

On studying protein phosphorylation patterns using bottom-up LC–MS/MS: the case of human α -casein

Frank Kjeldsen,*^a Mikhail M. Savitski,^a Michael L. Nielsen,^a Lei Shi^b and Roman A. Zubarev^a

Received 7th February 2007, Accepted 22nd May 2007

First published as an Advance Article on the web 13th June 2007

DOI: 10.1039/b701902e

Most proteomics studies involving mapping post-translational modifications, such as the phosphorylation of serine and threonine, are performed today using the ‘bottom-up’ approach. This approach involves enzymatic cleavage of proteins, most often by trypsin, with subsequent nano-LC–MS/MS. The occupancy rates of phosphosites in proteins may differ by orders of magnitude, and thus the occupancy rate must be reported for each occupied phosphosite. To highlight potential pitfalls in quantifying the occupancy rates, α_{s1} -casein from human milk was selected as a model molecule representing moderately phosphorylated proteins. For this purpose, human milk from one Caucasian woman in the eighth month of lactation was used. The phosphorylation level of caseins is believed to have major implications for the formation of micelles that are involved in delivering valuable calcium phosphate and other minerals to the newborn. Human α_{s1} -casein has been reported to be much less phosphorylated than ruminant caseins, which may indicate a different function of caseins in humans. Revealing the phosphorylation pattern in human casein can thus shed light on its function. The current study found that the sequence region between the residues Ser70 and Ser76 in human α_{s1} -casein is in fact phosphorylated, contrary to previous knowledge. The site of the most abundant phosphorylation is Ser75, in agreement with the known action of the mammary gland casein kinase. There is evidence for the second phosphorylation in that region, possibly at Ser73. Earlier reported positions of phosphorylations at Ser18 and Ser26 are also confirmed, but not the dominance of Ser18 phosphorylation. The occupancy rates at Ser18, Ser26 and Ser75 are estimated to be (7 ± 2) , (20 ± 6) and $(27 \pm 9)\%$, respectively. Owing to differences in the ionization efficiency between phosphorylated and unphosphorylated peptides a 30% error margin is added to the occupancy rates. The highlighted pitfalls of the bottom-up strategy include the sensitivity of enzymes to proximal acidic and phosphorylated residues and the presence of multiple isoforms, including unexpected ones, of the tryptic peptides. The utility of the earlier introduced PhosTS_hunter and ModifiComb approaches for evading the latter pitfall is demonstrated.

Introduction

Among the large number of post-translational modifications (PTMs) described to date, the few that are reversible and thus of regulatory importance in biological processes are of particular interest. Of these PTMs, protein phosphorylation is one of the best understood. The most common type of protein phosphorylation involves the formation of phosphate ester bonds with the hydroxy side chains of serine and threonine, and (much less frequently) tyrosine. Two counteracting enzyme systems, kinases and phosphatases, catalyze protein phosphorylation and dephosphorylation, respectively. In the human organism, there are thousands of protein kinases/phosphatases differing in their substrate specificities and other properties. Regulation *via* protein phosphorylation usually acts through altering the structure and thus the function of

target proteins.^{1–4} Even a single phosphorylation event can have dramatic effect on protein activity.^{5,6}

Essentially, any protein phosphorylation study conducted with mass spectrometry includes as a first step the determination of the amino acid residues that are phosphorylated *in vivo* in a protein present in a given biological system.⁷ The greatest challenge for analysis comes from the fact that proteins are frequently phosphorylated to low stoichiometry, *i.e.* only a small fraction of a given protein may be phosphorylated. Furthermore, every site of reversible phosphorylation is a dynamic balance between the kinase and phosphatase activities. Since the final state of dephosphorylation reactions is in practice never achieved completely, all such sites can be considered phosphorylated, albeit at a trace level. The reverse is also true: phosphorylated sites are never 100% occupied. Thus, tryptic cleavage of phosphoproteins results in a mixture of phosphorylated and unphosphorylated forms of tryptic peptides with the same sequence. Even if the phosphorylated form of a protein is isolated by, for example, immunoprecipitation, the isolated fraction will be a mixture of differentially phosphorylated forms of the same protein, and

^aLaboratory for Biological and Medical Mass Spectrometry, BMC, Uppsala University, Box 583, 75123 Uppsala, Sweden.

E-mail: frankk@bmb.sdu.dk; Fax: +46 18-471-22-44

^bCollege of Public Health, Jilin University, Changchun, China 130021

thus tryptic peptides of that mixture will have different occupancy rates of the phosphorylated sites.

These nuances did not matter much when MS instrumentation could only detect major phosphorylation events, *i.e.* highly occupied sites. But in the past few years, MS technology has significantly improved its sensitivity and dynamic range, enabling the detection and fragmentation of phosphorylated peptides at lower and lower levels. New approaches aimed at the detection of phosphopeptides have emerged that further enhance the efficiency.^{8–14} As a result of this development, the number of reported phosphosites has dramatically increased. In the bottom-up ('shotgun') approach, by far the most frequently used proteomics approach today,^{15,16} a single LC–MS run can reveal tens and even hundreds of phosphosites.^{9,13} However, in order to describe the dynamics of the proteome phosphorylation and retrieve truly valuable biological information, detecting the presence (or absence) of phosphorylation at a particular site is becoming less relevant without simultaneous measurement of the occupancy rate, *i.e.* the ratio of the abundances of phosphorylated and unphosphorylated molecules. Yet today, proteomics studies frequently treat phosphorylation sites in a Boolean manner, reporting them either occupied or not. This can partially be explained by the limitations of commonly used database search engines that assign the sequence and phosphorylation position for each to MS/MS spectrum independently of other spectra (in fact, such independence is considered an advantage as it simplifies statistics-based scoring¹⁵). The relative abundance and retention time (RT) are not used by most search engines, because extracting this information requires processing the chromatographic peak profile, *i.e.* going beyond a single mass spectrum. Yet RT is important, as unphosphorylated and phosphorylated peptides elute within a few minutes of each other,¹⁷ and this information can strengthen the sequence and phosphorylation site assignments made by the search engine.¹⁸ And of course, abundance information is critical for determination of the occupancy rate.

It is often assumed in the bottom-up approach that if a phosphopeptide is not detected in repeated analysis while the corresponding unphosphorylated analogue is detected, then the phosphorylation is either absent or present at a below-threshold level. However, the enzymatic action of trypsin that normally cleaves after basic Arg and Lys residues is known to be reduced in the presence of acidic and phosphorylated residues near the cleavage site.¹⁹ Thus, special care must be taken if the potential phosphorylation site is close to the trypsin cleavage site. Furthermore, it is assumed that HPLC columns do not bind phosphopeptides irreversibly, and that the latter are nicely ionized in electrospray ionization (ESI), so that the observed mass spectra of phosphopeptides are representative of the phosphorylation pattern in the protein. However, the quadruply phosphorylated N-terminal peptide from β -casein is usually missing in a straightforward nano-LC–MS/MS experiment.²⁰ But singly and doubly phosphorylated peptides are usually assumed to be reasonably well behaved in a standard bottom-up experiment.

To illustrate the importance of the above issues in the determination of the phosphorylation patterns of moderately phosphorylated proteins, we selected as a model molecule

α_{s1} -casein from human milk. Caseins (α_{s1} -, α_{s2} -, β - and κ -casein) are the predominant phosphoproteins found in milk of most mammalian species, especially in ruminants.²¹ In humans, β -casein is the primary casein and it comprises up to 70% of the total amount of caseins,²² whereas the α_{s1} -casein level is approximately 5%.²³ Casein is relatively hydrophobic, and poorly soluble in water. In many mammals the primary function of caseins is to form in milk large micelles upon binding of calcium.²⁴ The phosphorylation level of casein has major implications for the formation and stabilization of the micelles,^{24–26} which allow valuable calcium phosphate and other minerals to be delivered to the new-born. Human milk contains fewer caseins than most mammals, and a very low concentration of calcium phosphate. This raises the question of another function of caseins in humans, which is yet to be determined. Revealing the phosphorylation pattern in human casein may shed light on its function.

Human α_{s1} -casein contains 170 amino acids, including 15 serines and three threonines. Previously, a Danish study has found its phosphorylation pattern to be markedly different from the ruminant species.²⁷ Only two phosphorylations have been found in the human molecule – on Ser18 and Ser26. Significantly, no phosphorylation has been observed in humans in the conserved region SSISSSSEE(70–78), which is extensively phosphorylated in the ruminant α_{s1} -casein. The absence of phosphorylation in the region Ser70–Glu78 has been confirmed in analyses of milk from three individuals. Based on HPLC trace analysis of tryptic peptide mixtures, the existence of three variants of human α_{s1} -casein has been postulated: the unphosphorylated form (*ca.* 30%), with a single phosphorylation at Ser18 (*ca.* 50%), and the doubly phosphorylated variant at Ser18 and Ser26 (*ca.* 20%). The moderate degree of phosphorylation (0.9 phosphorylations per molecule on average), high abundance in the sample and thorough characterization by previous research made this molecule an ideal candidate for our purpose.

In the attempt to reproduce the conditions of earlier studies, the starting material was obtained from a Danish individual. We used a similar reversed-phase HPLC method, but different mass spectrometer and ionization technique (on-line ESI FT ICR instead of off-line MALDI TOF). The protein isolation consisted of coarse fractionation followed by separation on a 1D SDS gel. Two enzymes, trypsin and Glu-C, were used to prepare two separate batches of samples. The high mass accuracy of FT ICR and two employed fragmentation techniques ensured high fidelity of the peptide sequence and phosphorylation site assignment. Relative quantification of unphosphorylated and phosphorylated molecules was performed *via* chromatographic peaks (see Discussion below).

The results of the study were surprising, given how thoroughly the protein has been characterized in previous studies. Although abundant phosphorylation at Ser18 and Ser26 was confirmed, equally abundant phosphorylation on Ser75 was found in the region previously reported to be unphosphorylated. Moreover, evidence for double phosphorylation in that region was obtained. These results and their implication for the characterization of phosphorylation patterns by the bottom-up approach are discussed below.

Experimental

Sample collection and preparation

Human milk was collected from a healthy Caucasian woman in the eighth month of lactation. Samples were stored at -80°C until use. To minimize protease activity and degradation, the samples were defrosted in ice water and kept at *ca.* 4°C during the entire sample preparation. 52 mL of milk was centrifuged at $14\,000\text{ g}$ for 45 min at 0°C . The fractionated milk was divided into three parts, as customary in milk analysis: fat globules (3 mL), whey (*ca.* 48 mL) and pellet (1 mL) fractions.

SDS-PAGE and enzymatic hydrolysis

Proteins originating from the pellet and the fat globules were independently separated on one-dimensional SDS-polyacrylamide gels (30–200 kDa), and the protein bands were visualized with colloidal Coomassie Blue. To each SDS-PAGE were added two lanes of approximately 500 μg of proteins. For pellet proteins, each lane of the gel bands was cut into 26 equally sized pieces, and for fat globule proteins into 12 pieces. All gel pieces were washed, and the proteins in them were reduced, alkylated and in-gel digested. Proteins from one set of gel pieces were digested with modified sequence grade trypsin (Promega, Madison, WI) as described previously,²⁸ and proteins from the second set of gel pieces were digested with endoproteinase Glu-C Sequencing Grade (Roche, Netherlands). Proteins for Glu-C digestion were dissolved in a 25 mM ammonium carbonate buffer and Glu-C was added 1 : 30 and incubated for 16 h at 37°C . Finally, the samples were vacuum-centrifuged to remove all organic solvents and reconstituted prior to analysis in 30 μL of HPLC grade water containing 0.1% TFA (Sigma).

Instrumentation

LC-MS/MS analysis was performed on a 7 Tesla hybrid linear ion trap Fourier transform mass spectrometer (Finnigan LTQ FT, Thermo, Bremen, Germany) equipped with a nano-electrospray ion source (Proxeon Biosystems, Odense, Denmark). A nanoflow HPLC system was used on-line with the mass spectrometer. The system (Agilent 1100 nanoflow) included a solvent degasser, nanoflow pump, and a thermostated microautosampler. As solvent A, a mixture of 99.5% water and 0.5% acetic acid was used, and as solvent B, 90% acetonitrile, 9.5% water, and 0.5% acetic acid were used. The peptide sample was automatically loaded at a flow rate of 500 nL min^{-1} onto a 15 cm-long Proxeon nano-ESI emitter (75 μm inner diameter, 360 μm outer diameter) packed in-house with fully end-capped Reprosil-Pur C18 3 μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). After 20 min of loading time the peptides were eluted from the column with a 120 min gradient (4–45% buffer B for $t_1 = 24$ to $t_2 = 114$ min) at a flow rate of 200 nL min^{-1} . MS analysis was performed using unattended data-dependent acquisition mode in which the mass spectrometer automatically switched between a high-resolution survey scan (resolution at 100 000 in the m/z range 300–1500, the automatic gain control was set to $\text{AGC} = 3\,000\,000$) followed by medium-high resolution (25 000) fragmentation spectra of the two most abundant

peptides in the charge state two or higher. As fragmentation techniques, ECD ($\text{AGC} = 900\,000$ ions, two microscans summed together) was used followed by CAD ($\text{AGC} = 600\,000$ ions, one microscan).²⁹ The maximum accumulation time during CAD and ECD was set to 800 ms. The masses of fragmented ions were put on an exclusion list for 60 s.

Data processing: phosphopeptide identification and quantification

The original Xcalibur (Thermo Electron, Waltham, MA) binary data (RAW-files) were converted by BioWorks TurboSEQUENT (Thermo Electron) into a set of peak lists (DTA-files). These DTA-files contain the neutral mass and the charge state of the precursor, as well as the m/z values and intensities of all fragment ion peaks in the spectrum above a certain cut-off intensity value. For each peptide selected for MS/MS, two DTA-files were generated, one for ECD and another one for CAD. All software for processing the DTA-files was developed in-house in Java language.

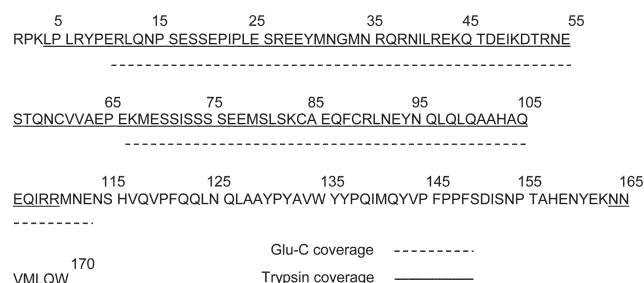
In the first step, the PhosTS_hunter program processed the CAD DTA-files as previously described to detect phosphopeptides.¹⁴ Briefly, this is done by testing if a fragment ion corresponding to a neutral loss of phosphoric acid with or without a water molecule is present in the datafile.

Simultaneously, ECD and CAD datafiles were merged as described²⁹ and submitted to the Mascot search engine using the IPI human protein database (Mascot determined threshold scores 25 for tryptic peptides and 20 for Glu-C peptides). These thresholds ensure >95% statistical confidence in the peptide identification.^{29,30} The quantification of the found phosphorylated peptides and their unphosphorylated analogues was performed by integration of the corresponding ion chromatographic peak using the program Xcalibur (Thermo).

Results and discussion

Peptide identification

Owing to imperfect gel fractionation, peptides from $\alpha_{1\text{s}}$ -casein were found in several gel fractions. The most abundant signal of these peptides was obtained from fraction 14 for pellet proteins and fraction 5 for fat globule proteins. The trypsin peptides covered a total of 67% of the $\alpha_{1\text{s}}$ -casein protein sequence, while Glu-C peptides covered 54% of the protein sequence. Combined, the total protein sequence coverage was 69% (Scheme 1). Importantly, the N-terminal part from the 4th to the 113th residue was covered.



Scheme 1

PhosTS_hunter revealed 114 different phosphopeptides in pellet protein samples and 57 phosphopeptides in fat globule samples. Eight of the phosphopeptides were attributed to α_{1s} -casein and will be the subject for this study. With PhosTS_hunter we retrieved more phosphopeptides than with Mascot search alone. This is because PhosTS_hunter pulls out phosphopeptides irrespective of any other additional unexpected modifications attached to peptides. To achieve this in the Mascot search we would have to allow all additional modifications in the search, which would raise the false positive rate and increase the search time dramatically.

Phosphorylation at Ser18 and Ser26

These sites of phosphorylation were expected, and were found both in the trypsin digest and the Glu-C digest. Fig. 1a shows the tryptic ion chromatogram for m/z values 931.93 and

891.95, corresponding to doubly charged ions of the peptide 12–27 without and with phosphorylation, respectively. The RT difference between these molecules is 5.2 min. On a closer inspection (right inset), there is a second peak in the chromatogram corresponding to a peptide with the same mass (m/z 931.93) eluting almost one minute earlier than the main component with RT = 68.44 min. Both these peptides were selected for MS/MS, and their identities were deduced unequivocally. Fig. 2 shows the corresponding CAD mass spectra. Both peptides have the same sequence and carry one phosphorylation each, but the sites of phosphorylation are different: Ser18 in the more abundant peptide (lower panel) and Ser26 in the less abundant one (upper panel). Furthermore, a small chromatographic peak was observed (Fig. 1a, left inset) at a 7.5 min longer retention time than the unphosphorylated form. This peak is due to peptide dicationations with m/z 971.921. This value is 8 ppm away from the theoretical m/z = 971.913 of the doubly phosphorylated peptide (the determined tolerable deviation in the survey scan was 10 ppm), to which this ion was attributed.

Fig. 1b shows the Glu-C digest chromatogram for m/z 686.83 and 726.81, corresponding to unphosphorylated and phosphorylated versions of the peptide SSEPIPLESREE(18–29). The RT difference between these peptides is 5.0 min, but here no chromatographic peak corresponding to doubly phosphorylated ions is observed (left inset). MS/MS of ions with m/z 726.81 revealed that the main site of phosphorylation is Ser26. The peptide with Ser18 phosphorylated is expected to elute later, but no corresponding peak is observed in the ion chromatogram (left inset). Discussion of this unexpected result follows below.

Phosphorylation at Ser75

The tryptic peptide MESSISSSEEMSLSK(68–83), containing the serine-rich region between Ser70 and Ser76, was present in several isoforms (detailed discussion below), the most abundant one being with the Met79 oxidized. The unphosphorylated peptide eluted at RT = 54.4 min, while phosphorylated one at RT = 59.3 min. The ECD spectrum of the latter is shown in Fig. 3: phosphorylation at Ser75 is clearly identified by z_8' and z_9' ions. The ion chromatogram (not shown) contains a smaller peak with a RT = 58.4 min corresponding to phosphorylation at a different residue. However, MS/MS data on this peak are absent.

Glu-C digest analysis confirmed phosphorylation at Ser75. Fig. 4 shows both CAD and ECD spectra of 2+ ions of the peptide KM(O)ESSISSSSEE(67–78) with oxidized Met68. CAD produced extensive backbone cleavage accompanied by abundant phosphate losses from b and y' fragments (Fig. 4a). The fraction of b and y' fragments that remained phosphorylated assigned this modification to the region Ser75–Ser76. ECD, as expected, gave no losses of phosphate group from c' ions and located the phosphorylation site unequivocally at Ser75 (Fig. 4b). This unambiguous assignment of the phosphorylation site by ECD is typical for the analyzed phosphopeptides in this study and demonstrates the advantage of performing both CAD and ECD fragmentation of phosphopeptides. The difference in the retention times between

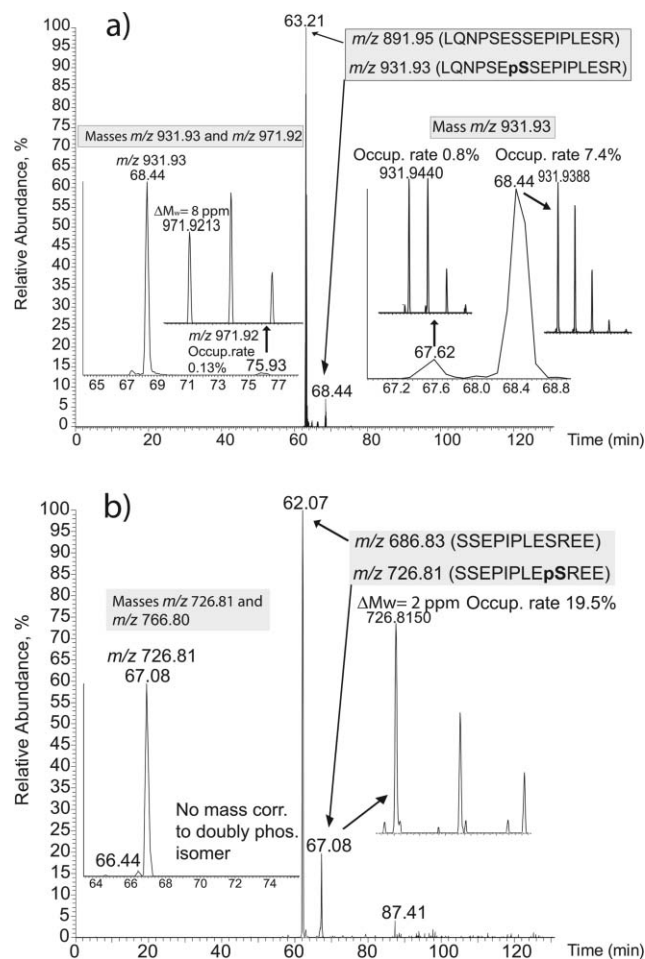


Fig. 1 (a) Ion chromatogram of the tryptic digest for m/z values 931.93 and 891.95 corresponding to doubly charged ions of the peptide 12–27 without and with phosphorylation, respectively. Right inset: two chromatographic peaks corresponding to phosphopeptides with different phosphorylation sites. Left inset: a small chromatographic peak is likely due to doubly phosphorylated peptide. (b) Ion chromatogram of the Glu-C digest for m/z 686.83 and 726.81, confirming the presence of phosphorylation at Ser26. Left inset: no chromatographic peak is present right of the main peak, which would be expected for doubly phosphorylated peptide.

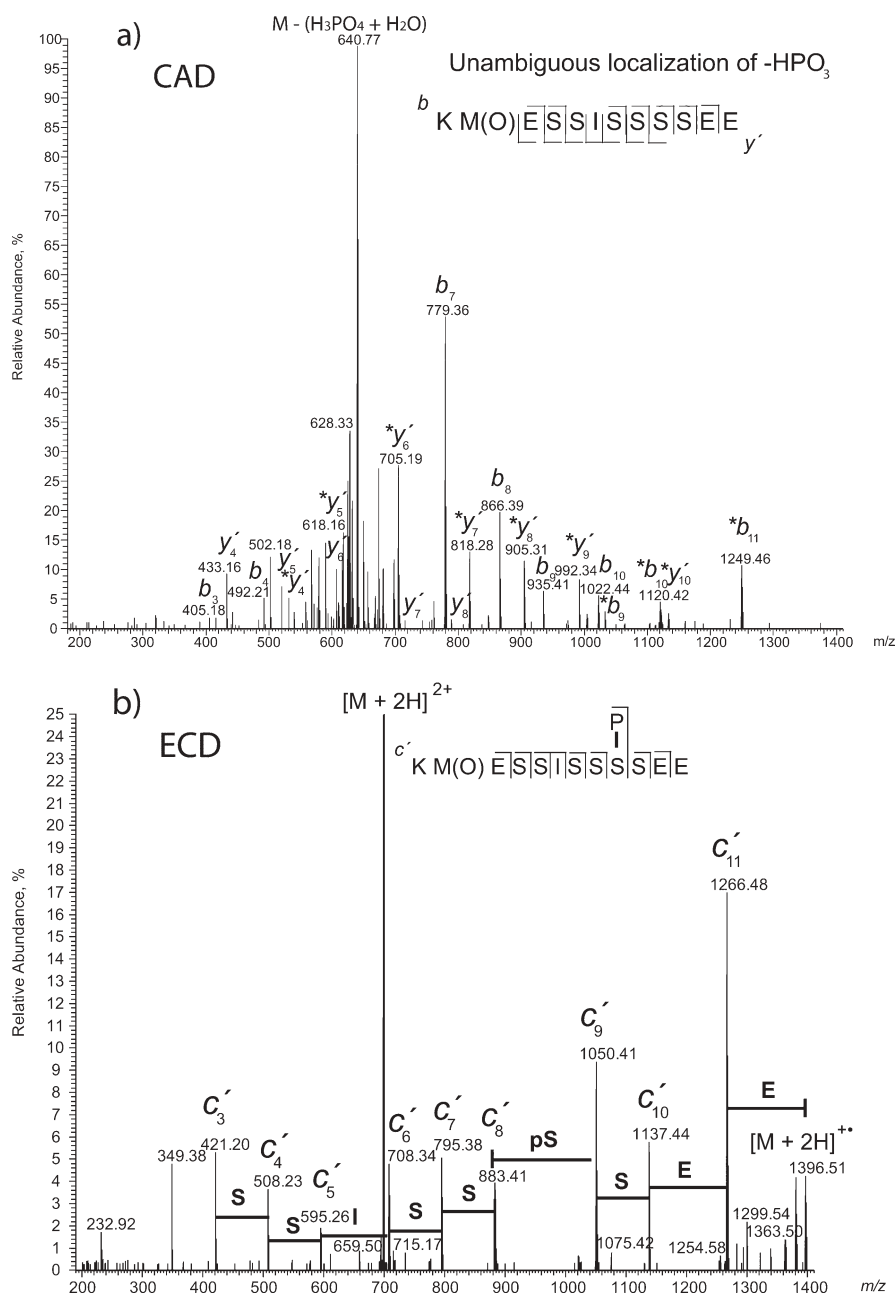


Fig. 4 (a) CAD mass spectrum of 2+ ions of the phosphorylated Glu-C peptide KMESSISSSSEE(67–78) with oxidized methionine [the asterisk (*) indicates fragment ions carrying the phosphate group]; (b) ECD of the same ions. While CAD identified Ser75–Ser76 as the phosphorylation site, ECD clearly located it at Ser75.

Quantification of the occupancy rates

In general, if both unphosphorylated and phosphorylated versions of a peptide can be detected in the same charge state, the ratio of the abundances of the corresponding ions (including all isotopes) can be used as a proxy for the occupancy rates. Steen *et al.* have found this method to give a statistical spread of 10% and average systematic errors of 25–30%.³¹ The main sources of the error are the different ionization efficiencies of unphosphorylated and phosphorylated peptides as well as the difference in average charge states of these peptides in electrospray ionization. Since the molecules are separated by the gradient HPLC, they elute and

become ionized at somewhat different solvent conditions, which add to the differences in ionization efficiencies. Different rates of chemical dephosphorylation as well as phosphate losses in the gas phase for different molecules also contribute to the error. The receipt of Steen *et al.* is to determine for each sequence relative ionization efficiencies of differently phosphorylated molecules using synthetic peptides.³¹ This is an accurate approach, but it can be prohibitively expensive in high-throughput measurements involving large numbers of phosphopeptides. However, a 30% error in the abundance ratio can be tolerated in many situations, including our study. The relative abundance of each site in the peptide 12–27 can be deduced from the integrals of the chromatographic peaks in

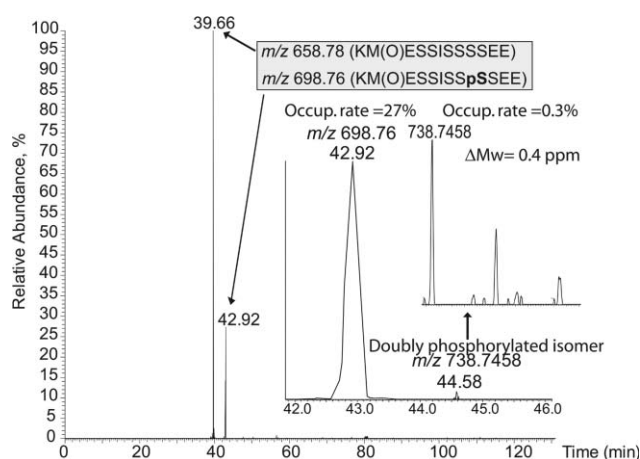


Fig. 5 Chromatographic peaks of the unphosphorylated and phosphorylated 2+ ions of the Glu-C peptide KM(O)ESSISSSSEE(67–78). Inset shows the chromatographic peak of a 2+ ion with m/z that deviates from the theoretical m/z 738.7461 of the doubly phosphorylated peptide by 0.4 ppm.

Fig. 1a. The abundance of the pSer18 (RT = 68.4 min) peptide is 7.4%, and of the pSer26 (RT = 67.6 min) peptide is 0.8% of the abundance of the unphosphorylated peptide. The doubly phosphorylated peptide 12–27 (RT = 75.93 min) was measured to have an occupancy rate of only 0.1%.

The latter estimate made for the tryptic mixture conflicts with that for the Glu-C digest in Fig. 1b, where phosphorylation at Ser26 appears to be dominant [occupancy rate $(20 \pm 6)\%$]. This conflict may be due to the specificity of respective enzymes. Since trypsin discriminates against cleavages next to acidic and phospho-sites, the peptide LQNPSESSEPIPLESR(12–27) with phosphorylation at Ser26 could have been produced in a relatively smaller quantity compared with the pSer18 variant (Fig. 1a). On the other hand,

Glu-C also appears to discriminate against cleavages next to phosphoserine, which reduced the relative quantity of pSSEPIPLESR(18–29) compared with the pSer26 analogue in Fig. 1b. Thus for the respective enzymes, only phosphorylations distant from the cleavage sites can be quantified: Ser19 for trypsin and Ser26 for Glu-C. Unfortunately, no peptide was found in either mixture that, due to a missed cleavage, would contain both phosphosites distant from the termini.

In quantitative terms, phosphorylation at Ser26 appears to be as abundant as at Ser18. The previous finding²⁶ that the pSer18 occupancy rate was 2.5 times higher than that for pSer26 could be due to the use of trypsin as the cleavage enzyme.

Quantification of phosphorylation in the peptide KM(O)ESSISSSSEE(67–78) by the area of the chromatographic peaks of 2+ ions is shown in Fig. 5. In that peptide the positions of the phosphosites were distant to both termini, and thus quantifications were considered reliable. The relative abundance of the singly phosphorylated peptide at Ser75 was $(27 \pm 9)\%$ and of doubly phosphorylated $(0.3 \pm 0.1)\%$ of that of the unmodified peptide. Because the 1-D gel separation of the milk proteins was imperfect we detected the peptide KM(O)ESSISSSSEE(67–78) in six different fractions. To test the reproducibility of the quantitative measurements we extracted the ion signal abundances of the 2+ unphosphorylated (m/z 658.780) and the 2+ phosphorylated peptides (m/z 698.763) from these fractions and obtained an average value of 22% with a SD of 9.8%. These values show that the quantification of the occupancy rate for this peptide is reproducible and within the expected error for repetitive measurements.

Isoforms of the peptide 68–83

As already mentioned, the peptide M⁶⁸ESSISSSSEEMSLSK⁸³ appeared in multiple isoforms. The presence of two methionines provided four chromatographically separated isoforms

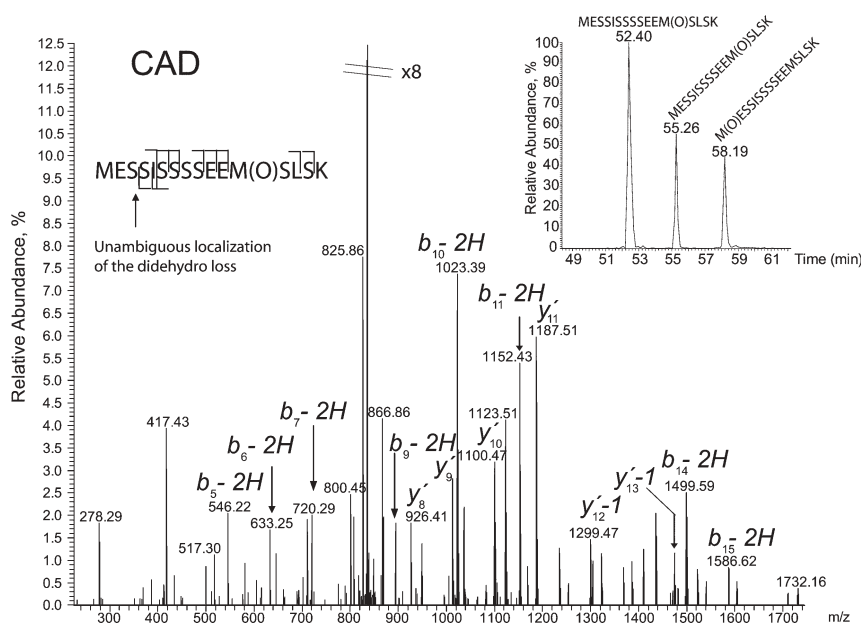
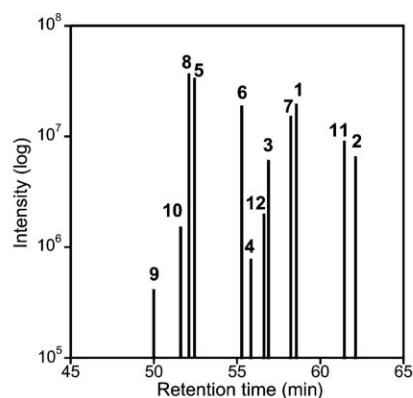


Fig. 6 CAD mass spectrum of the tryptic peptide MESSISSSSEEM(O)SLSK(68–83) eluting at RT = 52.40 min with Met79 oxidized and didehydro loss from either Ser70 or Ser71. Inset shows the chromatographic peaks of three isoforms of the peptide.

of unphosphorylated molecule: with no oxidation (RT = 60.4 min), double oxidation (RT = 51.5 min), and two singly-oxidized forms (RT = 54.4 min for Met79 oxidized and 57.4 min for Met68 oxidized). The most abundant form was that with a single methionine (Met68) oxidized. Moreover, each of the four molecules appears in an additional form corresponding to the loss of 2.019 ± 0.007 Da, or H_2 . The location of this loss was found to occur in the region Ser70–Ser71 (Fig. 6). Serine residues are known to undergo loss of H_2 as a PTM, forming an aldehyde (formylglycine). The abundance of the ions with the loss (RT = 52.40 min) was higher than that without the loss (RT = 55.26 min) by a factor of *ca.* 2 (Fig. 6, inset), suggesting that this was the dominant modification.

The presence of multiple isoforms of this peptide may be the reason why the previous study failed to detect phosphorylation in that region.²⁷ In that study, peptides were identified by their molecular masses, and the sequence was determined by Edman degradation. The presence of multiple isoforms of a peptide can complicate the analysis, as Fig. 7 demonstrates. On that



#	Isoform (modification site in Bold)	Modification(s) with unknown location	<i>m/z</i>
1	MESSISSSSEEMSLSK	Didehydro	858.8674
2	MESSISSSSEEMSLSK	Phosphorylation	898.8575
3	MESSISS p SSEEM(O)SLSK	Didehydro	906.8539
4	M(O) ESSISS p SSEEMSLSK	Didehydro	906.8538
5	MESSISSSSEEM(O)SLSK	Didehydro	866.8666
6	MESSISSSSEEM(O)SLSK		867.8771
7	M(O) ESSISSSSEEMSLSK		867.8776
8	M(O) ESSISSSSEEM(O)SLSK		875.8730
9	MESSISSSSEEM(Hse_lact)SLSK*		557.5875
10	ESSISSSSEEM(O)SLSK		802.3591
11	MESSISSSSEEMSLSK (base peptide)		859.8804
12	M(O) ESSISS p SSEEM(O)SLSK		915.8604

* Hse_lact = Homoserine lactone; loss of CH_4S

Fig. 7 (Top) Absolute abundances and retention times for 12 different isoforms of the tryptic peptide MESSISSSSEEMSLSK(68–83). Numbers in the diagram refer to the list of isoforms (bottom) found in the LC run.

figure, the relative abundances and retention time of all detected isoforms of the peptide MESSISSSSEEMSLSK are presented. It should be noted that not all modifications have been detected because not all peptides were selected for MS/MS, so that the total number of modified versions is surely much larger. Note that the unmodified peptide (entry 11) was only the sixth most abundant among the 12 related peptides. The retention times of these peptides stretch over 12 min, a significant portion of the overall RT range of peptides in our experiment (*ca.* 30 min).

Conclusions

Using advanced mass spectrometry techniques, we have measured the occupancy rates of phosphorylation sites in human α_{s1} -casein. α_{s1} -Casein was extracted from fresh milk from a healthy Caucasian woman in the eighth month of lactation. The current study found that the sequence region between the residues Ser70 and Ser76 in human α_{s1} -casein is phosphorylated, contrary to previous knowledge. The site of the most abundant phosphorylation is Ser75, in agreement with the known action of the mammary gland casein kinase. There is evidence for the second phosphorylation in that region, with Ser73 as the most likely site. Earlier reported phosphorylations of Ser18 and Ser26 are confirmed. The occupancy rates at Ser18, Ser26 and Ser75 are estimated to be $(7.4 \pm 2)\%$, $(19.5 \pm 6)\%$, and $(27 \pm 9)\%$, respectively. In the quantitative measurements a 30% error is accepted. The error is mainly due to differences in the ionization efficiency of phosphorylated and unphosphorylated peptides.

These findings became possible through the use of nano-LC-MS/MS based on a high-resolution FT ICR mass spectrometer with two complementary fragmentation techniques. This study confirmed the utility of the PhosTS_hunter approach and highlighted some of the pitfalls of the bottom-up strategy applied to profiling phosphorylated sites. The first pitfall is the occupancy rates that may differ by orders of magnitude. Thus, reporting an occupied phosphosite without simultaneous specification of the occupancy rate makes little sense. The second pitfall is the enzyme action. Both trypsin and Glu-C become much less efficient if the cleavage site precedes an acidic residue, including phosphorylation residues. Therefore, one should use quantification only for phosphorylations remote from the peptide termini, which often can be achieved only using an enzyme complementary to trypsin. Finally, the fact that peptides may appear in several different isoforms, including unexpected ones, can significantly complicate the analysis. Software for revealing the ‘family trees’ of peptides, such as ModifiComb³² or other spectral alignment algorithms³³ may prove invaluable for such analysis.

Acknowledgements

The authors are grateful to Karin Kjeldsen for providing the milk samples. This work was supported by the Knut and Alice Wallenberg Foundation and Wallenberg Consortium North (grant WCN2003-UU/SLU-009 to R. A. Z. and instrumental grant to R. A. Z. and Carol Nilsson) as well as the Swedish

research council (grants 621-2004-4897, 621-2002-5025 and 621-2003-4877 to R. A. Z.).

References

- 1 T. Hunter, *Cell*, 1987, **50**, 823–829.
- 2 B. Duclos, S. Marcandier and A. J. Cozzzone, *Methods Enzymol.*, 1991, **201**, 10–21.
- 3 M. Karin and T. Hunter, *Curr. Biol.*, 1995, **5**, 747–757.
- 4 L. N. Johnson and D. Barford, *J. Biol. Chem.*, 1990, **265**, 2409–2412.
- 5 I. Joung, T. U. Kim, L. A. Stolz, G. Payne, D. G. Winkler, C. T. Walsh, J. L. Strominger and J. Shin, *Proc. Natl. Acad. Sci. U. S. A.*, 1995, **92**, 5778–5782.
- 6 J. D. Watts, M. J. Welham, L. Kalt, J. W. Schrader and R. Aebersold, *J. Immunol.*, 1993, **151**, 6862–6871.
- 7 R. Aebersold and D. R. Goodlett, *Chem. Rev.*, 2001, **101**, 269–295.
- 8 L. Andersson and J. Porath, *Anal. Biochem.*, 1986, **154**, 250–254.
- 9 S. B. Ficarro, M. L. McClelland, P. T. Stukenberg, D. J. Burke, M. M. Ross, J. Shabanowitz, D. F. Hunt and F. M. White, *Nat. Biotechnol.*, 2002, **20**, 301–305.
- 10 Y. Ikeguchi and H. Nakamura, *Anal. Sci.*, 1997, **13**, 479–483.
- 11 S. A. Beausoleil, M. Jedrychowski, D. Schwartz, J. E. Elias, J. Villen, J. X. Li, M. A. Cohn, L. C. Cantley and S. P. Gygi, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 12130–12135.
- 12 M. Gronborg, T. Z. Kristiansen, A. Stensballe, J. S. Andersen, O. Ohara, M. Mann, O. N. Jensen and A. Pandey, *Mol. Cell. Proteomics*, 2002, **1**, 517–527.
- 13 A. Gruhler, J. V. Olsen, S. Mohammed, P. Mortensen, N. J. Faergeman, M. Mann and O. N. Jensen, *Mol. Cell. Proteomics*, 2005, **4**, 310–327.
- 14 T. Kocher, M. M. Savitski, M. L. Nielsen and R. A. Zubarev, *J. Proteome Res.*, 2006, **5**, 659–668.
- 15 A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik and J. R. Yates, *Nat. Biotechnol.*, 1999, **17**, 676–682.
- 16 E. Lasonder, Y. Ishihama, J. S. Andersen, A. M. W. Vermunt, A. Pain, R. W. Sauerwein, W. M. C. Eling, N. Hall, A. P. Waters, H. G. Stunnenberg and M. Mann, *Nature*, 2002, **419**, 537–542.
- 17 P. R. Liu, C. L. Feasley and F. E. Regnier, *J. Chromatogr. A*, 2004, **1047**, 221–227.
- 18 M. Palmblad, M. Ramstrom, K. E. Markides, P. Hakansson and J. Bergquist, *Anal. Chem.*, 2002, **74**, 5826–5830.
- 19 N. J. Morris, M. Bushfield, B. E. Lavan and M. D. Houslay, *Biochem. J.*, 1994, **301**, 693–702.
- 20 A. Stensballe, S. Andersen and O. N. Jensen, *Proteomics*, 2001, **1**, 207–222.
- 21 R. Jenness and D. E. Sloan, *Dairy Sci. Abstr.*, 1970, **32**, 599–610.
- 22 R. J. Carroll and J. J. Basch, *Food Microstruct.*, 1985, **4**, 323–331.
- 23 L. K. Rasmussen, H. A. Due and T. E. Petersen, *Comp. Biochem. Physiol., B: Biochem. Mol. Biol.*, 1995, **111**, 75–81.
- 24 H. M. Farrell, E. L. Malin, E. M. Brown and P. X. Qi, *Curr. Opin. Colloid Interface Sci.*, 2006, **11**, 135–147.
- 25 S. M. Sood, P. J. Herbert and C. W. Slattery, *J. Dairy Sci.*, 1998, **81**, 3149–3153.
- 26 S. M. Sood, G. Erickson and C. W. Slattery, *Protein J.*, 2005, **24**, 227–232.
- 27 E. S. Sorensen, L. Moller, M. Vinther, T. E. Petersen and L. K. Rasmussen, *Eur. J. Biochem.*, 2003, **270**, 3651–3655.
- 28 M. Wilm, A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis and M. Mann, *Nature*, 1996, **379**, 466–469.
- 29 M. L. Nielsen, M. M. Savitski and R. A. Zubarev, *Mol. Cell. Proteomics*, 2005, **4**, 835–845.
- 30 D. N. Perkins, D. J. C. Pappin, D. M. Creasy and J. S. Cottrell, *Electrophoresis*, 1999, **20**, 3551–3567.
- 31 H. Steen, J. A. Jebanathirajah, M. Springer and M. W. Kirschner, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 3948–3953.
- 32 M. M. Savitski, M. L. Nielsen and R. A. Zubarev, *Mol. Cell. Proteomics*, 2006, **5**, 935–948.
- 33 D. Tsur, S. Tanner, E. Zandi, V. Bafna and P. A. Pevzner, *Nat. Biotechnol.*, 2005, **23**, 1562–1567.