarnaast geven we een aantal specifieke voorbeelden van mogelijke wisselwerkingen in de humane biologie.

Door de bijzondere structuur van proline is het aminozuur vaak een uitdaging voor proteases, gespecialiseerde eiwitten die aminozuur ketens van andere eiwitten kunnen doorknippen. In de massa spectrometrie gebruiken we graag proteases om de eiwitten te knippen tot wat behapbaardere stukken: kortere aminozuurketens genaamd peptides. Meestal kiezen we voor trypsine, omdat deze protease knipt na arginine of lysine en dus een positieve lading achterlaat op het laatste aminozuur van het peptide. Gecombineerd met de basische amine aan het begin van elk peptide creëert deze aanpak condities die zeer geschikt zijn voor het fragmenteren van de peptides in de massa spectrometer.

Trypsine heeft ook enkele nadelen. Zo komen arginine en lysine residuen eigenlijk te frequent voor, waardoor de resulterende peptiden vrij kort zijn. Voor deze korte peptiden is de kans veel groter dat ze niet uniek zijn en het daardoor niet mogelijk is om zonder twijfel te zeggen uit welk eiwitten ze zijn gekomen. Daarnaast knipt trypsine vaak niet als er een proline of een modificatie in de buurt zit. Trypsine is al zo lang de favoriete protease, dat we alle stukken eiwitten en modificaties die goed zichtbaar zijn met deze methode inmiddels wel gezien hebben. Omdat er ook peptiden te kort zijn of juist geen arginine of lysine hebben, blijft er altijd een stuk van het proteoom onzichtbaar. Voor trypsine is dat ongeveer 13% van het humane proteoom. Om dit, voor trypsine onzichtbare deel van de humane eiwitten, zichtbaar te maken zullen we moeten knippen met een protease die een hele andere selectiviteit heeft dan trypsine zelf. Op die manier kan een complementaire zet peptiden worden gemaakt. In

hoofdstuk drie bekijken we een proline-specifieke protease voor het meten van het humane proteoom en fosfoproteoom. We vergelijken de geïdentificeerde eiwitten met trypsine en zien dat ca. 30% van de eiwitten uniek zijn voor de proline-specifieke protease, EndoPro. Daarnaast vinden we een groot deel van de unieke fosforylaties met EndoPro, dat zijn dus allemaal fosforylaties die we met trypsine niet konden zien. Deze complementaire stukken eiwitten en modificaties maken EndoPro zeer geschikt voor gebruik in (fosfo)proteomics projecten.

Een andere unieke eigenschap van proline is dat het aminozuur relatief vaak gevonden wordt in *cis*-conformatie. Wanneer twee aminozuren een binding aangaan met elkaar wordt de rotatie rond de C-N binding beperkt en zien we in theorie nog maar twee vormen: de *cis* en de *trans* vorm. We bekijken in hoofdstuk vier hoe deze verdeling is voor alle combinaties van aminozuren in het humane proteoom. We zien dat ongeveer 0.35% van de peptidyl bindingen in de *cis* vorm bestaan in onze eiwitten, waarvan de meesten zich voor een proline bevinden. Om de vorm van peptidyl-bindingen voor proline residuen te kunnen bekijken in eiwitten, evalueren we de mate waarin een proline-specifieke protease een voorkeur heeft voor de een van de twee vormen. Onze initiële data lijkt te ondersteunen dat EndoPro inderdaad liever *trans*-proline residuen knipt dan *cis*-proline residuen, waardoor de protease zeer geschikt zou zijn om de isomeer specificiteit in de biologie verder te onderzoeken.

In het laatste hoofdstuk beginnen we aan een combinatie van fosforylatie en isomeer specifieke biologie. Van meerdere eiwitten weten we dat ze isomeer specifiek zijn in hun reacties. Vooral als de reactie plaatsvindt vlak bij de proline. Dit komt waarschijnlijk door de grote structurele verandering in een eiwit als de proline verandert van *cis*- naar *trans* of andersom. We onderzoeken Pin1, een eiwit wat de *cis*/*trans* isomerisatie versnelt van eiwitten die gefosforyleerd zijn op het aminozuur voor de proline. Door de wisselwerking van fosforylatie en conformatie kan een eiwit niet alleen gefosforyleerd zijn of niet, maar voor elk van die toestanden ook nog *cis* of *trans* zijn. Dit mechanisme zorgt voor een extra niveau van regulatie op bepaalde eiwitten. In hoofdstuk 5 kijken we naar de substraten van Pin1 om in beeld te brengen welke eiwitten mogelijk op deze manier gereguleerd worden. De initiële data laat een verrijking zien van RNA gerelateerde eiwitten, die wellicht extra streng gereguleerd worden. Vervolg experimenten zijn nodig om dit te bevestigen.

Addendum

About the author
List of Publications
Acknowledgements

About the author

Saar studied Chemistry at the Vrije Universiteit in Amsterdam, where she was a board member of the study association VCSVU and subsequently chaired their external affairs committee. During her masters, she focussed on Analytical Chemistry, supplemented by the COAST MSc+ program. She developed analytical methods to follow the degradation of tricyclic antidepressants with liquid chromatography and for the analysis of heavy asphaltenes using nonaqueous capillary electrophoresis combined with ion mobility mass spectrometry. In 2016, she started her PhD research in the lab of Albert Heck. During this time, she was also a member of the Future Medicine Fellows, who organised multiple seminars and a well-attended summer school focussing on gene- and celltherapies.

List of Publications

Saar A M van der Laarse*, Aneika C Leney* and Albert J R Heck, **Crosstalk between phosphorylation and O-GlcNAcylation: friend or foe**, FEBS J., 2018 DOI: 10.1111/febs.14491

Saar A. M. van der Laarse*, Charlotte A. G. H. van Gelder*, Marshall Bern, Michiel Akeroyd, Maurien M. A. Olsthoorn and Albert J. R. Heck, **Targeting proline in (phospho)proteomics**, FEBS J., 2020, DOI: 10.1111/febs.15190

Jean-François Greisch, Saar A M van der Laarse, Albert J R Heck, **Enhancing Top-Down Analysis Using Chromophore-Assisted Infrared Multiphoton Dissociation from (Phospho)peptides to Protein Assemblies**, Anal Chem, 2020, DOI: 10.1021/acs.analchem.0c03412

Nadine Prust, Saar van der Laarse, Henk W.P. van den Toorn, Nina M.van Sorge, Simone Lemeer, **In-Depth Characterization of the Staphylococcus aureus Phosphoproteome Reveals New Targets of Stk1**, MCP, 2021, DOI: 10.1074/mcp.RA120.002232

A

* authors contributed equally

Acknowledgements

I've spent about four years in the lab of Albert Heck: an inspiring environment with a lot of kind, smart and unique people and countless amazing instruments. I can honestly say that I've felt like a kid in a candy store and - after finding it - really enjoyed the freedom to pursue my scientific curiosity.

Although often frustrating, my PhD has been an experience that I look back on fondly. I would like to thank my professor, the staff, the technicians and all of my fellow lab members for making it a great journey. I've learned and grown thanks to you all and found some wonderful friends along the way.

As we say in Dutch, "De laatste loodjes wegen het zwaarst". I think that has been especially true for my PhD adventure. I loved doing research, but combining some of that research into this thesis was a lot less enjoyable. A special few ensured - using any means necessary - that I kept moving up my imaginary mountain and it is thanks to them that I have reached this point today.

Therefore, I would like to extend a special thank you to my family and a few close friends. Firstly, I'm grateful to my parents, Annelie en Willem, for my privileged upbringing, the freedom to study whatever I liked, the great atmosphere in our house and the continued support and advice.

Thank you to Charlotte, for dragging me across the finish line and for always drowning our sorrows and celebrating our victories together. You always know when to feed or caffeinate me, or when a little tough love is called for. Thank you for all the fun we had in the lab, for the great discussions, for always being brutally honest with me and for all the weekends you sacrificed helping me to write. Looking forward to many more sorrows to be drowned and victories to be enjoyed together.

Forever grateful to Dallas, for keeping Kyle in check. Thank you for being my friend, for always being flexible around my hectic mess of plans and for all the great times we've had so far. Looking forward to many more in the future (please don't count them out now). I'm confident my career will flourish because of the soldering skills you've taught me and the Hecklab will forever be enlightened with our beautiful, dimmable creation.

I'd like to thank Romy, for pretending to understand my weird chemistry talk and for always going on roadtrips with me. I've stopped counting the years we have been friends, because it makes me feel way too old. You always have my back and I promise to attempt to minimize my throwing up in your company.

My sister Benny deserves a special thank you for being my closest friend and showing me what it means to follow your heart no matter what others may think. I'm proud to call you my sister and I have a deep respect for your perseverance. I'm looking forward many more adventures with you by my

side.

Finally, Jeffrey deserves a special mention for surviving me and my craziness for over a decade now, for being my trusted co-malloot and for having my back no matter what – ok, horses excluded. Thank you for your patience, for powering through my messes and most of all, for your continued support. I'm looking forward to many more years of stealing your blanket and annoying you to no end.