# Top Down Milk Protein Identification and Relative Quantification by Q Exactive Mass Spectrometer

*Terry Zhang, David M. Horn, Charles Yang and Dipankar Ghosh*
*Thermo Fisher Scientific, San Jose CA*

Thermo SCIENTIFIC

# Overview

**Purpose:** Mixtures of casein and whey proteins were characterized by top-down mass spectrometry in order to determine forms to be quantified in milk samples. These identified forms have been used to calculate the whey protein/casein protein ratio studies.

**Methods:** Whey and casein proteins were mixed and top-down LC-MS/MS data of these samples were acquired on a benchtop quadrupole-Orbitrap mass spectrometer. ProSightPC 3.0 was used for intact protein identification. SIEVE 2.1 was used for whey protein/casein protein ratio calculations.

**Results** With top-down approach, we identified α –lactalbumin, beta-lactoglobulin, α–s1-casein, α–s2-casein, k-casein and β–casein in these samples. Different forms of the milk proteins, including intact, truncation products, and with unexpected modifications forms were identified. Using PCA software, whey protein to casein ratios were calculated. Consistent whey to casein ratios were obtained for all different forms of proteins.

# Introduction

Whey proteins, most notably alpha-lactalbumin and beta-lactoglobulin, are currently being investigated for their potential positive health benefits. Whey proteins are a substantial percentage of human milk (~60%) while such proteins are less abundant in cow milk (~20%). Currently, the nutritional requirements of infant formulas require comparison of their amino acid composition in comparison to human milk, but the calculation of the ratio of whey protein/casein protein currently is not required in most countries. However, in China, there is an emerging requirement that at least 60% of protein content in infant formula are from whey proteins but there currently is no standard method to calculate this ratio. The work described in this poster describes the identification of the various forms of caseins and whey proteins as a first step and then intact protein-based method for determination of the whey vs. casein protein ratio. A surprising number of casein protein forms were identified, indicating that quantification based on a single target casein or whey mass or using bottom up approach with peptides will underestimate the amount of that protein in a given sample.

# Methods

Whey (Hilmar Ingredients) and casein (Sigma-Aldrich, Cat No. C8654-550G) powders were prepared in 5% acetonitrile, 95% water with 0.1% formic acid and were mixed with different whey to casein ratios (1:3, 1:2, 1:1 and 2:1). Each protein solution was analyzed in duplicate by ESI-LC-MS/MS analyses using a Thermo Scientific™ Hypersil™ BioBasic C8 reversed phase 1 mm x 100 mm column packed with 5 um particles in conjunction with a Thermo Scientific™ Q Exactive™ mass spectrometer at 140,000 resolution. The LC/MS/MS runs were subsequently analyzed using a prerelease version of ProSightPC 3.0. ProSightHT was first used to deconvolve the datasets using Xtract as the data reduction algorithm. The resulting .puf files were loaded individually and all spectra were analyzed by appending and searching the three searches described in the next section. Results from each dataset were exported to a ProSightPC repository and the results were reviewed in a single repository report. Proteins with expectation values better than 1e-6 were considered to be valid. The LC/MS elution chromatogram were extracted to get protein's molecular weight. From the extracted MS full chromatogram , the whey to casein protein relative ratios were then calculated using Thermo Scientific™ SIEVE™ 2.1.

## Results

### Protein Identification Using ProSightPC

The datasets were first searched in ProSightPC against a full bovine protein database with no indexed modifications and a 60 kDa precursor search tolerance. This type of search allows ProSightPC to identify the proteins in the casein and whey samples without fully characterizing them against a large database such as bos taurus (87708 proteoforms). As expected, alpha-S1-casein, alpha-S2-casein, beta-casein, kappa-casein, and beta-lactoglobulin were identified as well as glycosylation-dependent cell adhesion molecule 1 (GLCM1). Alpha-lactalbumin, a major component of whey proteins, was confidently identified after reduced and alkylated.

To identify and characterize as many different forms of the identified proteins above, a flatfile database using the 7 accession numbers above was created using up to 20 concurrent modifications per sequence allowing for all annotated modifications. The resulting sequence database contained 39 basic sequences and 33651 proteoforms. Each spectrum in each dataset was searched 3 times:

1) Absolute mass search with 1.02 Da precursor mass accuracy and 10 ppm fragment accuracy to find forms that exactly match predicted intact proteoforms in the database.

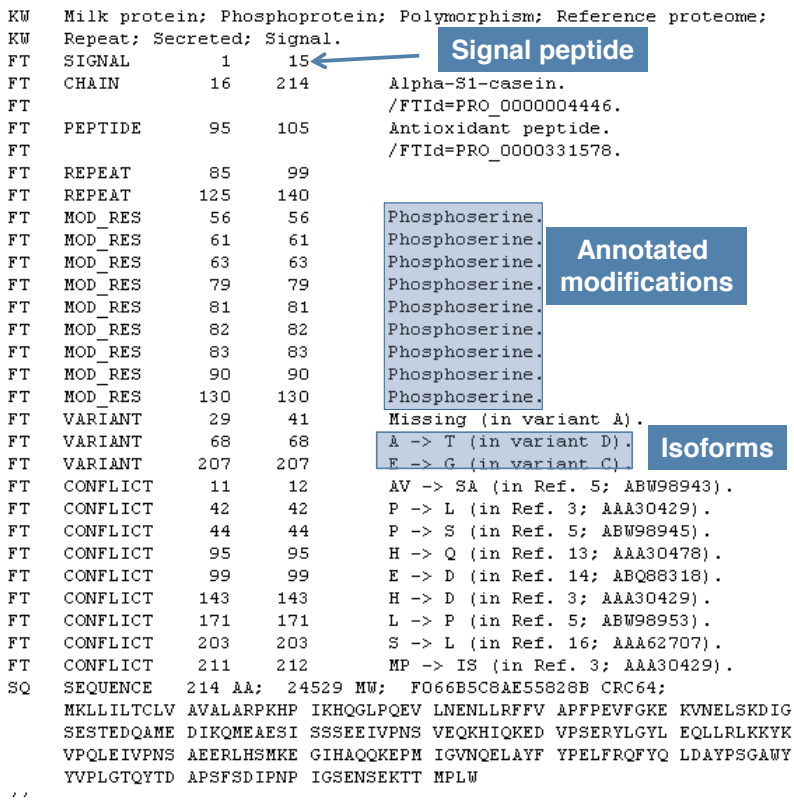
2) Biomarker search with 10 ppm precursor mass accuracy and 10 ppm fragment mass accuracy to find truncation products.

3) Absolute mass search "delta m" with 25000 Da precursor tolerance and 15 ppm fragment tolerance to find forms of the target proteins that were not annotated in the original flatfile and do not match a truncated form.

The identification with the lowest expectation value from the three searches above was chosen as the best candidate for each MS/MS spectrum. The results of the searches produced 14 matched intact forms for the 6 proteins excluding alpha-lactalbumin and in addition 59 truncated and partially characterized forms.

The whey proteins where for the most part detected in only a few primarily intact forms, while the casein proteins were detected in many different forms that included a substantial number of truncation products. One such example of such complexity is alpha-casein. For this protein, there are 9 known phosphorylation sites, a signal peptide that is removed in the mature form, and 3 known sequence variants, as annotated in the Uniprot flatfile (Figure 1).

**FIGURE 1. UniProt flatfile entry for alpha-S1-casein, accession number P02662. The entry indicates that up to 9 phosphoserines may be identified on specific sites and there are at least four different known variants. Intact forms of the primary sequence with 6, 7, 8, and 9 phosphorylations were also detected. Also, less abundant proteoforms with the variant C sequence were also detected.**



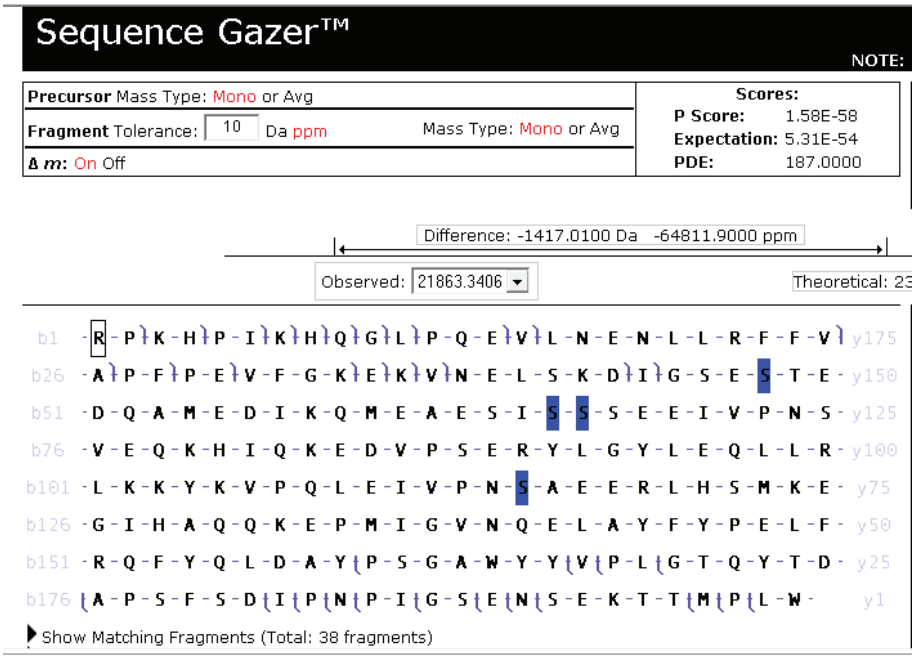
The list of identified forms of alpha-casein can be found in Table 1. For the first search with narrow precursor and product ion tolerances, 6 different intact forms of alpha-casein were detected, all with the signal peptide removed, including two forms of variant C (E->G) with 7 and 8 phosphorylations. The subsequent biomarker search identified 16 different truncated forms of alpha-S1-casein that contained both the N-terminus and the C-terminus. Depending on the site of truncation, many of these forms were also highly phosphorylated. Finally, the error tolerant absolute mass search identified 5 forms that could not be completely characterized. The entries in table 1 denoted as "not fully characterized" were identified by search number 3, the absolute mass search with a large precursor mass tolerance. For these identifications, large stretches of sequence were identified from one or both termini, but the mass difference between the target protein sequence and the measured monoisotopic

**Table 1. List of identified forms of alpha-casein. The best E-value column indicates the protein spectral match that produced the most confident result across the 10 raw data files.**

| Nominal Mass (Da) | RT (min) | Identified form | Best E-value |
|---|---|---|---|
| 2617 | 4.1 | Truncation | 3.35e-39 |
| 7684 | 5.1 | Not fully characterized | 5.07e-27 |
| 7851 | 5.6 | Not fully characterized | 3.27e-29 |
| 8018 | 6.1 | Not fully characterized | 5.20e-15 |
| 8561 | 7.3 | Truncation | 2.26e-24 |
| 8633 | 7.3 | Truncation | 1.05e-59 |
| 9519 | 6.4 | Truncation, 7 phosphorylations | 4.28e-27 |
| 10871 | 6.7 | Truncation, 1 phosphorylation | 5.02e-31 |
| 11162 | 6.5 | Truncation, 1 phosphorylation | 5.64e-31 |
| 11290 | 6.3 | Truncation, 1 phosphorylation | 1.30e-32 |
| 12747 | 5.3 | Truncation, 7 phosphorylations | 7.01e-35 |
| 13348 | 6.0 | Truncation, 1 phosphorylation | 3.12e-22 |
| 13734 | 5.7 | Truncation, 5 phosphorylations | 9.93e-13 |
| 14099 | 5.9 | Truncation, 1 phosphorylation | 5.80e-39 |
| 14750 | 5.8 | Truncation, 2 phosphorylations | 6.07e-26 |
| 19338 | 5.0 | Not fully characterized | 1.54e-08 |
| 19416 | 5.1 | Truncation, 7 phosphorylations | 9.34-15 |
| 19496 | 5.4 | Truncation, 8 phosphorylations | 1.07e-12 |
| 21863 | 7.0 | Not fully characterized | 5.31e-34 |
| 23347 | 7 | Truncation, 8 phosphorylations | 3.71e-45 |
| 23427 | 7.1 | Truncation, 9 phosphorylations | 1.53e-23 |
| 23440 | 6.4 | Intact, 6 phosphorylations | 8.48e-56 |
| 23448 | 6.5 | Intact, 7 phosphorylations, variant C | 4.13e-49 |
| 23520 | 6.5 | Intact, 7 phosphorylations | 5.02e-65 |
| 23528 | 6.7 | Intact, 8 phosphorylations, variant C | 5.74e-45 |
| 23600 | 6.6 | Intact, 8 phosphorylations | 4.87e-60 |
| 23680 | 6.9 | Intact, 9 phosphorylations | 1.39e-52 |

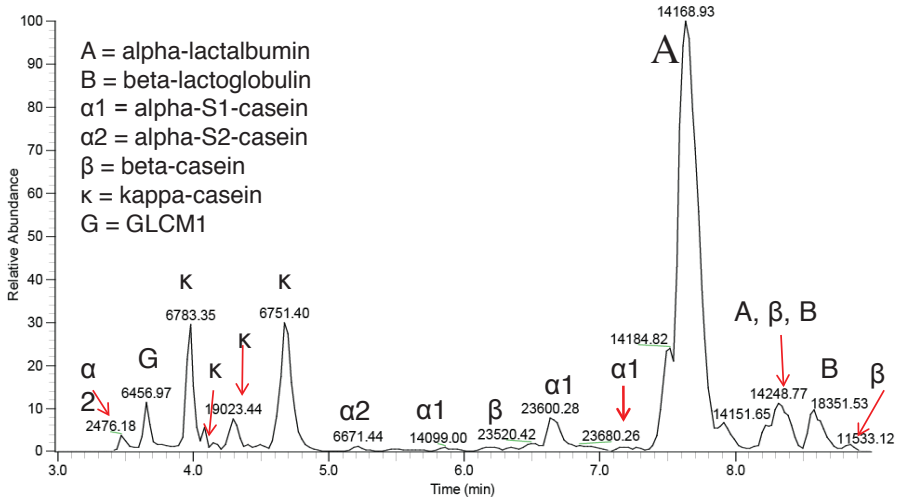
mass could not be exactly localized on the protein sequence. One interesting example of this is the alpha-s1-casein proteoform with nominal monoisotopic mass of 21863 Da. This protein form is relatively abundant and the Sequence Gazer result (Figure 2) shows substantial sequence coverage with a very confident expectation value of 5.31e-54. Given that the N- and C-termini are confirmed by many matching fragments, the mass discrepancy can be localized to a region in the middle of the sequence. This mass discrepancy is actually rather large (>1000 Da) and suggests that this could be due to a removal of a stretch of amino acids.

**FIGURE 2. Sequence Gazer results for an unexpected form of alpha casein. There is substantial sequence coverage from both termini and thus the modification to the sequence is somewhere in the middle. Due to the size of the mass discrepancy, it is likely that the modification is a stretch of missing sequence, perhaps due to a splice variant.**




Alpha-S2, beta- and kappa-casein were also determined to be highly heterogeneous with 5, 25 and 7 forms detected, respectively. In contrast, only 3 forms were detected for beta-lactoglobulin, 2 of which were intact, as well as 3 forms of GLCM1. Two forms of intact alpha-lactabumin and some truncated forms were identified when the whey proteins were reduced and alkylated. Figure 3 is the annotated extracted chromatogram of one of the whey casein mixture elution profile.

**FIGURE 3. Annotated extracted chromatogram for one of the whey:casein protein mixture datasets. There are multiple abundant chromatographic peaks for each of protein, indicating extensive heterogeneity for each of these proteins.**



### Protein Quantitation Using SIEVE 2.1

For relative protein quantitation, the extracted chromatogram of different whey to casein protein ratio were analyzed by SIEVE 2.1. Figure 4 was the trend intensities of whey protein, α-lactalbumin and casein protein, α-S1-casein. It reproducibly showed clear relation that the α-lactalbumin percentage decreased while the casein concentration increased . Since the proteins are so heterogeneous, using one protein will underestimate the real protein concentration. We sum the peak intensities from whole whey proteins and casein proteins. The whey to casein protein response ratios were calculated using the sum intensities of whey and casein proteins. Surprisingly, it seems yield the linear trend response with both injections (Figure 5). This leads us to the possibility of protein quantitation with intact protein approach. For more accurate protein quantitation, the response factor of each pure protein should be applied.

**FIGURE 4. Different ratio whey to casein protein mixture trend intensities analysis by SIEVE 2.1. Whey to casein ratio: Blue (2 to 1), Red (1 to 1), Green (1 to 2) Yellow (1 to 3)**

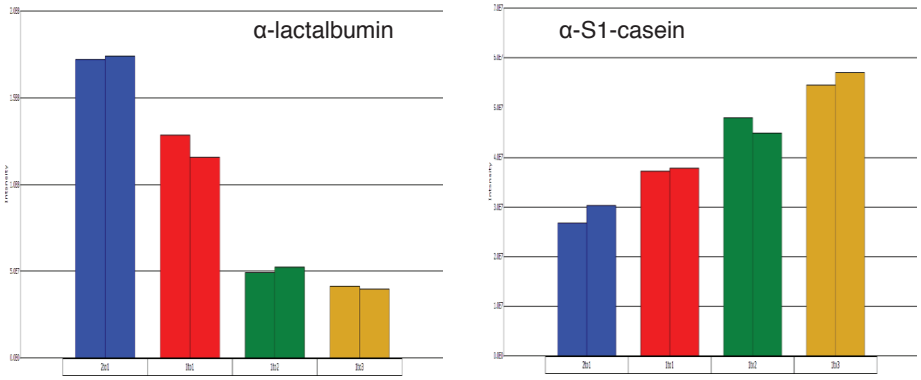
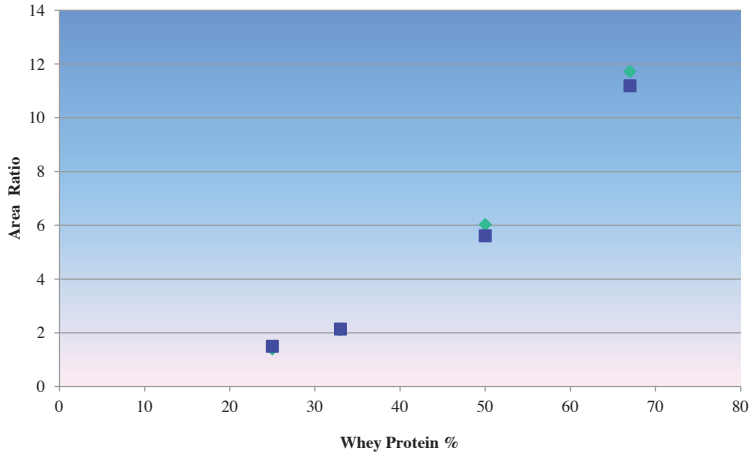



**FIGURE 5. Total whey protein to total casein protein response intensity ratios verses percentage of whey protein in protein solutions.**




## Conclusions

- The casein proteins are surprisingly complex, with numerous truncated forms of the expected proteins
- Alpha-s1-casein is the most complex of the proteins in the mixture, with 27 different identified forms including two different sequence variants.
- The different whey:protein ratios were calculated by using the extracted ion chromatogram.

**www.thermoscientific.com**




*Thermo Fisher Scientific,*
*San Jose, CA USA*
*is ISO 9001:2008 Certified.*

**Africa-Other** +27 11 570 1840
**Australia** +61 3 9757 4300
**Austria** +43 1 333 50 34 0
**Belgium** +32 53 73 42 41
**Canada** +1 800 530 8447
**China** +86 10 8419 3588
**Denmark** +45 70 23 62 60
**Europe-Other** +43 1 333 50 34 0
**Finland/Norway/Sweden**
+46 8 556 468 00
**France** +33 1 60 92 48 00
**Germany** +49 6103 408 1014
**India** +91 22 6742 9434
**Italy** +39 02 950 591
**Japan** +81 45 453 9100
**Latin America** +1 561 688 8700
**Middle East** +43 1 333 50 34 0
**Netherlands** +31 76 579 55 55
**New Zealand** +64 9 980 6700
**Russia/CIS** +43 1 333 50 34 0
**South Africa** +27 11 570 1840
**Spain** +34 914 845 965
**Switzerland** +41 61 716 77 00
**UK** +44 1442 233555
**USA** +1 800 532 4752

Thermo SCIENTIFIC

Part of Thermo Fisher Scientific

ASMS13_W348_TZhang_E 07/13S