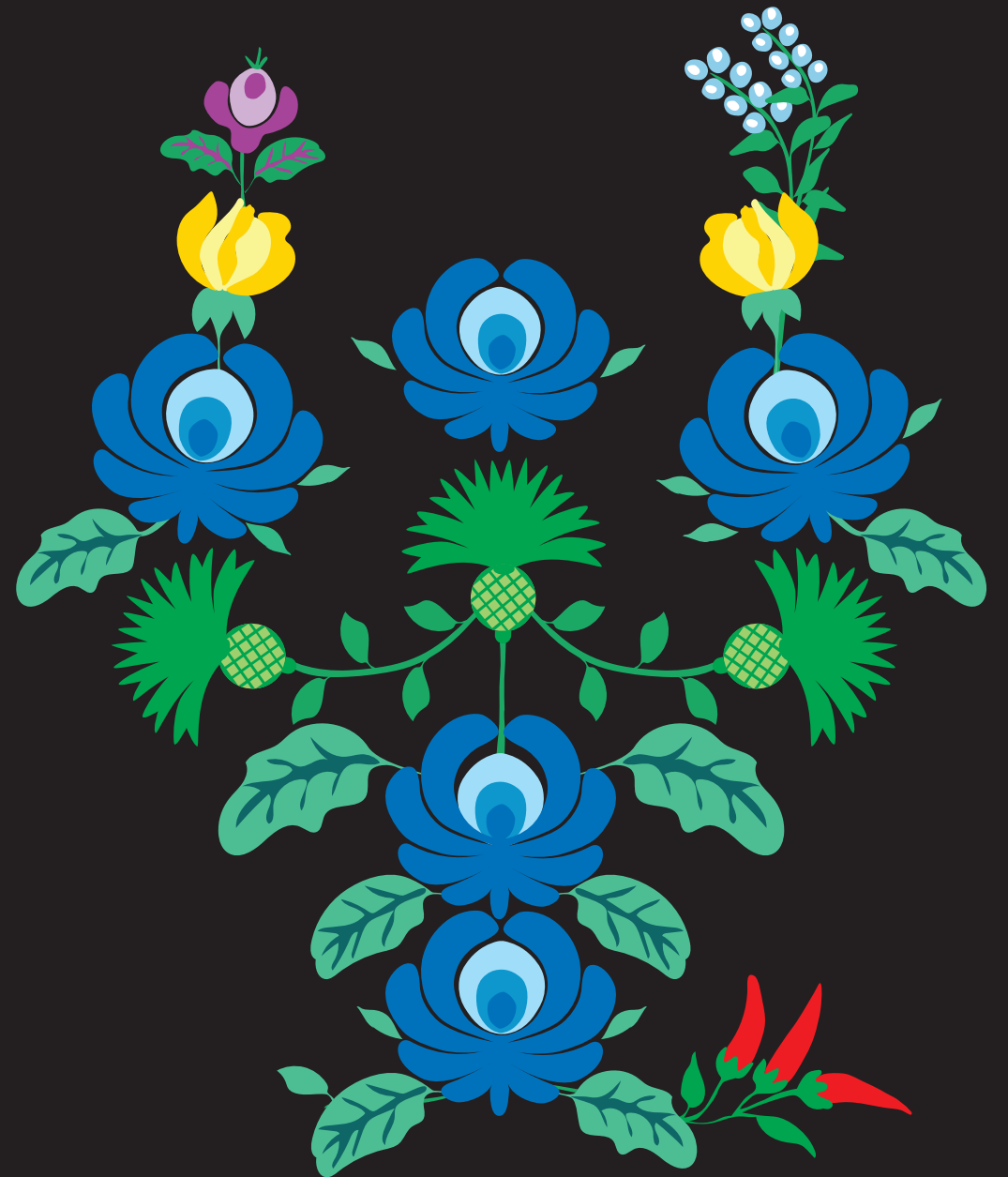


UNRAVELLING ELUSIVE MYSTERIES OF BOVINE MILK PROTEINS BY MASS SPECTROMETRY



Inge Gazi

UNRAVELLING ELUSIVE MYSTERIES OF BOVINE MILK PROTEINS BY MASS SPECTROMETRY

INGE GAZI • 2023



Unravelling elusive mysteries of bovine milk proteins by mass spectrometry

From colostrum to mature milk, during processing and
throughout storage

Inge Gazi

ISBN: 978-90-393-7572-3

DOI: <https://doi.org/10.33540/1904>

© Copyright 2023 Inge Gazi.

Cover design by Inês Vilalva | <https://inesvilalva.com/>.

Printed by Ridderprint | <https://www.ridderprint.nl/>.

The research presented in this thesis was performed in the Biomolecular Mass Spectrometry and Proteomics Group, Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, The Netherlands.

Unravelling elusive mysteries of bovine milk proteins by mass spectrometry

De mysteries van koemelkeiwitten ontcijferd met massaspectrometrie
(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

maandag 11 september 2023 des middags te 2.15 uur

door

Inge Gazi

geboren op 13 februari 1987
te Iași, Roemenië

Promotoren:

Prof. dr. A.J.R. Heck
Prof. dr. T. Huppertz

Copromotor:

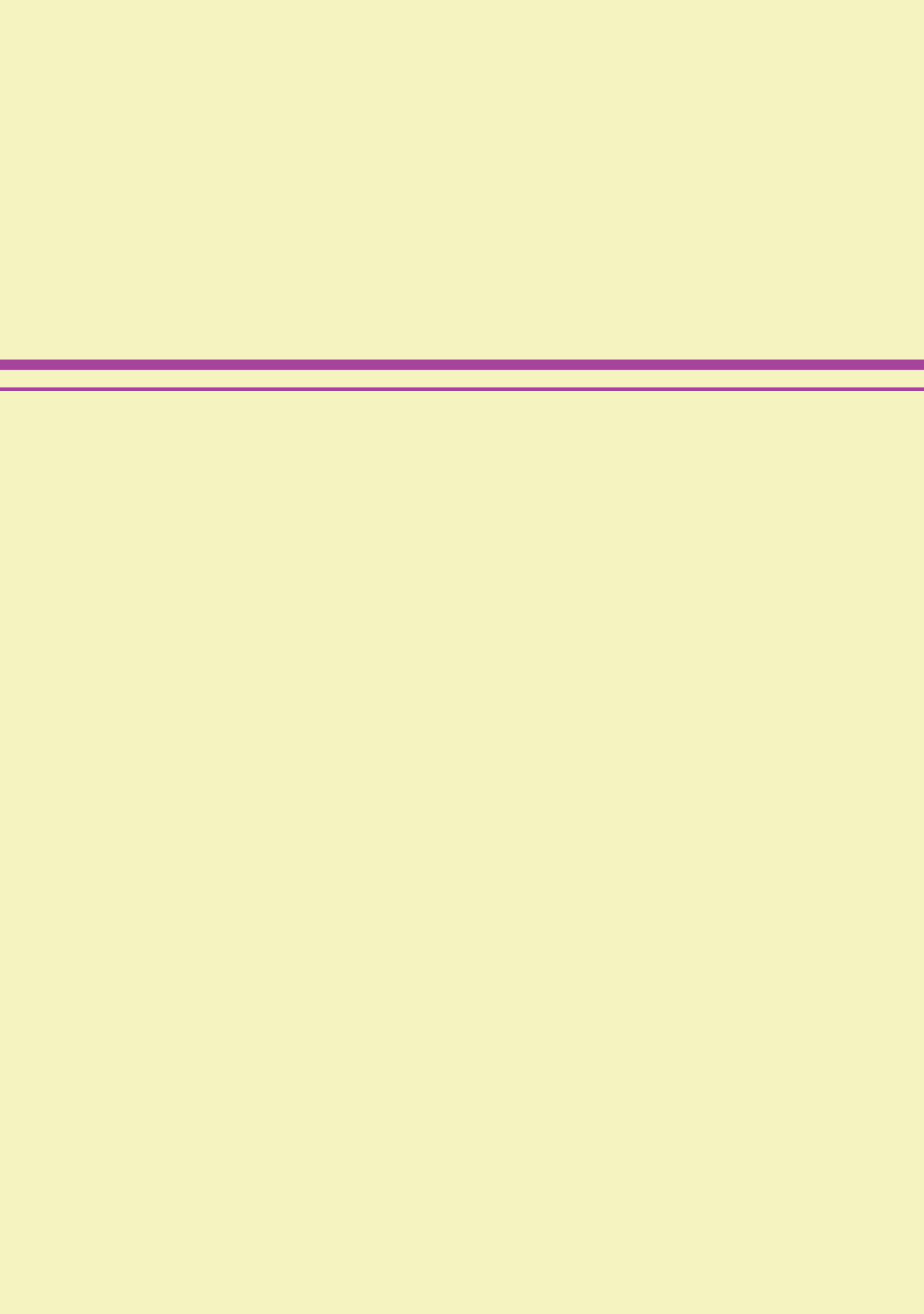
Dr. K.R. Reiding

Beoordelingscommissie:

Prof. dr. J. Garssen
Prof. dr. K. Hettinga
Prof. dr. G. Vidarsson
Prof. dr. M.H.M. Wauben
Prof. dr. M. Wuhler

Table of Contents

Chapter 1. Introduction	7
Chapter 2. Identifying glycation hot-spots in bovine milk proteins during production and storage of skim milk powder	55
Chapter 3. Key changes in bovine milk immunoglobulin G during lactation: NeuAc sialylation is a hallmark of colostrum immunoglobulin G <i>N</i> -glycosylation	93
Chapter 4. The <i>N</i> -glycoproteome of bovine mammary secretions from colostrum to mature milk, with detailed insights into the <i>N</i> -glycosylation of α -lactalbumin	137
Chapter 5. Summary and Outlook	167
About the author	185
Acknowledgements	193



CHAPTER 1

Introduction



Chapter 1. Introduction

Part of this chapter has been published as:

Gazi, I., Johansen, L.B., Huppertz, T., 2022. Heterogeneity, fractionation, and isolation. In: McSweeney, P.L.H., McNamara, J.P. (Eds.), *Encyclopedia of Dairy Sciences*, 3rd Ed., Academic Press, Oxford, pp. 881–893.

<https://doi.org/10.1016/B978-0-12-818766-1.00278-6>

Milk, in the form of human milk or cow milk-based infant formula, is the first source of nutrition of the human neonate. With the domestication of cattle starting approximately 10,500 years ago, the earliest evidence for the consumption of milk and dairy products dates 9000 years back (Evershed *et al.*, 2008, Roffet-Salque *et al.*, 2018). This marks 9 millennia during which milk and dairy products have constituted a major food group in the human diet, also beyond infancy. The nutritional and functional properties of milk are the main reason why it has maintained a core position in a balanced diet until the present day (Muehlhoff *et al.*, 2013). Dairy food consumption in *e.g.*, The Netherlands, accounts for 30% of the mean total food consumption of young children in g/d, and 10-15% for the older age groups, providing on average 15% of total dietary energy, and contributing macronutrients such as carbohydrates, fats, high-quality protein (15% dietary intake, DI), and micronutrients including calcium (58% DI), phosphorus (32% DI), vitamins B2 and B12 (39% DI), vitamin A (33% DI), zinc (23% DI) and iodine (27%) (Van Rossum *et al.*, 2020).

The milk of all mammalian species studied in detail to date contains a wide variety of proteins. This is not surprising given that milk, as the sole source of nutrition for the mammalian neonate, should provide the required levels of nitrogen and essential amino acids. In addition to being a highly digestible source of nitrogen and essential amino acids, milk proteins also have various other key biological functions. They act, for instance, as a carrier of salts and are essential in the development of the immune system of the neonate and have protective anti-microbial and anti-inflammatory activity. Alongside dietary proteins from other origin, milk proteins play a key role in supporting growth, development, maintenance, recovery, as well as other biological processes and functions (Horstman and Huppertz, 2022).

The proteins of bovine milk

The genome of any given organism contains the blueprint to the baseline amino acid sequence of every protein that it is able to produce. The complete sequencing of the human genome (Venter *et al.*, 2001) also brought on a whole new era of protein research, and specifically proteomics. Shortly after the complete sequencing of the human genome followed also the complete sequencing of other mammalian genomes, including the bovine genome (Elsik *et al.*, 2009). The bovine genome was estimated to contain 22,000 protein-encoding genes, of which approximately 16,000 were found orthologous with human genes (Elsik *et al.*, 2009). The diversity of proteins, however, does not stop at the number of protein-encoding genes. Genetic polymorphism and alternative splicing events can both lead to the production of proteins with differences in the amino acid sequences by

the same gene. Further increases in protein diversity can result from post-translational modifications. The combination of these factors raises the possible number of proteins far above the theoretically 22,000 encoded proteins. The primary sequence of the protein, the genotype and the post-translational modifications can all have a direct impact on the function of the protein, further highlighting the importance of not only knowing the genomic blueprint, but of also studying the proteins themselves.

Some of the earliest scientific publications describing the discovery of milk proteins, and particularly caseins, include the works of Berzelius (1814) and Braconnot (1830). Both publications describe acidification of milk resulting in a curd that can be separated from a clear solution, as well as the investigation of physical properties of this acid curd. Berzelius refers to this curd as “cheese” (“Käse”) while Braconnot calls it “cheese material” (“Käsestoff”), illustrating the origin of the name “casein”. This fractionation is still used to define milk proteins, wherein a pH adjustment to 4.6 of milk separates the milk into a whey protein fraction (liquid) and the casein fraction (precipitate). The ratio at which caseins and whey proteins occur varies between species, but the milk of most species is casein-dominant. However, human milk is a notable exception here, being whey-protein dominant. In the milk of the major dairying species (*i.e.*, bovine, buffalo, sheep and goat) caseins and whey proteins occur at a mass ratio of ~80:20. In addition, a separate class of milk fat globule membrane proteins may be distinguished, although these are present at much lower concentrations compared to caseins and whey proteins. The casein and whey protein classes consist of a number of gene products, for which several genetic variants can exist, as well as various degrees of post-translational modification, primarily in the form of phosphorylation and *N*- and *O*-glycosylation.

Structurally, caseins and whey proteins also differ notably, with the whey proteins characterized as proteins with a defined secondary, tertiary, and quaternary structure which is often stabilized by a number of disulphide bonds. By contrast, owing at least in part to the high amounts of Pro residues, caseins have little or no defined secondary and tertiary structure. Differences are also observed in the oligomeric state of the milk proteins, with whey proteins occurring primarily in monomeric and dimeric form, whereas caseins occur predominantly in the form of casein micelles, which are colloidal particles that can consist of tens of thousands of casein molecules, assembled together with calcium and inorganic phosphate. The very different properties of caseins and whey proteins have also led to their exploration and utilization in the form of isolates or enriched fractions, which are widely applied in the dairy sector.

In this chapter, we will describe the different milk protein fractions and their properties. Within this scope, we will focus on the major proteins present in the milk of *Bos taurus*, *i.e.*, the caseins α_{s1} -, α_{s2} -, β - and κ -casein, as well as the whey proteins α -lactalbumin, β -lactoglobulin, bovine serum albumin, lactoferrin, and last but not least, the immunoglobulins.

Classification and nomenclature of bovine milk proteins

As outlined above, the milk proteins are a collection of different gene products that show further heterogeneity based on the genetic variants of each protein, and their different types and degrees of post-translational modification. In this chapter, we will use the nomenclature originally proposed by the American Dairy Science Committee on the Nomenclature, Classification and Methodology of Milk Proteins with respect to indicating genetic variants and phosphorylation. For instance, α_{s1} -casein B-8P refers to genetic variant B of α_{s1} -casein containing 8 phosphorylated amino acid residues and β -casein A2-5P refers to the A2 variant of β -casein containing 5 phosphorylated amino acid residues. For each casein and whey protein for which genetic variation is discussed, we will do this based on a so-called reference protein, as defined in the UniProtKB/SwissProt database (<https://www.uniprot.org/>).

Caseins

The primary structure and gene sequence of the four caseins (α_{s1} -, α_{s2} -, β - and κ -casein) have been well established. In Table 1.1 some properties of caseins in bovine milk are summarized based on the reference proteoform of each casein. The four caseins are all subject to different degrees of phosphorylation whereas κ -casein is also subject to *O*-glycosylation; furthermore, α_{s2} -casein can form inter- and intra-molecular disulphide bonds and κ -casein forms intermolecular disulphide bonds.

α_{s1} -Casein

The primary structure of the reference proteoform for α_{s1} -casein (α_{s1} -CN B-8P) is shown in Figure 1.1. In addition to the 8P-variant, 9P-variants for α_{s1} -casein are also commonly found, with additional phosphorylation at Ser41 (Huppertz, 2013). Previously, α_{s1} -CN B-9P was believed to be a separate gene product, rather than a phosphoform of α_{s1} -casein, and was thus formerly known as α_{s0} -CN. In a study of Montbéliarde cows by Fang *et al.* (2016), they found that 77% of the α_{s1} -CN B was 8P and 23% was 9P. A 7P-variant of α_{s1} -casein has also been reported, but this phosphoform appears to be rare and to occur only at low amounts (Nilsson *et al.*, 2020). At least 10 different genetic variants of α_{s1} -casein have been described in bovine milk; however, variant B is most commonly-encountered.

Table 1.1 – Properties of the caseins in bovine milk

Protein	α_{s1} -casein	α_{s2} -casein	β -casein	κ -casein
Gene name	CSN1S1	CSN1S2	CSN2	CSN3
SwissProt accession	P02662	P02663	P02666	P02668
Conc. in milk (g/L) ^a	12-15	3-4	9-11	2-4
Phosphorylation range ^b	7-9	9-15	4-5	0-3
Number of glycans ^c	-	-	-	0-6
Reference proteoform	α_{s1} -CN B-8P	α_{s2} -CN A-11P	β -CN A2-5P	κ -CN A-1P
Number of amino acids	199	207	209	169
Number of Pro residues	17	10	35	20
Monoisotopic mass (Da) ^d	23600.21	25212.98	23968.15	19025.53
Average mass (Da) ^d	23614.43	25228.03	23982.89	19037.14
pI ^e	4.43	4.95	4.66	5.62

^aGoulding et al. (2020)

^bFang et al. (2017); Farrell Jr et al. (2004); Leonil et al. (1995); Miranda et al. (2020)

^cHolland et al. (2005); Sunds et al. (2019)

^dProtein masses calculated based on amino acid sequence, individual amino acid residue masses from http://www.matrixscience.com/help/aa_help.html, and masses of post-translational modifications from <https://www.sigmaaldrich.com/life-science/proteomics/post-translational-analysis/phosphorylation/mass-changes.html>. For the mass calculation it was assumed that all cysteine residues are reduced.

^eThe pI of the proteins was calculated using the ProtPi protein tool that also takes into account the presence of post-translational modification: <https://www.protpi.ch/Calculator/ProteinTool>.

The monoisotopic and average masses are essential for correct identification of proteoforms by mass spectrometry (MS). The isoelectric point (pI) is an important

	Met	Lys	Leu	Leu	Ile	Leu	Thr	Cys	Leu	Val	1	Ala	Val	Ala	Leu	Ala	Arg	Pro	Lys	His	Pro	5
6	Ile	Lys	His	Gln	Gly	Leu	Pro	Gln	Glu	Val	15	Leu	Asn	Glu	Asn	Leu	Leu	Arg	Phe	Phe	Val	25
26	Ala	Pro	Phe	Pro	Glu	Val	Phe	Gly	Lys	Glu	35	Lys	Val	Asn	Glu	Leu	Ser	Lys	Asp	Ile	Gly	45
46	pSer	Glu	pSer	Thr	Glu	Asp	Gln	Ala	Met	Glu	55	Asp	Ile	Lys	Gln	Met	Glu	Ala	Glu	pSer	Ile	65
66	pSer	pSer	pSer	Glu	Glu	Ile	Val	Pro	Asn	pSer	75	Val	Glu	Gln	Lys	His	Ile	Gln	Lys	Glu	Asp	85
86	Val	Pro	Ser	Glu	Arg	Tyr	Leu	Gly	Tyr	Leu	95	Glu	Gln	Leu	Leu	Arg	Leu	Lys	Lys	Tyr	Lys	105
106	Val	Pro	Gln	Leu	Glu	Ile	Val	Pro	Asn	pSer	115	Ala	Glu	Glu	Arg	Leu	His	Ser	Met	Lys	Glu	125
126	Gly	Ile	His	Ala	Gln	Gln	Lys	Glu	Pro	Met	135	Ile	Gly	Val	Asn	Gln	Glu	Leu	Ala	Tyr	Phe	145
146	Tyr	Pro	Glu	Leu	Phe	Arg	Gln	Phe	Tyr	Gln	155	Leu	Asp	Ala	Tyr	Pro	Ser	Gly	Ala	Trp	Tyr	165
166	Tyr	Val	Pro	Leu	Gly	Thr	Gln	Tyr	Thr	Asp	175	Ala	Pro	Ser	Phe	Ser	Asp	Ile	Pro	Asn	Pro	185
186	Ile	Gly	Ser	Glu	Asn	Ser	Glu	Lys	Thr	Thr	195	Met	Pro	Leu	Trp							199

Figure 1.1 – Primary sequence of α_{s1} -CN B-8P (Grosclaude et al., 1973, Mercier et al., 1971), SwissProt accession P02662. The following residues are highlighted on the sequence: Xxx = amino acids of the signal peptide; pSer = phosphorylated serine. Numbering starts at the N-terminus of the mature protein.

factor influencing physical-chemical properties. Differences in post-translational modifications such as the degree of phosphorylation and the degree and type of glycosylation change the pI of the molecule. For example, the pI values of α_{s1} -CN B-8P and B-9P are 4.43 and 4.37, respectively.

The distribution of charge and hydrophobicity on the α_{s1} -CN B-8P amino acid sequence is shown in Figure 1.2. At natural milk pH, α_{s1} -casein carries a net-negative charge, with areas of high net-negative charge found around the two centres of phosphorylation located between residues 31 and 80. In contrast, the 30 N-terminal amino acids carry a net-positive charge, while region 81-179 is balanced in relation to positively- and negatively-charged residues. The C-terminus (180-199) carries a net-negative charge. Some patches of hydrophobicity in the α_{s1} -casein sequence are found for residues 20-35 and 140-170. Like all other caseins, α_{s1} -casein shows self-association behaviour, from dimers to tetramers, hexamers, octamers, *etc.* Monomers are found at only very small quantities, unless ionic strength is very low (<3 mM).

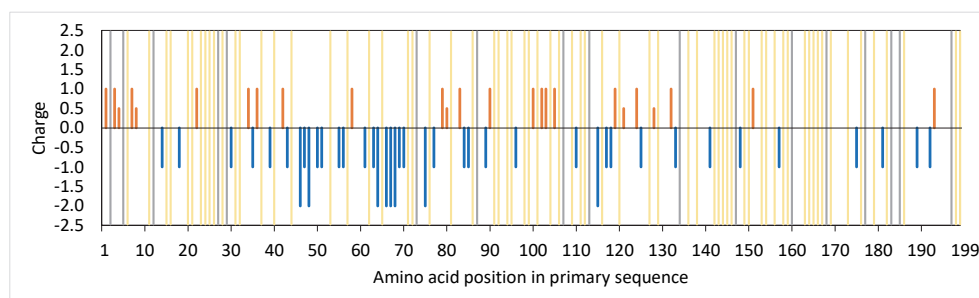


Figure 1.2 – Distribution of charged and hydrophobic residues on the amino acid chain of α_{s1} -CN B-8P. Negatively-charged residues (Asp = -1; Glu = -1; pSer = -2); positively-charged residues (Arg = 1; His = 0.5; Lys = 1); proline residues; hydrophobic residues (Ala, Ile, Leu, Phe, Val, Trp, Tyr).

α_{s1} -Casein, like all other caseins, is also prone to enzymatic hydrolysis. In this chapter, 3 enzymes will be considered in relation to casein hydrolysis, *i.e.*, the indigenous milk proteases plasmin and cathepsin D, and chymosin, the protease used in cheese making. Focus will be on the major protease cleavage sites and on the most abundant or functionally-relevant proteolysis products. Plasmin is a serine protease highly specific for Lys-Xxx or Arg-Xxx peptide bonds, whereas cathepsin D and chymosin are both aspartic acid proteases that preferentially cleave peptide bonds between two hydrophobic residues (*i.e.*, Ala, Leu, Tyr, Val, Ile, Phe, Trp). Cathepsin D cleaves α_{s1} -casein at residue Phe23-Phe24 and Phe24-Val25; fragment 24-199 is also known as α_{s1} -I. Chymosin can also hydrolyse the Phe23-Phe24-bond in α_{s1} -casein. A total of 14 cleavage sites for plasmin were identified in α_{s1} -casein.

α_{s2} -Casein

The primary structure for the reference proteoform of α_{s2} -casein (α_{s2} -CN A-11P) is shown in Figure 1.3. As for the other caseins, at least 5 different genetic variants of α_{s2} -casein have been described to exist. The focus in this chapter will be on variant A, which is also the most prevalent one. Of the caseins, α_{s2} -casein is the most highly phosphorylated one. Next to the 11P variant of α_{s2} -casein, various other phosphoforms of the protein also exist, with levels of phosphorylation from 9 to 15 having been reported for α_{s2} -casein (Fang *et al.*, 2017, Nilsson *et al.*, 2020). Consequently, α_{s2} -casein is also the casein that exhibits the highest variability in the degree of phosphorylation. α_{s2} -Casein contains two Cys-residues at positions 36 and 40, which can occur in either intra- or intermolecular disulphide bonds in homodimers. Figure 1.4 shows the distribution of charge and hydrophobicity on the amino acid sequence of α_{s2} -CN B-11P. The primary structure of α_{s2} -casein shows two areas of high sequence homology, *i.e.*, residues 42-122 and 124-207. Residues 1-40 and 41-80 both contain a centre of phosphorylation and high net-negative charge, whereas the region of 81-125 contains small numbers of positively-charged residues and a high proportion of hydrophobic residues. The region of 126-170 contains a third phosphorylation cluster and is net-negatively-charged. The C-terminal region of 171-207 consists mostly of positively-charged and hydrophobic residues. α_{s2} -Casein also shows self-association behaviour, which is most pronounced at 20 °C and an ionic strength of 0.2-0.3 M.

	Met	Lys	Phe	Phe	Ile	Phe	Thr	Cys	Leu	Leu	1	Ala	Val	Ala	Leu	Ala	Lys	Asn	Thr	Met	Glu	5
6	His	Val	pSer	pSer	pSer	Glu	Glu	Ser	Ile	Ile	15	pSer	Gln	Glu	Thr	Tyr	Lys	Gln	Glu	Lys	Asn	25
26	Met	Ala	Ile	Asn	Pro	Ser	Lys	Glu	Asn	Leu	35	Cys	Ser	Thr	Phe	Cys	Lys	Glu	Val	Val	Arg	45
46	Asn	Ala	Asn	Glu	Glu	Glu	Tyr	Ser	Ile	Gly	55	pSer	pSer	pSer	Glu	Glu	pSer	Ala	Glu	Val	Ala	65
66	Thr	Glu	Glu	Val	Lys	Ile	Thr	Val	Asp	Asp	75	Lys	His	Tyr	Gln	Lys	Ala	Leu	Asn	Glu	Ile	85
86	Asn	Gln	Phe	Tyr	Gln	Lys	Phe	Pro	Gln	Tyr	95	Leu	Gln	Tyr	Leu	Tyr	Gln	Gly	Pro	Ile	Val	105
106	Leu	Asn	Pro	Trp	Asp	Gln	Val	Lys	Arg	Asn	115	Ala	Val	Pro	Ile	Thr	Pro	Thr	Leu	Asn	Arg	125
126	Glu	Gln	Leu	pSer	Thr	pSer	Glu	Glu	Asn	Ser	135	Lys	Lys	Thr	Val	Asp	Met	Glu	pSer	Thr	Glu	145
146	Val	Phe	Thr	Lys	Lys	Thr	Lys	Leu	Thr	Glu	155	Glu	Glu	Lys	Asn	Arg	Leu	Asn	Phe	Leu	Lys	165
166	Lys	Ile	Ser	Gln	Arg	Tyr	Gln	Lys	Phe	Ala	175	Leu	Pro	Gln	Tyr	Leu	Lys	Thr	Val	Tyr	Gln	185
186	His	Gln	Lys	Ala	Met	Lys	Pro	Trp	Ile	Gln	195	Pro	Lys	Thr	Lys	Val	Ile	Pro	Tyr	Val	Arg	205
206	Tyr	Leu									207											

Figure 1.3 – Primary sequence of α_{s2} -CN A-11P (Brignon *et al.*, 1977), SwissProt accession P02663. The following residues are highlighted on the sequence: Xxx = amino acids of the signal peptide; Cys = cysteine; pSer = phosphorylated serine. Numbering starts at the N-terminus of the mature protein.

α_{s2} -Casein is also prone to enzymatic digestion. Plasmin-derived peptides in milk systems were found originating from the N-terminus, *i.e.*, Lys1-Lys21 and Lys1-Lys24, as well as from C-terminus of the protein, *i.e.*, Ala189-Leu207 and The198-Leu207 (Rauh *et al.*, 2014).

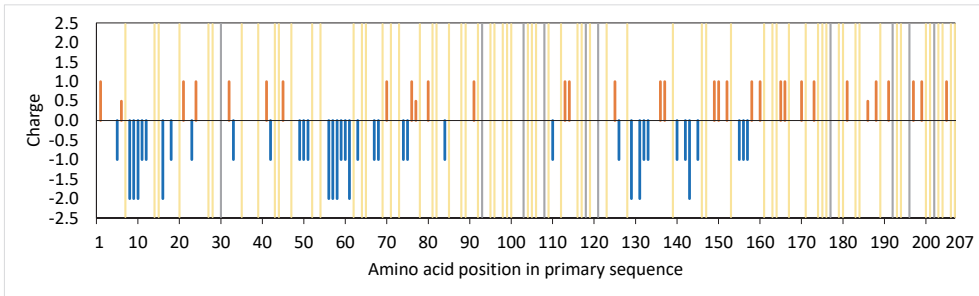


Figure 1.4 – Distribution of charged and hydrophobic residues on the amino acid chain of α_2 -CN A-11P. █ Negatively-charged residues (Asp = -1; Glu = -1; pSer = -2); █ positively-charged residues (Arg = 1; His = 0.5; Lys = 1); █ proline residues; █ hydrophobic residues (Ala, Ile, Leu, Phe, Val, Trp, Tyr).

β -Casein

The reference proteoform for β -casein is β -CN A2-5P, the sequence of which is shown in Figure 1.5. The 5P form is the main phosphoform for β -casein; the 4P is typically also present in milk, but at a considerably lower concentration. As is also the case for the other caseins, several genetic variants have been described for β -casein. The most commonly occurring ones are shown in Table 1.2.

	Met	Lys	Val	Leu	Ile	Leu	Ala	Cys	Leu	Val	1	Ala	Leu	Ala	Leu	Ala	Arg	Glu	Leu	Glu	Glu	5
6	Leu	Asn	Val	Pro	Gly	Glu	Ile	Val	Glu	pSer	15	Leu	pSer	pSer	pSer	Glu	Glu	Ser	Ile	Thr	Arg	25
26	Ile	Asn	Lys	Lys	Ile	Glu	Lys	Phe	Gln	pSer	35	Glu	Glu	Gln	Gln	Gln	Thr	Glu	Asp	Glu	Leu	45
46	Gln	Asp	Lys	Ile	His	Pro	Phe	Ala	Gln	Thr	55	Gln	Ser	Leu	Val	Tyr	Pro	Phe	Pro	Gly	Pro	65
66	Ile	Pro	Asn	Ser	Leu	Pro	Gln	Asn	Ile	Pro	75	Pro	Leu	Thr	Gln	Thr	Pro	Val	Val	Val	Pro	85
86	Pro	Phe	Leu	Gln	Pro	Glu	Val	Met	Gly	Val	95	Ser	Lys	Val	Lys	Glu	Ala	Met	Ala	Pro	Lys	105
106	His	Lys	Glu	Met	Pro	Phe	Pro	Lys	Tyr	Pro	115	Val	Glu	Pro	Phe	Thr	Glu	Ser	Gln	Ser	Leu	125
126	Thr	Leu	Thr	Asp	Val	Glu	Asn	Leu	His	Leu	135	Pro	Leu	Pro	Leu	Leu	Gln	Ser	Trp	Met	His	145
146	Gln	Pro	His	Gln	Pro	Leu	Pro	Pro	Thr	Val	155	Met	Phe	Pro	Pro	Gln	Ser	Val	Leu	Ser	Leu	165
166	Ser	Gln	Ser	Lys	Val	Leu	Pro	Val	Pro	Gln	175	Lys	Ala	Val	Pro	Tyr	Pro	Gln	Arg	Asp	Met	185
186	Pro	Ile	Gln	Ala	Phe	Leu	Leu	Tyr	Gln	Glu	195	Pro	Val	Leu	Gly	Pro	Val	Arg	Gly	Pro	Phe	205
206	Pro	Ile	Ile	Val							209											

Figure 1.5 – Primary sequence of β -CN A2-5P (Dumas et al., 1972, Grosclaude et al., 1973), SwissProt accession P02666. The following residues are highlighted on the sequence: Xxx = amino acids of the signal peptide; pSer = phosphorylated serine. Numbering starts at the N-terminus of the mature protein.

The charge and hydrophobicity distribution on the sequence of β -CN A2-5P is shown in Figure 1.6. The primary structure of β -casein shows two distinct regions, *i.e.*, a strongly negatively-charged N-terminus (residues 1-40) which contains the centre of phosphorylation and a C-terminal fragment (residues 41-209) of moderate charge and notable hydrophobicity. As a result of this amphipathic character, β -casein is prone to micellisation which is driven by hydrophobic interactions and thus strongly increases with increasing temperature. Of the caseins, β -casein is the one that is most prone to plasmin-induced hydrolysis. The major plasmin cleavage sites are Lys28-Lys29, Lys105-His106 and Lys107- Glu108,

Table 1.2 – Differences in the amino acid sequences of the most common genetic variants of β -casein found in the milk of *Bos taurus* compared to the reference proteoform β -CN A2-5P

Genetic variant	Amino acid position in mature protein					Proteoform	Mass (Da)		
	67	93	106	122	152		Monoisotopic	Average	pI
A2	Pro	Met	His	Ser	Pro	β -CN A2-5P	23968.15	23982.89	4.66
A1	His					β -CN A1-5P	24008.16	24022.91	4.73
A3			Gln			β -CN A3-5P	23959.15	23973.88	4.58
B	His			Arg		β -CN B-5P	24077.23	24092.02	4.82
F	His				Leu	β -CN F-5P	24024.19	24038.96	4.73
I		Leu				β -CN I-5P	23950.19	23964.85	4.66

leading to the formation of so-called γ -caseins (residues 29-209, 106-209 and 108-209) and proteose peptones (residues 1-28, 1-105, 1-107, 29-105 and 29-107). β -Casein can also be cleaved by both cathepsin D (Hurley *et al.*, 2000) and chymosin. The most common sites where β -casein is cleaved by the aforementioned aspartic proteases are Leu58-Val59, Leu127-Thr128, Trp143-Met144, Leu165-Ser166 and Leu192-Tyr193.

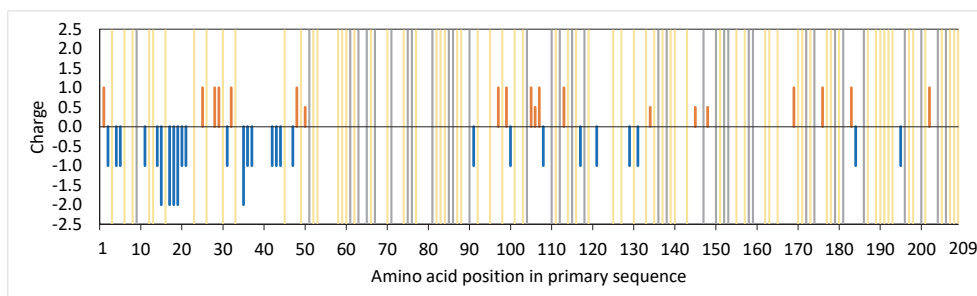


Figure 1.6 – Distribution of charged and hydrophobic residues on the amino acid chain of β -CN A2-5P. Negatively-charged residues (Asp = -1; Glu = -1; pSer = -2); positively-charged residues (Arg = 1; His = 0.5; Lys = 1); proline residues; hydrophobic residues (Ala, Ile, Leu, Phe, Val, Trp, Tyr).

κ -Casein

Among the caseins, κ -casein takes a unique position in that it is the least frequently and least abundantly phosphorylated, but also the only bovine casein thus far known to be glycosylated. In addition, it is found only on the surface, and not in the core of the casein micelles and it is crucial in providing the micelles with steric stabilization. The reference proteoform for κ -casein is κ -CN A-1P and its amino acid

sequence is shown in Figure 1.7. The C-terminal glutamine is cyclized to a pyroglutamic acid, marked in the figure as Pyr.

	Met	Met	Lys	Ser	Phe	Phe	Leu	Val	Val	Thr		Ile	Leu	Ala	Leu	Thr	Leu	Pro	Phe	Leu	Gly	
1	Ala	Pyr	Glu	Gln	Asn	Gln	Glu	Gln	Pro	Ile	9	Arg	Cys	Glu	Lys	Asp	Glu	Arg	Phe	Phe	Ser	19
21	Asp	Lys	Ile	Ala	Lys	Tyr	Ile	Pro	Ile	Gln	29	Tyr	Val	Leu	Ser	Arg	Tyr	Pro	Ser	Tyr	Gly	39
41	Leu	Asn	Tyr	Tyr	Gln	Gln	Lys	Pro	Val	Ala	49	Leu	Ile	Asn	Asn	Gln	Phe	Leu	Pro	Tyr	Pro	59
61	Tyr	Tyr	Ala	Lys	Pro	Ala	Ala	Val	Arg	Ser	69	Pro	Ala	Gln	Ile	Leu	Gln	Trp	Gln	Val	Leu	79
81	Ser	Asn	Thr	Val	Pro	Ala	Lys	Ser	Cys	Gln	89	Ala	Gln	Pro	Thr	Thr	Met	Ala	Arg	His	Pro	99
101	His	Pro	His	Leu	Ser	Phe	Met	Ala	Ile	Pro	109	Pro	Lys	Lys	Asn	Gln	Asp	Lys	Thr	Glu	Ile	119
121	Pro	Thr	Ile	Asn	Thr	Ile	Ala	Ser	Gly	Glu	129	Pro	Thr	Ser	Thr	Pro	Thr	Thr	Glu	Ala	Val	139
141	Glu	Ser	Thr	Val	Ala	Thr	Leu	Glu	Asp	pSer	149	Pro	Glu	Val	Ile	Glu	Ser	Pro	Pro	Glu	Ile	159
161	Asn	Thr	Val	Gln	Val	Thr	Ser	Thr	Ala	Val	169											

Figure 1.7 – Primary sequence of κ -CN A-1P (Grosclaude *et al.*, 1972, Mercier *et al.*, 1973), SwissProt accession P02668. The following residues are highlighted on the sequence: Xxx = amino acids of the signal peptide; Cys = cysteine; pSer = phosphorylated serine; Pyr = cyclization of protein N-terminal glutamine to pyroglutamic acid. Numbering starts at the N-terminus of the mature protein.

The charge and hydrophobicity distribution on the amino acid sequence of κ -CN A-1P are shown in Figure 1.8. κ -Casein possesses an interesting charge distribution, with the N-terminal para- κ -casein region (residues 1-105) carrying a net-positive charge, whereas the C-terminal caseinomacropeptide (CMP; residues 106-169) carries a net-negative charge. Next to the most commonly observed 1P variant for κ -casein, the 0P, 2P and 3P phosphoforms can also be found, but at considerably lower frequencies (Nilsson *et al.*, 2020). All known phosphorylation sites are found in the CMP region of the protein. Several genetic variants have been described for κ -casein as well, with the most frequently occurring ones shown in Table 1.3.

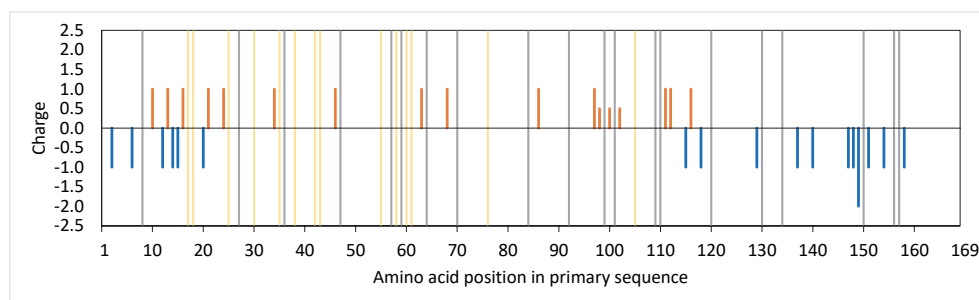


Figure 1.8 – Distribution of charged and hydrophobic residues on the amino acid chain of κ -CN A-1P. Negatively-charged residues (Asp = -1; Glu = -1; pSer = -2); positively-charged residues (Arg = 1; His = 0.5; Lys = 1); proline residues; hydrophobic residues (Ala, Ile, Leu, Phe, Val, Trp, Tyr).

κ -Casein contains two Cys residues (positions 11 and 88), which predominantly form intermolecular disulphide bonds. Dimers and trimers of κ -casein are most common, but higher-order disulphide-linked oligomers are also found.

Table 1.3 – Differences in the amino acid sequences of the most common genetic variants of κ -casein found in the milk of *Bos taurus* compared to the reference proteoform κ -CN A-1P













Genetic variant	Amino acid position in mature protein			Proteoform	Mass (Da)		pI
	136	148	155		Monoisotopic	Average	
A	Thr	Asp	Ser	κ -CN A-1P	19025.53	19037.14	5.62
B	Ile	Ala		κ -CN B-1P	18993.58	19005.18	5.91
E			Gly	κ -CN E-1P	18995.52	19007.11	5.62






Next to phosphorylation, κ -casein can contain a variable (0-6) number of *O*-glycans, all of which have been reported to occur in the CMP region of the protein. The most commonly-encountered glycoforms contain 1-3 glycans *per* protein, with abundance of the glycoforms decreasing with increasing number of glycans. Higher degrees of glycosylation have also been detected at low abundance in milk with Vreeman *et al.* (1986) observing up to 5 glycans and Holland *et al.* (2006) up to 6 glycans. All glycans occur on Thr residues *via O*-linked glycosylation. The following glycosylation sites have been identified: Thr121, Thr131, Thr133, Thr136 (not in variant B in light of the Thr136 \rightarrow Ile136 mutation), Thr142, Thr145 and Thr165 (Sunds *et al.*, 2019). When considering glycosylation of κ -casein quantitatively, two aspects can be distinguished, *i.e.*, the percentage of κ -casein molecules in a sample that is glycosylated, and the number of glycans attached to the κ -casein molecules. It is interesting to note that in all work published to date, neither the complete presence nor the complete absence of glycosylation of bovine κ -casein have been reported. The percentage of κ -casein molecules that are glycosylated has been reported to be as high as 60% for individual bovines, but such high values are rare (Bonfatti *et al.*, 2019). Further heterogeneity can arise from the type of glycans that can be found on κ -casein (Farrell Jr *et al.*, 2004, Saito and Itoh, 1992, Sunds *et al.*, 2019). An overview of reported glycans is shown in Table 1.4, along with the monoisotopic and average glycan residue masses.

As is the case of β -casein, κ -casein is also an amphipathic molecule prone to micellisation. Unlike β -casein, the N-terminal portion of κ -casein associates while the negatively-charged C-terminus provides steric stabilization. In addition, κ -casein micellisation is not temperature-dependent, indicating that it is not driven by hydrophobic interactions. Contrary to the other caseins, κ -casein is not prone to hydrolysis by plasmin or cathepsin D. It is, however, hydrolysed by chymosin during cheese-making at the Phe105-Met106 bond, leading to the release of CMP in

solution and the loss of steric stabilization of the casein micelle surface by the remaining para- κ -casein.

Table 1.4 – O-glycan residue structures found on glycosylated κ -casein

Neutral glycans				Acidic glycans			
	Glycan structure	Monoisotopic Δ mass (Da)	Average Δ mass (Da)		Glycan structure	Monoisotopic Δ mass (Da)	Average Δ mass (Da)
a		203.08	203.20	f		494.17	494.45
b		365.13	365.34	g		656.23	656.60
c		568.21	568.53	h		656.23	656.60
d		730.26	730.67	i		947.32	947.85
e		876.32	876.82	j		859.31	859.79
				k		1021.36	1021.93
				l		1312.46	1313.19

Glycan residue structures were built using GlycoWorkBench 2.1 build 146, according to the symbol nomenclature for glycan representation of the Consortium for Functional Glycomics (Varki *et al.*, 2009):  *N*-acetylgalactosamine (GalNAc),  *N*-acetylglucosamine (GlcNAc),  galactose (Gal),  *N*-acetylneuraminic acid (NeuAc), and  fucose (Fuc). The masses of the glycan residues are calculated excluding a water molecule each.

*Reported by Nwosu *et al.* (2010). All other structures are reproduced from O'Riordan *et al.* (2014).

Whey proteins

Table 1.5 summarises properties of the major whey proteins found in bovine milk in relation to reference proteoforms, where applicable. Because of their increased complexity due to the diversity of their variable regions, the immunoglobulins were not included in this table.

Table 1.5 – Properties of the major whey proteins in bovine milk

Protein	β -lactoglobulin	α -lactalbumin	Bovine serum albumin	Lactoferrin
Gene name	LGB	LALBA	ALBU	LTF
SwissProt accession	P02754	P00711	P02769	P24627
Conc. in milk (g/L)	2.00 - 4.00	0.60 - 1.70	0.40	0.02 - 0.10
Number of amino acid residues	162	123	583	689
Phosphorylation	-	-	Yes	-
Number of glycans	-	0-1	0-1	4-5
Disulfide bridges	2	4	17	17
Reference proteoform	b-Ig B	a-la B	-	-
Monoisotopic mass (Da) ^a	18269.42	14176.81	66389.86 ^c	76095.09 ^c
Average mass (Da) ^a	18281.01	14185.92	66432.27 ^c	76143.08 ^c
pI ^b	4.8	4.84	5.64 ^c	8.07 ^c

^aProtein masses calculated based on amino acid sequence, individual amino acid residue masses from http://www.matrixscience.com/help/aa_help.html, and masses of post-translational modifications from <https://www.sigmaaldrich.com/life-science/proteomics/post-translational-analysis/phosphorylation/mass-changes.html>. For the mass calculation it was assumed that all cysteine residues are reduced.

^bThe pI of the proteins was calculated using the ProtPi protein tool that also takes into account the presence of post-translational modification: <https://www.protpi.ch/Calculator/ProteinTool>.

^cThe calculations were performed on the amino acid sequence without including any PTMs.

α -Lactalbumin

α -Lactalbumin is the second most-abundant whey protein in bovine milk, representing ~20% of the total whey protein in milk. It is a regulatory subunit of lactose synthase that enables the synthesis of lactose in milk. While several genetic variants of α -lactalbumin have been reported in literature, bovine milk is dominated by variant B. The primary sequence of α -la B is shown in Figure 1.9. A small proportion of up to 10% of the total α -lactalbumin has been reported to be N-glycosylated. The protein contains 8 Cys-residues which are all interlinked in 4 intramolecular disulphide bonds. The structure is further stabilised by the binding

of one Ca²⁺ ion in a pocket containing 4 Asp residues. Removal of the calcium ion loosens the structure generating a molten globule state upon heating, in which the protein has increased susceptibility to denaturation and aggregation. This property is exploited in industrial-scale isolation of α -lactalbumin from milk.

β -Lactoglobulin

β -Lactoglobulin is the major whey protein, representing ~50% (*m/m*) of total whey proteins in bovine milk. While several genetic variants have been described for β -lactoglobulin, the most commonly encountered ones are variants A and B. The amino acid sequence of the reference protein, *i.e.*, β -lg B, is shown in Figure 1.10. β -Lactoglobulin A differs from variant B by two mutations, *i.e.*, Gly64 → Asp64 and Ala118 → Val118.

β -Lactoglobulin contains 5 Cys residues, of which 4 form 2 intramolecular disulphide bonds. The fifth, *i.e.*, Cys121, is buried in the native protein structure but becomes exposed upon unfolding and can participate in intermolecular thiol-thiol and thiol-disulphide interchange reactions. This is key in the heat-induced aggregation of whey proteins.

Bovine serum albumin

Bovine serum albumin is the third most abundant whey protein in mature milk after α -lactalbumin and β -lactoglobulin, amounting to up to 10% of total whey protein. The primary sequence of bovine serum albumin is shown in Figure 1.11. The protein contains 35 Cys residues, which form 17 intramolecular disulphide bridges, leaving Cys34 in the mature protein the only residue not engaged in an intramolecular disulphide bridge. In human serum albumin, the homologous Cys34 is often encountered cysteinylated (Nagumo *et al.*, 2014). The disulphide bridges are located in relatively close proximity, thereby creating a number of short loops in the structure. Although bovine serum albumin can bind metal ions and fatty

	Met	Met	Ser	Phe	Val	Ser	Leu	Leu	Leu	Val	1	Gly	Ile	Leu	Phe	His	Ala	Thr	Gln	Ala	Glu	1
2	Gln	Leu	Thr	Lys	Cys	Glu	Val	Phe	Arg	Glu	11	Leu	Lys	Asp	Leu	Lys	Gly	Tyr	Gly	Gly	Val	21
22	Ser	Leu	Pro	Glu	Trp	Val	Cys	Thr	Thr	Phe	31	His	Thr	Ser	Gly	Tyr	Asp	Thr	Gln	Ala	Ile	41
42	Val	Gln	Asn	Asn	Asp	Ser	Thr	Glu	Tyr	Gly	51	Leu	Phe	Gln	Ile	Asn	Asn	Lys	Ile	Trp	Cys	61
62	Lys	Asp	Asp	Gln	Asn	Pro	His	Ser	Ser	Asn	71	Ile	Cys	Asn	Ile	Ser	Cys	Asp	Lys	Phe	Leu	81
82	Asp	Asp	Asp	Leu	Thr	Asp	Asp	Ile	Met	Cys	91	Val	Lys	Lys	Ile	Leu	Asp	Lys	Val	Gly	Ile	101
102	Asn	Tyr	Trp	Leu	Ala	His	Lys	Ala	Leu	Cys	111	Ser	Glu	Lys	Leu	Asp	Gln	Trp	Leu	Cys	Glu	121
122	Lys	Leu									123											

Figure 1.9 – Primary sequence of α -la B (Brew, Castellino, Vanaman & Hill, 1970; SwissProt accession P00711). The following residues are highlighted on the sequence: Xxx = amino acids of the signal peptide; Cys = cysteine. Numbering starts at the N-terminus of the mature protein.

Chapter 1

	Met	Lys	Cys	Leu	Leu	Leu	Ala	Leu	Ala	Leu	1	Thr	Cys	Gly	Ala	Gln	Ala	Leu	Ile	Val	Thr	4
5	Gln	Thr	Met	Lys	Gly	Leu	Asp	Ile	Gln	Lys	14	Val	Ala	Gly	Thr	Trp	Tyr	Ser	Leu	Ala	Met	24
25	Ala	Ala	Ser	Asp	Ile	Ser	Leu	Leu	Asp	Ala	34	Gln	Ser	Ala	Pro	Leu	Arg	Val	Tyr	Val	Glu	44
45	Glu	Leu	Lys	Pro	Thr	Pro	Leu	Gly	Asp	Leu	54	Glu	Ile	Leu	Leu	Gln	Lys	Trp	Glu	Asn	Gly	64
65	Glu	Cys	Ala	Gln	Lys	Lys	Ile	Ile	Ala	Glu	74	Lys	Thr	Lys	Ile	Pro	Ala	Val	Phe	Lys	Ile	84
85	Asp	Ala	Leu	Asn	Glu	Asn	Lys	Val	Leu	Val	94	Leu	Asp	Thr	Asp	Tyr	Lys	Lys	Tyr	Leu	Leu	104
105	Phe	Cys	Met	Glu	Asn	Ser	Ala	Glu	Pro	Glu	114	Gln	Ser	Leu	Ala	Cys	Gln	Cys	Leu	Val	Arg	124
125	Thr	Pro	Glu	Val	Asp	Asp	Glu	Ala	Leu	Glu	134	Lys	Phe	Asp	Lys	Ala	Leu	Lys	Ala	Leu	Pro	144
145	Met	His	Ile	Arg	Leu	Ser	Phe	Asn	Pro	Thr	154	Gln	Leu	Glu	Glu	Gln	Cys	His	Ile			162

Figure 1.10 – Primary sequence of β -lg B (Braunitzer, Chen, Schrank & Stangl, 1972; SwissProt accession P02754). The following residues are highlighted on the sequence: Xxx = amino acids of the signal peptide; Cys = cysteine. Numbering starts at the N-terminus of the mature protein.

acids, no clear physiological role of bovine serum albumin in milk has been established to date. Likewise, the influence of bovine serum albumin on functional properties of milk protein preparations is also believed to be limited.

	Met	Lys	Trp	Val	Thr	Phe	Ile	Ser	Leu	Leu	6	Leu	Leu	Phe	Ser	Ser	Ala	Tyr	Ser	Arg	Gly	16
1	Val	Phe	Arg	Arg	Asp	Thr	His	Lys	Ser	Glu	6	Ile	Ala	His	Arg	Phe	Lys	Asp	Leu	Gly	Glu	16
17	Glu	His	Phe	Lys	Gly	Leu	Val	Leu	Ile	Ala	26	Phe	Ser	Gln	Tyr	Leu	Gln	Gln	Cys	Pro	Phe	36
37	Asp	Glu	His	Val	Lys	Leu	Val	Asn	Glu	Leu	46	Thr	Glu	Phe	Ala	Lys	Thr	Cys	Val	Ala	Asp	56
57	Glu	Ser	His	Ala	Gly	Cys	Glu	Lys	Ser	Leu	66	His	Thr	Leu	Phe	Gly	Asp	Glu	Leu	Cys	Lys	76
77	Val	Ala	Ser	Leu	Arg	Glu	Thr	Tyr	Gly	Asp	86	Met	Ala	Asp	Cys	Cys	Glu	Lys	Gln	Glu	Pro	96
97	Glu	Arg	Asn	Glu	Cys	Phe	Leu	Ser	His	Lys	106	Asp	Asp	Ser	Pro	Asp	Leu	Pro	Lys	Leu	Lys	116
117	Pro	Asp	Pro	Asn	Thr	Leu	Cys	Asp	Glu	Phe	126	Lys	Ala	Asp	Glu	Lys	Lys	Phe	Trp	Gly	Lys	136
137	Tyr	Leu	Tyr	Glu	Ile	Ala	Arg	Arg	His	Pro	146	Tyr	Phe	Tyr	Ala	Pro	Glu	Leu	Leu	Tyr	Tyr	156
157	Ala	Asn	Lys	Tyr	Asn	Gly	Val	Phe	Gln	Glu	166	Cys	Cys	Gln	Ala	Glu	Asp	Lys	Gly	Ala	Cys	176
177	Leu	Leu	Pro	Lys	Ile	Glu	Thr	Met	Arg	Glu	186	Lys	Val	Leu	Ala	Ser	Ser	Ala	Arg	Gln	Arg	196
197	Leu	Arg	Cys	Ala	Ser	Ile	Gln	Lys	Phe	Gly	206	Glu	Arg	Ala	Leu	Lys	Ala	Trp	Ser	Val	Ala	216
217	Arg	Leu	Ser	Gln	Lys	Phe	Pro	Lys	Ala	Glu	226	Phe	Val	Glu	Val	Thr	Lys	Leu	Val	Thr	Asp	236
237	Leu	Thr	Lys	Val	His	Lys	Glu	Cys	Cys	His	246	Gly	Asp	Leu	Leu	Glu	Cys	Ala	Asp	Asp	Arg	256
257	Ala	Asp	Leu	Ala	Tyr	Ile	Cys	Asp	Asn		266	Gln	Asp	Thr	Ile	Ser	Lys	Leu	Lys	Lys		276
277	Cys	Cys	Asp	Lys	Pro	Leu	Leu	Glu	Lys	Ser	286	His	Cys	Ile	Ala	Glu	Val	Glu	Lys	Asp	Ala	296
297	Ile	Pro	Glu	Asn	Leu	Pro	Pro	Leu	Thr	Ala	306	Asp	Phe	Ala	Glu	Asp	Lys	Asp	Val	Cys	Lys	316
317	Asn	Tyr	Gln	Glu	Ala	Lys	Asp	Ala	Phe	Leu	326	Gly	Ser	Phe	Leu	Tyr	Glu	Tyr	Ser	Arg	Arg	336
337	His	Pro	Glu	Tyr	Ala	Val	Ser	Val	Leu	Leu	346	Arg	Leu	Ala	Lys	Glu	Tyr	Glu	Ala	Thr	Leu	356
357	Glu	Glu	Cys	Cys	Ala	Lys	Asp	Asp	Pro	His	366	Ala	Cys	Tyr	Ser	Thr	Val	Phe	Asp	Lys	Leu	376
377	Lys	His	Leu	Val	Asp	Glu	Pro	Gln	Asn	Leu	386	Ile	Lys	Gln	Asn	Cys	Asp	Gln	Phe	Glu	Lys	396
397	Leu	Gly	Glu	Tyr	Gly	Phe	Gln	Asn	Ala	Leu	406	Ile	Val	Arg	Tyr	Thr	Arg	Lys	Val	Pro	Gln	416
417	Val	Ser	Thr	Pro	Thr	Leu	Val	Glu	Val	Ser	426	Arg	Ser	Leu	Gly	Lys	Val	Gly	Thr	Arg	Cys	436
437	Cys	Thr	Lys	Pro	Glu	Ser	Glu	Arg	Met	Pro	446	Cys	Thr	Glu	Asp	Tyr	Leu	Ser	Leu	Ile	Leu	456
457	Asn	Arg	Leu	Cys	Val	Leu	His	Glu	Lys	Thr	466	Pro	Val	Ser	Glu	Lys	Val	Thr	Lys	Cys	Cys	476
477	Thr	Glu	Ser	Leu	Val	Asn	Arg	Arg	Pro	Cys	486	Phe	Ser	Ala	Leu	Thr	Pro	Asp	Glu	Thr	Tyr	496
497	Val	Pro	Lys	Ala	Phe	Asp	Glu	Lys	Leu	Phe	506	Thr	Phe	His	Ala	Asp	Ile	Cys	Thr	Leu	Pro	516
517	Asp	Thr	Glu	Lys	Gln	Ile	Lys	Lys	Gln	Thr	526	Ala	Leu	Val	Glu	Leu	Leu	Lys	His	Lys	Pro	536
537	Lys	Ala	Thr	Glu	Glu	Gln	Leu	Lys	Thr	Val	546	Met	Glu	Asn	Phe	Val	Ala	Phe	Val	Asp	Lys	556
557	Cys	Cys	Ala	Ala	Asp	Asp	Lys	Glu	Ala	Cys	566	Phe	Ala	Val	Glu	Gly	Pro	Lys	Leu	Val	Val	576
577	Ser	Thr	Gln	Thr	Ala	Leu	Ala				583											

Figure 1.11 – Primary sequence of bovine serum albumin (Brown, 1975; SwissProt accession P02769). The following residues are highlighted on the sequence: Xxx = amino acids of the signal peptide; Xxx = amino acids of the propeptide; Cys = cysteine. Numbering starts at the N-terminus of the mature protein.

Lactoferrin

Lactoferrin is an iron-binding protein that is found at low concentrations (0.03-0.30 g·L⁻¹) in bovine milk but at considerably higher concentrations (up to 7 g·L⁻¹) in human milk. The amino acid sequence of bovine milk lactoferrin is shown in Figure 1.12. All Cys residues in the protein are involved in intra-molecular disulphide bonds. Lactoferrin in bovine milk is *N*-glycosylated at 4 Asn residues, *i.e.*, Asn233, Asn368, Asn476 and Asn545. In colostrum a 5th *N*-glycosylation site is also occupied, *i.e.*, Asn281. Unlike most other milk proteins, lactoferrin carries a net-positive charge at milk pH, a feature which is utilized in its isolation from milk or whey. Although lactoferrin is known as an iron-binding protein, it can also bind other cations, *e.g.*, copper or manganese. Biological functions of lactoferrin include, amongst others, immune modulatory, anti-bacterial and anti-viral activities.

	Met	Lys	Leu	Phe	Val	Pro	Ala	Leu	Leu	Ser	1	Leu	Gly	Ala	Leu	Gly	Leu	Cys	Leu	Ala	Ala	1
2	Pro	Arg	Lys	Asn	Val	Arg	Trp	Cys	Thr	Ile	11	Ser	Gln	Pro	Glu	Trp	Phe	Lys	Cys	Arg	Arg	21
22	Trp	Gln	Trp	Arg	Met	Lys	Lys	Leu	Gly	Ala	31	Pro	Ser	Ile	Thr	Cys	Val	Arg	Arg	Ala	Phe	41
42	Ala	Leu	Glu	Cys	Ile	Arg	Ala	Ile	Ala	Glu	51	Lys	Lys	Ala	Asp	Ala	Val	Thr	Leu	Asp	Gly	61
62	Gly	Met	Val	Phe	Glu	Ala	Gly	Arg	Asp	Pro	71	Tyr	Lys	Leu	Arg	Pro	Val	Ala	Ala	Glu	Ile	81
82	Tyr	Gly	Thr	Lys	Glu	Ser	Pro	Gln	Thr	His	91	Tyr	Tyr	Ala	Val	Ala	Val	Val	Lys	Lys	Gly	101
102	Ser	Asn	Phe	Gln	Leu	Asp	Gln	Leu	Gln	Gly	111	Arg	Lys	Ser	Cys	His	Thr	Gly	Leu	Gly	Arg	121
122	Ser	Ala	Gly	Trp	Ile	Ile	Pro	Met	Gly	Ile	131	Leu	Arg	Pro	Tyr	Leu	Ser	Trp	Thr	Glu	Ser	141
142	Leu	Glu	Pro	Leu	Gln	Gly	Ala	Val	Ala	Lys	151	Phe	Phe	Ser	Ala	Ser	Cys	Val	Pro	Cys	Ile	161
162	Asp	Arg	Gln	Ala	Tyr	Pro	Asn	Leu	Cys	Gln	171	Leu	Cys	Lys	Gly	Glu	Gly	Glu	Asn	Gln	Cys	181
182	Ala	Cys	Ser	Ser	Arg	Glu	Pro	Tyr	Phe	Gly	191	Tyr	Ser	Gly	Ala	Phe	Lys	Cys	Leu	Gln	Asp	201
202	Gly	Ala	Gly	Asp	Val	Ala	Phe	Val	Lys	Glu	211	Thr	Thr	Val	Phe	Glu	Asn	Leu	Pro	Glu	Lys	221
222	Ala	Asp	Arg	Asp	Gln	Tyr	Glu	Leu	Leu	Cys	231	Leu	Asn	Asn	Ser	Arg	Ala	Pro	Val	Asp	Ala	241
242	Phe	Lys	Glu	Cys	His	Leu	Ala	Gln	Val	Pro	251	Ser	His	Ala	Val	Val	Ala	Arg	Ser	Val	Asp	261
262	Gly	Lys	Glu	Asp	Leu	Ile	Trp	Lys	Leu	Leu	271	Ser	Lys	Ala	Gln	Glu	Lys	Phe	Gly	Lys	Asn	281
282	Lys	Ser	Arg	Ser	Phe	Gln	Leu	Phe	Gly	Ser	291	Pro	Pro	Gly	Gln	Arg	Asp	Leu	Leu	Phe	Lys	301
302	Asp	Ser	Ala	Leu	Gly	Phe	Leu	Arg	Ile	Pro	311	Ser	Lys	Val	Asp	Ser	Ala	Leu	Tyr	Leu	Gly	321
322	Ser	Arg	Tyr	Leu	Thr	Thr	Leu	Lys	Asn	Leu	331	Arg	Glu	Thr	Ala	Glu	Glu	Val	Lys	Ala	Arg	341
342	Tyr	Thr	Arg	Val	Val	Trp	Cys	Ala	Val	Gly	351	Pro	Glu	Glu	Gln	Lys	Lys	Cys	Gln	Gln	Trp	361
362	Ser	Gln	Gln	Ser	Gly	Gln	Asn	Val	Thr	Cys	371	Ala	Thr	Ala	Ser	Thr	Thr	Asp	Asp	Cys	Ile	381
382	Val	Leu	Val	Leu	Lys	Gly	Glu	Ala	Asp	Ala	391	Leu	Asn	Leu	Asp	Gly	Gly	Tyr	Ile	Tyr	Thr	401
402	Ala	Gly	Lys	Cys	Gly	Leu	Val	Pro	Val	Leu	411	Ala	Glu	Asn	Arg	Lys	Ser	Ser	Lys	His	Ser	421
422	Ser	Leu	Asp	Cys	Val	Leu	Arg	Pro	Thr	Glu	431	Gly	Tyr	Leu	Ala	Val	Ala	Val	Val	Lys	Lys	441
442	Ala	Asn	Glu	Gly	Leu	Thr	Trp	Asn	Ser	Leu	451	Lys	Asp	Lys	Lys	Ser	Cys	His	Thr	Ala	Val	461
462	Asp	Arg	Thr	Ala	Gly	Trp	Asn	Ile	Pro	Met	471	Gly	Leu	Ile	Val	Asn	Gln	Thr	Gly	Ser	Cys	481
482	Ala	Phe	Asp	Glu	Phe	Ser	Gln	Ser	Cys	491	Ala	Pro	Gly	Ala	Asp	Pro	Lys	Ser	Arg	Leu	501	
502	Cys	Ala	Leu	Cys	Ala	Gly	Asp	Asp	Gln	Gly	511	Leu	Asp	Lys	Cys	Val	Pro	Asn	Ser	Lys	Glu	521
522	Lys	Tyr	Tyr	Gly	Tyr	Thr	Gly	Ala	Phe	Arg	531	Cys	Leu	Ala	Glu	Asp	Val	Gly	Asp	Val	Ala	541
542	Phe	Val	Lys	Asn	Asp	Thr	Val	Trp	Glu	Asn	551	Thr	Asn	Gly	Glu	Ser	Thr	Ala	Asp	Trp	Ala	561
562	Lys	Asn	Leu	Asn	Arg	Glu	Asp	Phe	Arg	Leu	571	Leu	Cys	Leu	Asp	Gly	Thr	Arg	Lys	Pro	Val	581
582	Thr	Glu	Ala	Gln	Ser	Cys	His	Leu	Ala	Val	591	Ala	Pro	Asn	His	Ala	Val	Val	Ser	Arg	Ser	601
602	Asp	Arg	Ala	Ala	His	Val	Lys	Gln	Val	Leu	611	Leu	His	Gln	Gln	Ala	Leu	Phe	Gly	Lys	Asn	621
622	Gly	Lys	Asn	Cys	Pro	Asp	Lys	Phe	Cys	Leu	631	Phe	Lys	Ser	Glu	Thr	Lys	Asn	Leu	Leu	Phe	641
642	Asn	Asp	Asn	Thr	Glu	Cys	Leu	Ala	Lys	Leu	651	Gly	Gly	Arg	Pro	Thr	Tyr	Glu	Glu	Tyr	Leu	661
662	Gly	Thr	Glu	Tyr	Val	Thr	Ala	Ile	Ala	Asn	671	Leu	Lys	Lys	Cys	Ser	Thr	Ser	Pro	Leu	Leu	681
682	Glu	Ala	Cys	Ala	Phe	Leu	Thr	Arg			689											

Figure 1.12 – Primary sequence of lactoferrin (SwissProt accession P24627). The following residues are highlighted on the sequence: Xxx = amino acids of the signal peptide; Asn = *N*-glycosylated asparagine; Cys = cysteine. Numbering starts at the *N*-terminus of the mature protein.

Immunoglobulins

In order to provide passive immunity to the new-born calf, immunoglobulins are the key constituent of bovine colostrum (Weström *et al.*, 2020), *i.e.*, the mammary secretion produced in the first 3 d post-partum (McGrath *et al.*, 2016). Bovine colostrum has a protein concentration up to ~8 times higher than that of mature milk, and up to 80% of the total protein of colostrum is made up of immunoglobulins, particularly immunoglobulin G (IgG) (Marnila and Korhonen, 2011). During the first month of lactation, the mammary secretion gradually transitions from colostrum to mature milk (Madsen *et al.*, 2004). The immunoglobulins still remain present throughout lactation, but contrary to colostrum, they only constitute a minor proportion of the whey proteins in mature milk (Stelwagen *et al.*, 2009). Nevertheless, the immunoglobulins are amongst the most important functional proteins in bovine milk. As a result of their genetic polymorphism, post-translational modifications, and diversity of the variable regions, they are possibly also the most heterogeneous proteins of milk.

The major immunoglobulins in bovine milk are the IgGs with subclasses 1, 2 (Butler, 1983) and 3 being present in decreasing order of abundance (Gazi *et al.*, 2023, Zhang *et al.*, 2022). Bovine milk immunoglobulins are particularly dominated by IgGs due to the fact that unlike humans, the anatomy of the bovine placenta does not allow the passive transfer of IgGs to the unborn calf (Chucru *et al.*, 2010). The calf is therefore born without systemic immunity, and its survival depends on receiving passive immunity *via* the maternal IgGs through milk post-partum (Godden *et al.*, 2019). By contrast, the major immunoglobulin class found in human milk is secretory IgA (sIgA) (Atyeo and Alter, 2021, Dingess *et al.*, 2022). Bovine milk also contains sIgA, though at lower abundances relative to the IgG concentrations. Next to sIgA, sIgM is also present in bovine mammary secretions, particularly abundantly in colostrum (Butler, 1983, Gazi *et al.*, 2023).

As depicted in Figure 1.13, an immunoglobulin monomer is comprised of 2 identical heavy chains and 2 identical light chains that are interlinked through several disulphide bridges. Both the light chains and the heavy chains are divided into constant and variable domains. In immunoglobulin biosynthesis, variable (V), diversity (D) and joining (J) DNA segments are selected and recombined to form the code for the variable domain, a process known as V(D)J recombination (Chi *et al.*, 2020). Next, during class switch recombination, gene segments on the heavy chain locus are excised to create the code for a specific antibody class defining the constant domain (Chi *et al.*, 2020). These gene segments are part of the germline genome, *i.e.*, genetic material that can be passed down to offspring. Consequently, the constant regions of immunoglobulins are well characterized and can be

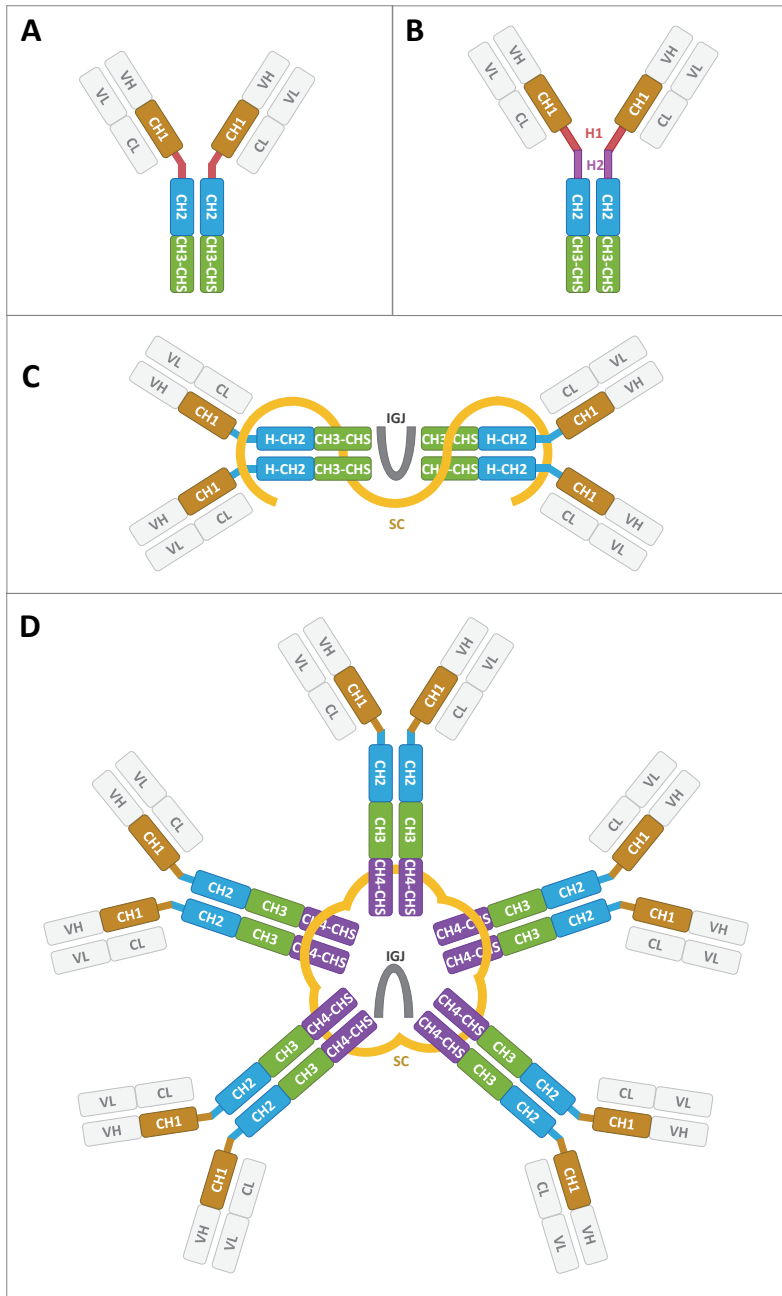


Figure 1.13 – Artistic impression of bovine milk immunoglobulin structures. A) Monomeric IgG1 and IgG2; B) Monomeric IgG3; C) Dimeric sIgA; D) Pentameric sIgM1 and sIgM2. The colour coding used in the illustrations corresponds to the colour coding used in Figures 1.14–1.21 that show the amino acid sequences of the constant heavy regions and of the immunoglobulin J chain and polymeric immunoglobulin receptor, respectively. VL = variable light; VH = variable heavy; CL = constant light; CH = constant heavy; H = hinge; IGJ = immunoglobulin J chain; SC = secretory component.

identified when searched against a protein database. The complementarity-determining regions (CDRs), also known as the hypervariable regions in the variable domain that play a dominant role in antigen recognition, undergo further somatic hypermutation (Chi *et al.*, 2020), which ensures the adaptation of the immune system to new antigens. Somatic hypermutation occurs in individual cells and is not part of the germline genome, *i.e.*, cannot be passed down to the offspring. Consequently, it is extremely challenging to study the exact amino acid sequences of these variable domains. As a result of these biochemical mechanisms that increase their diversity, and of their genetic polymorphism and post-translational modifications, immunoglobulins are arguably the most heterogeneous proteins in milk (as well as in circulation).

The sequences of the bovine milk immunoglobulin constant heavy chains shown in Figures 1.14-1.16 and 1.19-1.21 were downloaded on 01.02.2021 from the website of the International ImMunoGeneTics Information System (<http://www.imgt.org>). The constant heavy chains can further be divided into several regions: the constant heavy domain 1 (CH1) is located in the fragment antigen-binding (Fab) region along with the light chain and the variable domains. CH1 is followed by the hinge (H) region of the antibody. The tail region of the antibody also known as the fragment crystallisable (Fc) region contains the C-terminal CH2-CH3 domains in IgG and IgA, and CH2-CH4 domains in IgM.

1	Xaa	Ser	Thr	Thr	Ala	Pro	Lys	Val	Tyr	Pro	10	Leu	Ser	Ser	Cys	Cys	Gly	Asp	Lys	Ser	Ser	20
21	Ser	Thr	Val	Thr	Leu	Gly	Cys	Leu	Val	Ser	30	Ser	Tyr	Met	Pro	Glu	Pro	Val	Thr	Val	Thr	40
41	Trp	Asn	Ser	Gly	Ala	Leu	Lys	Ser	Gly	Val	50	His	Thr	Phe	Pro	Ala	Val	Leu	Gln	Ser	Ser	60
61	Gly	Leu	Tyr	Ser	Leu	Ser	Ser	Met	Val	Thr	70	Val	Pro	Gly	Ser	Thr	Ser	Gly	Gln	Thr	Phe	80
81	Thr	Cys	Asn	Val	Ala	His	Pro	Ala	Ser	Ser	90	Thr	Lys	Val	Asp	Lys	Ala	Val	Asp	Pro	Thr	100
101	Cys	Lys	Pro	Ser	Pro	Cys	Asp	Cys	Cys	Pro	110	Pro	Pro	Glu	Leu	Pro	Gly	Gly	Pro	Ser	Val	120
121	Phe	Ile	Phe	Pro	Pro	Lys	Pro	Lys	Asp	Thr	130	Leu	Thr	Ile	Ser	Gly	Thr	Pro	Glu	Val	Thr	140
141	Cys	Val	Val	Asp	Val	Gly	His	Asp	Asp	150	Pro	Glu	Val	Lys	Phe	Ser	Trp	Phe	Val	Asp	160	
161	Asp	Val	Glu	Val	Asn	Thr	Ala	Thr	Thr	Lys	170	Pro	Arg	Glu	Glu	Gln	Phe	Asn	Ser	Thr	Tyr	180
181	Arg	Val	Val	Ser	Ala	Leu	Arg	Ile	Gln	His	190	Gln	Asp	Trp	Thr	Gly	Gly	Lys	Glu	Phe	Lys	200
201	Cys	Lys	Val	His	Asn	Glu	Gly	Leu	Pro	Ala	210	Pro	Ile	Val	Arg	Thr	Ile	Ser	Arg	Thr	Lys	220
221	Gly	Pro	Ala	Arg	Glu	Pro	Gln	Val	Tyr	Val	230	Leu	Ala	Pro	Pro	Gln	Glu	Glu	Leu	Ser	Lys	240
241	Ser	Thr	Val	Ser	Leu	Thr	Cys	Met	Val	Thr	250	Ser	Phe	Tyr	Pro	Asp	Tyr	Ile	Ala	Val	Glu	260
261	Trp	Gln	Arg	Asn	Gly	Gln	Pro	Glu	Ser	Glu	270	Asp	Lys	Tyr	Gly	Thr	Thr	Pro	Pro	Gln	Leu	280
281	Asp	Ala	Asp	Ser	Ser	Tyr	Phe	Leu	Tyr	Ser	290	Lys	Leu	Arg	Val	Asp	Arg	Asn	Ser	Trp	Gln	300
301	Glu	Gly	Asp	Thr	Tyr	Thr	Cys	Val	Val	Met	310	His	Glu	Ala	Leu	His	Asn	His	Tyr	Thr	Gln	320
321	Lys	Ser	Thr	Ser	Lys	Ser	Ala	Gly	Lys		329											

Figure 1.14 – Primary sequence of the IgG1 constant heavy chain allele 1 (IMGT accession number X16701). The following residues are highlighted on the sequence: **Asn** = N-glycosylated asparagine; **Cys** = cysteine; **Xxx** = heavy constant 1 region; **Xxx** = hinge region; **Xxx** = heavy constant 2 region; **Xxx** = heavy constant 3 region; numbering starts at the N-terminus of the constant region of the heavy chain.

IgGs are monomeric immunoglobulins. The structures of bovine IgGs are depicted in Figure 1.13 where panel A is representative of IgG1 and IgG2, whereas panel B is representative of IgG3. Figures 1.14-1.16 show the amino acid sequences of the

IgG1-IgG3 heavy constant chains. While bovines only produce 3 classes of IgGs and humans 4, similarities can be drawn between the bovine and human IgGs. Particularly in the case of IgG3 it can be observed that the hinge region is longer than those of the other IgGs. From the amino acid sequence of IgG3 (Figure 1.16) it is apparent that the hinge region is rich in proline and threonine residues, which are typical amino acids found in mucin-like domains (Brockhausen *et al.*, 2022). A parallel can be drawn to the human IgG3, whose hinge region is also mucin-like and is rich in *O*-glycosylation (Plomp *et al.*, 2015) which protects the immunoglobulin from proteolytic degradation (Wandall *et al.*, 2021). It can therefore be expected for *O*-glycosylation to also be found in the bovine IgG3 hinge region.

Compared to *O*-glycosylation, *N*-glycosylation commonly occurs on all immunoglobulins (de Haan *et al.*, 2020). Unlike *O*-glycosylation which may occur on any Ser or Thr residue, *N*-glycosylation typically follows the canonical motif of Asn-Xxx-Ser/Thr, where the Asn that is glycosylated is followed by any amino acid with the exception of Pro at position +1 and by either a Ser or Thr at position +2 (Stanley *et al.*, 2022). *N*-Glycosylation motifs are highlighted in all of the immunoglobulin heavy chain constant region sequences and related proteins depicted in Figures 1.14-1.21. Further increases in protein diversity result from glycosylation micro-, macro- and meta-heterogeneity, i.e., site-specific occupancy, site-specific variation in glycoforms and protein-specific variation across all its glycosylation sites (Čaval *et al.*, 2021).

Next to variable domain sequence, disulphide bridging, *O*- and *N*-glycosylation micro-, macro- and meta-heterogeneity, immunoglobulin diversity can further be increased by genetic polymorphism, as previously also described for the major bovine milk proteins. Tables 1.6-1.10 show the differences in primary structure of

Table 1.6 – Differences in the amino acid sequences of the alleles of IgG1 heavy chain compared to IGHG1*01

Allele	Amino acid position								
	1	75	Insertion between 77-78	100	103	104	105	284	291
IGHG1*01	Xaa	Thr		Thr	Pro	Ser	Pro	Ser	Lys
IGHG1*02	Ala								
IGHG1*03	Ala		Thr	Arg	Thr	deletion	Thr	Gly	Arg
IGHG1*04		Ser		Arg	Thr	deletion	Thr	Gly	Arg

Chapter 1

1	Xaa	Ser	Thr	Thr	Ala	Pro	Lys	Val	Tyr	Pro	10	Leu	Ser	Ser	Cys	Cys	Gly	Asp	Lys	Ser	Ser	20
21	Ser	Thr	Val	Thr	Leu	Gly	Cys	Leu	Val	Ser	30	Ser	Tyr	Met	Pro	Glu	Pro	Val	Thr	Val	Ser	40
41	Trp	Asn	Ser	Gly	Ala	Leu	Lys	Ser	Gly	Val	50	His	Thr	Phe	Pro	Ala	Val	Leu	Gln	Ser	Ser	60
61	Gly	Leu	Tyr	Ser	Leu	Ser	Ser	Met	Val	Thr	70	Val	Pro	Gly	Ser	Thr	Ser	Gly	Gln	Thr	Phe	80
81	Thr	Cys	Asn	Val	Ala	His	Pro	Ala	Ser	Ser	90	Thr	Lys	Val	Asp	Lys	Ala	Val	Gly	Val	Ser	100
101	Ser	Asp	Cys	Ser	Lys	Pro	Asn	Asn	Gln	His	110	Cys	Xaa	Arg	Glu	Pro	Ser	Val	Phe	Ile	Phe	120
121	Pro	Pro	Lys	Pro	Lys	Asp	Thr	Leu	Met	Ile	130	Thr	Gly	Thr	Pro	Glu	Val	Thr	Cys	Val	Val	140
141	Val	Asn	Val	Gly	His	Asp	Asn	Pro	Glu	Val	150	Gln	Phe	Ser	Trp	Phe	Val	Asp	Asp	Val	Glu	160
161	Val	His	Thr	Ala	Arg	Thr	Lys	Pro	Arg	Glu	170	Glu	Gln	Phe	Asn	Ser	Thr	Tyr	Arg	Val	Val	180
181	Ser	Ala	Leu	Pro	Ile	Gln	His	Gln	Asp	Trp	190	Thr	Gly	Gly	Lys	Glu	Phe	Lys	Cys	Lys	Val	200
201	Asn	Ile	Lys	Gly	Leu	Ser	Ala	Ser	Ile	Val	210	Arg	Ile	Ile	Ser	Arg	Ser	Lys	Gly	Pro	Ala	220
221	Arg	Glu	Pro	Gln	Val	Tyr	Val	Leu	Asp	Pro	230	Pro	Lys	Glu	Glu	Leu	Ser	Lys	Ser	Thr	Val	240
241	Ser	Val	Thr	Cys	Met	Val	Ile	Gly	Phe	Tyr	250	Pro	Glu	Asp	Val	Asp	Val	Glu	Trp	Gln	Arg	260
261	Asp	Arg	Gln	Thr	Glu	Ser	Glu	Asp	Lys	Tyr	270	Arg	Thr	Thr	Pro	Pro	Gln	Leu	Asp	Ala	Asp	280
281	Arg	Ser	Tyr	Phe	Leu	Tyr	Ser	Lys	Leu	Arg	290	Val	Asp	Arg	Asn	Ser	Trp	Gln	Arg	Gly	Asp	300
301	Thr	Tyr	Thr	Cys	Val	Val	Met	His	Glu	Ala	310	Leu	His	Asn	His	Tyr	Met	Gln	Lys	Ser	Thr	320
321	Ser	Lys	Ser	Ala	Gly	Lys					326											

Figure 1.15 – Primary sequence of the IgG2 constant heavy chain allele 1 (IMGT accession number M36946). The following residues are highlighted on the sequence: **Asn** = N-glycosylated asparagine; **Cys** = cysteine; **Xxx** = heavy constant 1 region; **Xxx** = hinge region; **Xxx** = heavy constant 2 region; **Xxx** = heavy constant 3 region; numbering starts at the N-terminus of the constant region of the heavy chain.

the immunoglobulin heavy chain alleles described in IMGT compared to allele 1 of each immunoglobulin.

The secretory immunoglobulins sIgA and sIgM are typically present in larger complexes when compared to monomeric IgGs (Figure 1.13C and D). sIgA is typically present as an immunoglobulin dimer that is linked together by the immunoglobulin J-chain (IGJ) and secretory component (SC) through disulphide

Table 1.7 – Differences in the amino acid sequences of the alleles of IgG2 heavy chain compared to IGHG2*01

Allele	Amino acid position																							
	1	12	14	18	22	73	75	Insertion between 77-78	101	106	107	110	112	166	202	208	240	242	247	255	261	279	298	301
IGHG2*01	Xaa	Ser	Cys	Lys	Thr	Gly	Thr		Ser	Pro	Asn	His	Xaa	Thr	Ile	Ser	Val	Val	Ile	Asp	Asp	Ala	Arg	Thr
IGHG2*02					Gly	Ala	Ser	Thr					Val					Leu						
IGHG2*03	Ala	Ala	Ser	Thr		Ala	Ser		Ile	Cys	His	Pro	Val	Ser	Asn	Pro	Leu		Thr	Ala	Asn	Thr	Glu	Ala
IGHG2*04		Ala	Arg	Thr			Ser		Ile	Cys	His	Pro	Val	Ser	Asn	Pro	Leu		Thr	Ala	Asn	Thr	Glu	

1	Xaa	Ser	Thr	Thr	Ala	Pro	Lys	Val	Tyr	Pro	10	Leu	Ala	Ser	Ser	Cys	Gly	Asp	Thr	Ser	Ser	20
21	Ser	Thr	Val	Thr	Leu	Gly	Cys	Leu	Val	Ser	30	Ser	Tyr	Met	Pro	Glu	Pro	Val	Thr	Val	Thr	40
41	Trp	Asn	Ser	Gly	Ala	Leu	Lys	Ser	Gly	Val	50	His	Thr	Phe	Pro	Ala	Val	Arg	Gln	Ser	Ser	60
61	Gly	Leu	Tyr	Ser	Leu	Ser	Ser	Met	Val	Thr	70	Val	Pro	Ala	Ser	Ser	Ser	Glu	Thr	Gln	Thr	80
81	Phe	Thr	Cys	Asn	Val	Ala	His	Pro	Ala	Ser	90	Ser	Thr	Lys	Val	Asp	Lys	Ala	Val	Thr	Ala	100
101	Arg	Arg	Pro	Val	Pro	Thr	Thr	Pro	Lys	Thr	110	Thr	Ile	Pro	Pro	Gly	Lys	Pro	Thr	Thr	Pro	120
121	Lys	Ser	Glu	Val	Glu	Lys	Thr	Pro	Cys	Gln	130	Cys	Ser	Lys	Cys	Pro	Glu	Pro	Leu	Gly	Gly	140
141	Leu	Ser	Val	Phe	Ile	Phe	Pro	Pro	Lys	Pro	150	Lys	Asp	Thr	Leu	Thr	Ile	Ser	Gly	Thr	Pro	160
161	Glu	Val	Thr	Cys	Val	Val	Val	Asp	Val	Gly	170	Gln	Asp	Asp	Pro	Glu	Val	Gln	Phe	Ser	Trp	180
181	Phe	Val	Asp	Asp	Val	Glu	Val	His	Thr	Ala	190	Arg	Thr	Lys	Pro	Arg	Glu	Glu	Gln	Phe	Asn	200
201	Ser	Thr	Tyr	Arg	Val	Val	Ser	Ala	Leu	Arg	210	Ile	Gln	His	Gln	Asp	Trp	Leu	Gln	Gly	Lys	220
221	Glu	Phe	Lys	Cys	Lys	Val	Asn	Asn	Lys	Gly	230	Leu	Pro	Ala	Pro	Ile	Val	Arg	Thr	Ile	Ser	240
241	Arg	Thr	Lys	Gly	Gln	Ala	Arg	Glu	Pro	Gln	250	Val	Tyr	Val	Leu	Ala	Pro	Pro	Arg	Glu	Glu	260
261	Leu	Ser	Lys	Ser	Thr	Leu	Ser	Leu	Thr	Cys	270	Leu	Ile	Thr	Gly	Phe	Tyr	Pro	Glu	Thr	Ile	280
281	Asp	Val	Glu	Trp	Gln	Arg	Asn	Gly	Gln	Pro	290	Glu	Ser	Glu	Asp	Lys	Tyr	His	Thr	Thr	Ala	300
301	Pro	Gln	Leu	Asp	Ala	Asp	Gly	Ser	Tyr	Phe	310	Leu	Tyr	Ser	Lys	Leu	Arg	Val	Asn	Lys	Ser	320
321	Ser	Trp	Gln	Glu	Gly	Asp	His	Tyr	Thr	Cys	330	Ala	Val	Met	His	Glu	Ala	Leu	Arg	Asn	His	340
341	Tyr	Lys	Glu	Lys	Ser	Ile	Ser	Arg	Ser	Pro	350	Gly	Lys									352

Figure 1.16 – Primary sequence of the IgG3 constant heavy chain allele 1 (IMGT accession number U63638). The following residues are highlighted on the sequence: **Asn** = N-glycosylated asparagine; **Cys** = cysteine; **Xxx** = heavy constant 1 region; **Xxx** = hinge region 1; **Xxx** = hinge region 2; **Xxx** = heavy constant 2 region; **Xxx** = heavy constant 3 region; numbering starts at the N-terminus of the constant region of the heavy chain.

bridging (Kumar *et al.*, 2020). Higher order oligomeric structures have been described for human sIgA (Kumar *et al.*, 2020) and it is therefore expected that bovine sIgA may occur as not only a dimer, but also higher order oligomers. As is the case of the immunoglobulin heavy chains, IGJ (Figure 1.17) and SC (Figure 1.18) also contain N-glycosylation motifs, can therefore be glycosylated and further increase the heterogeneity of immunoglobulin proteoforms.

Table 1.8 – Differences in the amino acid sequences of the alleles of IgG3 heavy chain compared to IGHG3*01

Allele	Amino acid position									
	14	57	75	77	120	121	122	138	250	314
IGHG3*01	Ser	Arg	Ser	Glu	Pro	Lys	Ser	Leu	Gln	Lys
IGHG3*02	Arg	Leu	Thr		Gln	Glu				Arg
IGHG3*03			Thr	Gly	Gln	Glu	Pro	Pro	Leu	Arg

1	<i>Met</i>	<i>Lys</i>	<i>Asn</i>	<i>Cys</i>	<i>Leu</i>	<i>Leu</i>	<i>Phe</i>	<i>Trp</i>	<i>Gly</i>	<i>Val</i>	8	Leu	Ala	Ile	Phe	Val	Met	Ala	Val	Leu	Val	18
21	<i>Thr</i>	<i>Ser</i>	<i>Arg</i>	<i>Ile</i>	<i>Ile</i>	<i>Pro</i>	<i>Ser</i>	<i>Ala</i>	<i>Glu</i>	<i>Asp</i>	28	Pro	Ser	Gln	Asp	Ile	Val	Glu	Arg	Asn	Val	38
41	<i>Arg</i>	<i>Ile</i>	<i>Ile</i>	<i>Val</i>	<i>Pro</i>	<i>Leu</i>	<i>Asn</i>	<i>Ser</i>	<i>Arg</i>	<i>Glu</i>	48	Asn	Ile	Ser	Asp	Pro	Thr	Ser	Pro	Met	Arg	58
61	<i>Lys</i>	<i>Phe</i>	<i>Val</i>	<i>Tyr</i>	<i>His</i>	<i>Leu</i>	<i>Ser</i>	<i>Asp</i>	<i>Leu</i>		68	Cys	Lys	Lys	Cys	Asp	Thr	Thr	Glu	Val	Glu	78
81	<i>Leu</i>	<i>Glu</i>	<i>Asp</i>	<i>Gln</i>	<i>Val</i>	<i>Val</i>	<i>Thr</i>	<i>Ala</i>	<i>Ser</i>	<i>Gln</i>	88	Ser	Asn	Ile	Cys	Asp	Ser	Asp	Ala	Glu	Thr	98
101	<i>Cys</i>	<i>Tyr</i>	<i>Thr</i>	<i>Tyr</i>	<i>Asp</i>	<i>Arg</i>	<i>Asn</i>	<i>Lys</i>	<i>Cys</i>	<i>Tyr</i>	108	Thr	Asn	Arg	Val	Lys	Leu	Ser	Tyr	Arg	Gly	118
121	<i>Gln</i>	<i>Thr</i>	<i>Lys</i>	<i>Met</i>	<i>Val</i>	<i>Glu</i>	<i>Thr</i>	<i>Ala</i>	<i>Leu</i>	<i>Thr</i>	128	Pro	Asp	Ser	Cys	Tyr	Pro	Asp				135

Figure 1.17 – Primary sequence of immunoglobulin J chain (UniProt accession number Q3SYR8, 01.02.2021). The following residues are highlighted on the sequence: **Asn** = N-glycosylated asparagine; **Cys** = cysteine. The first 22 amino acids depicted in italic represent the signal peptide. Numbering starts at the N-terminus of the mature protein.

The primary sequence of bovine IgA heavy constant is shown in Figure 1.19. Interestingly, bovines produce a single subclass of IgA as opposed to humans that produce IgA1 and IgA2 (Stanfield *et al.*, 2018). Based on sequence similarity and length and composition of the hinge region, bovine IgA appears to be more closely related to human IgA2 than IgA1 (Cakebread *et al.*, 2015). In case of IgM, while humans produce a single subclass, bovine produce 2 subclasses: IgM1 (Figure 1.20) and IgM2 (Figure 1.21) (Stanfield *et al.*, 2018). The sIgM structures, unlike the sIgA, comprise 5 immunoglobulin monomers, 1 IGJ and 1 SC interlinked through disulphide bridges (Kumar *et al.*, 2021). A further distinction of IgM from the other immunoglobulin classes is that its Fc portion is longer and contains an additional CH domain.

	<i>Met</i>	<i>Ser</i>	<i>Arg</i>	<i>Leu</i>	<i>Phe</i>	<i>Leu</i>	<i>Ala</i>	<i>Cys</i>	<i>Leu</i>	<i>Leu</i>		<i>Ala</i>	<i>Ile</i>	<i>Phe</i>	<i>Pro</i>	<i>Val</i>	<i>Val</i>	<i>Ser</i>	<i>Met</i>	<i>Lys</i>	<i>Ser</i>	2
3	Pro	Ile	Phe	Gly	Pro	Glu	Glu	Val	Thr	Ser	12	Val	Glu	Gly	Arg	Ser	Val	Ser	Ile	Lys	Cys	22
23	Tyr	Tyr	Pro	Pro	Thr	Ser	Val	Asn	Arg	His	32	Thr	Arg	Lys	Tyr	Trp	Cys	Arg	Gln	Gly	Ala	42
43	Gln	Gly	Arg	Cys	Thr	Thr	Leu	Ile	Ser	Ser	52	Glu	Gly	Tyr	Val	Ser	Asp	Asp	Tyr	Val	Gly	62
63	Arg	Ala	Asn	Leu	Thr	Asn	Phe	Pro	Glu	Ser	72	Gly	Thr	Phe	Val	Ser	Asp	Ile	Ser	His	Leu	82
83	Thr	His	Lys	Asp	Ser	Gly	Arg	Tyr	Lys	Cys	92	Gly	Leu	Gly	Ile	Val	Ser	Arg	Gly	Leu	Asn	102
103	Phe	Asp	Val	Ser	Leu	Glu	Val	Ser	Gln	Asp	112	Pro	Ala	Gln	Ala	Ser	His	Ala	His	Val	Tyr	122
123	Thr	Val	Asp	Leu	Gly	Arg	Thr	Val	Thr	Ile	132	Asn	Cys	Pro	Phe	Thr	Arg	Ala	Asn	Ser	Glu	142
143	Lys	Arg	Lys	Ser	Leu	Cys	Lys	Lys	Thr	Ile	152	Gln	Asp	Cys	Phe	Gln	Val	Val	Asp	Ser	Thr	162
163	Gly	Tyr	Val	Ser	Asn	Ser	Tyr	Lys	Asp	Arg	172	Ala	His	Ile	Ser	Ile	Leu	Gly	Thr	Asn	Thr	182
183	Gln	Val	Phe	Ser	Val	Val	Ile	Asn	Arg	Val	192	Lys	Leu	Ser	Asp	Ala	Gly	Met	Tyr	Val	Cys	202
203	Glu	Ala	Gly	Asp	Asp	Ala	Lys	Ala	Asp	Lys	212	Ile	Asn	Ile	Asp	Leu	Gln	Val	Leu	Glu	Pro	222
223	Glu	Pro	Glu	Leu	Val	Tyr	Gly	Asp	Leu	Arg	232	Ser	Ser	Val	Thr	Phe	Asp	Cys	Ser	Leu	Gly	242
243	Pro	Glu	Val	Ala	Asn	Val	Pro	Lys	Phe	Leu	252	Cys	Gln	Lys	Lys	Asn	Gly	Gly	Ala	Cys	Asn	262
263	Val	Val	Ile	Asn	Thr	Leu	Gly	Lys	Lys	Ala	272	Gln	Asp	Phe	Gln	Gly	Arg	Ile	Val	Ser	Val	282
283	Pro	Lys	Asp	Asn	Gly	Val	Phe	Ser	Val	His	292	Ile	Thr	Ser	Leu	Arg	Lys	Glu	Asp	Ala	Gly	302
303	Arg	Tyr	Val	Cys	Gly	Ala	Gln	Pro	Glu	Gly	312	Glu	Pro	Gln	Asp	Gly	Trp	Pro	Val	Gln	Ala	322
323	Trp	Gln	Leu	Phe	Val	Asn	Glu	Glu	Thr	Ala	332	Ile	Pro	Ala	Ser	Pro	Ser	Val	Val	Lys	Gly	342
343	Val	Arg	Gly	Gly	Ser	Val	Thr	Val	Ser	Cys	352	Pro	Tyr	Asn	Pro	Lys	Asp	Ala	Asn	Ser	Ala	362
363	Lys	Tyr	Trp	Cys	His	Trp	Glu	Glu	Ala	Gln	372	Asn	Gly	Arg	Cys	Pro	Arg	Leu	Val	Glu	Ser	382
383	Arg	Gly	Leu	Ile	Lys	Gln	Gln	Tyr	Glu	Gly	392	Arg	Leu	Ala	Leu	Leu	Thr	Glu	Pro	Gly	Asn	402
403	Gly	Thr	Tyr	Thr	Val	Ile	Leu	Asn	Gln	Leu	412	Thr	Asp	Gln	Asp	Thr	Gly	Phe	Tyr	Trp	Cys	422
423	Val	Thr	Asp	Gly	Asp	Thr	Arg	Trp	Ile	Ser	432	Thr	Val	Glu	Leu	Lys	Val	Val	Gln	Gly	Glu	442
443	Pro	Ser	Leu	Lys	Val	Pro	Lys	Asn	Val	Thr	452	Ala	Trp	Leu	Gly	Glu	Pro	Leu	Lys	Leu	Ser	462
463	Cys	His	Phe	Pro	Cys	Lys	Phe	Tyr	Ser	Phe	472	Glu	Lys	Tyr	Trp	Cys	Lys	Trp	Ser	Asn	Arg	482
483	Gly	Cys	Ser	Ala	Leu	Pro	Thr	Gln	Asn	Asp	492	Gly	Pro	Ser	Gln	Ala	Phe	Val	Ser	Cys	Asp	502
503	Gln	Asn	Ser	Gln	Val	Val	Ser	Leu	Asn	Leu	512	Asp	Thr	Val	Thr	Lys	Glu	Asp	Glu	Gly	Trp	522
523	Tyr	Trp	Cys	Gly	Val	Lys	Glu	Gly	Pro	Arg	532	Tyr	Gly	Glu	Thr	Ala	Ala	Val	Tyr	Val	Ala	542
543	Val	Glu	Ser	Arg	Val	Lys	Gly	Ser	Gln	Gly	552	Ala	Lys	Gln	Val	Lys	Ala	Ala	Pro	Ala	Gly	562
563	Ala	Ala	Ile	Gln	Ser	Arg	Ala	Gly	Glu	Ile	572	Gln	Asn	Lys	Ala	Leu	Leu	Asp	Pro	Ser	Phe	582
583	Phe	Ala	Lys	Glu	Ser	Val	Lys	Asp	Ala	Ala	592	Gly	Gly	Pro	Gly	Ala	Pro	Ala	Asp	Pro	Gly	602
603	Arg	Pro	Thr	Gly	Tyr	Ser	Gly	Ser	Ser	Lys	612	Ala	Leu	Val	Ser	Thr	Leu	Val	Pro	Leu	Ala	622
623	Leu	Val	Leu	Val	Ala	Gly	Val	Val	Ala	Ile	632	Gly	Val	Val	Arg	Ala	Arg	His	Arg	Lys	Asn	642
643	Val	Asp	Arg	Ile	Ser	Ile	Arg	Ser	Tyr	Arg	652	Thr	Asp	Ile	Ser	Met	Ser	Asp	Phe	Glu	Asn	662
663	Ser	Arg	Asp	Phe	Glu	Gly	Arg	Asp	Asn	Met	672	Gly	Ala	Ser	Pro	Glu	Ala	Gln	Glu	Thr	Ser	682
683	Leu	Gly	Gly	Lys	Asp	Glu	Phe	Ala	Thr	Thr	692	Thr	Glu	Asp	Thr	Val	Glu	Ser	Lys	Glu	Pro	702
703	Lys	Lys	Ala	Lys	Arg	Ser	Lys	Glu	Glu	Thr	712	Ala	Asp	Glu	Ala	Phe	Thr	Thr	Phe	Leu	Pro	722
723	Gln	Ala	Lys	Asn	Leu	Ala	Ser	Ala	Ala	Thr	732	Gln	Asn	Gly	Pro	Thr	Glu	Ala				739

Figure 1.18 – Primary sequence of the polymeric immunoglobulin receptor (PIGR; UniProt accession number P81265, 01.02.2021). The following residues are highlighted on the sequence: **Asn** = N-glycosylated asparagine; **Cys** = cysteine. The first 18 amino acids depicted in italic represent the signal peptide. Numbering starts at the N-terminus of the mature protein. Highlighted in yellow is fragment 1-581 of the mature protein representing the secretory component (depicted in yellow in Figure 1.13C and D).

Immunoglobulin glycosylation modulates, amongst others, the function, interactions and activity of the immunoglobulins (Arnold *et al.*, 2007, Jennewein and Alter, 2017). Furthermore, secretory immunoglobulin glycans, including the glycans of the secretory component, can bind pathogens (Arnold *et al.*, 2007). In pathogen binding, the glycans can act as decoys, preventing pathogen adhesion to

1	Xaa	Ser	Glu	Thr	Ser	Pro	Ser	Ile	Phe	Pro	10	Leu	Ser	Leu	Gly	Asn	Asn	Asp	Pro	Ala	Gly	20
21	Gln	Val	Val	Ile	Gly	Cys	Leu	Val	Gln	Gly	30	Phe	Phe	Pro	Ser	Ala	Pro	Leu	Ser	Val	Thr	40
41	Trp	Asn	Gln	Asn	Gly	Asp	Ser	Val	Ser	Val	50	Arg	Asn	Phe	Pro	Ala	Val	Leu	Ala	Gly	Ser	60
61	Leu	Tyr	Thr	Met	Ser	Ser	Gln	Leu	Thr	Leu	70	Pro	Ala	Ser	Leu	Cys	Pro	Lys	Gly	Gln	Ser	80
81	Val	Thr	Cys	Gln	Val	Gln	His	Leu	Ser	Lys	90	Ala	Ser	Lys	Thr	Val	Ala	Val	Pro	Cys	Ile	100
101	Ile	Gln	Asp	Ser	Ser	Ser	Cys	Cys	Val	Pro	110	Asn	Cys	Glu	Pro	Ser	Leu	Ser	Val	Gln	Pro	120
121	Pro	Ala	Leu	Glu	Asp	Leu	Leu	Leu	Gly	Ser	130	Asn	Ala	Ser	Leu	Thr	Cys	Thr	Leu	Ser	Gly	140
141	Leu	Lys	Ser	Ala	Glu	Gly	Ala	Ser	Phe	Thr	150	Trp	Asn	Pro	Thr	Gly	Gly	Lys	Thr	Ala	Val	160
161	Gln	Gly	Ser	Pro	Lys	Arg	Asp	Ser	Cys	Gly	170	Cys	Tyr	Ser	Val	Ser	Ser	Val	Leu	Pro	Gly	180
181	Cys	Ala	Asp	Pro	Trp	Asn	Ser	Gly	Gln	Thr	190	Phe	Ser	Cys	Ser	Val	Thr	His	Pro	Glu	Ser	200
201	Lys	Ser	Ser	Leu	Thr	Ala	Thr	Ile	Lys	Lys	210	Asp	Leu	Gly	Asn	Thr	Phe	Arg	Pro	Gln	Val	220
221	His	Leu	Leu	Pro	Pro	Ser	Glu	Glu	Leu	230	Ala	Leu	Asn	Glu	Leu	Val	Thr	Leu	Thr	Cys	240	
241	Leu	Val	Arg	Gly	Phe	Ser	Pro	Lys	Glu	Val	250	Leu	Val	Arg	Trp	Leu	Gly	Asn	Gln	Glu	260	
261	Leu	Pro	Arg	Glu	Lys	Tyr	Pro	Leu	Trp	Gly	270	Pro	Leu	Pro	Glu	Ala	Gln	Ser	Val	Thr	280	
281	Thr	Phe	Ala	Val	Thr	Ser	Val	Leu	Arg	Val	290	Asp	Ala	Glu	Val	Trp	Lys	Gln	Gly	Asp	Thr	300
301	Phe	Ser	Cys	Met	Val	Gly	His	Glu	Ala	Leu	310	Pro	Leu	Ala	Phe	Thr	Gln	Lys	Thr	Ile	Asp	320
321	Arg	Leu	Ala	Gly	Lys	Pro	Thr	His	Val	Asn	330	Val	Ser	Val	Val	Met	Ser	Glu	Val	Asp	Gly	340
341	Val	Cys	Tyr								343											

Figure 1.19 – Primary sequence of IgA constant heavy chain allele 1 (IMGT accession number KT723008). The following residues are highlighted on the sequence: Asn = N-glycosylated asparagine; Cys = cysteine; Xxx = heavy constant 1 region; Xxx = hinge + heavy constant 2 regions; Xxx = heavy constant 3 region; numbering starts at the N-terminus of the constant region of the heavy chain.

1	Xaa	Gly	Glu	Ser	His	Pro	Lys	Val	Phe	Pro	10	Leu	Val	Ser	Cys	Val	Ser	Ser	Pro	Ser	Asp	20
21	Glu	Ser	Thr	Val	Ala	Leu	Gly	Cys	Leu	Ala	30	Arg	Asp	Phe	Val	Pro	Asn	Ser	Val	Ser	Phe	40
41	Ser	Trp	Lys	Phe	Asn	Asn	Ser	Thr	Val	Ser	50	Ser	Glu	Arg	Phe	Thr	Phe	Pro	Glu	Val	60	
61	Leu	Arg	Asp	Gly	Leu	Trp	Ser	Ala	Ser	Ser	70	Gln	Val	Val	Leu	Pro	Ser	Ser	Ser	Ala	Phe	80
81	Gln	Gly	Pro	Asp	Asp	Tyr	Leu	Val	Cys	Glu	90	Val	Gln	His	Pro	Lys	Gly	Gly	Lys	Thr	Val	100
101	Gly	Thr	Val	Arg	Val	Ile	Ala	Thr	Lys	Ala	110	Glu	Val	Leu	Ser	Pro	Val	Val	Ser	Val	Phe	120
121	Val	Pro	Pro	Arg	Asn	Ser	Leu	Ser	Gly	Asp	130	Gly	Asn	Ser	Lys	Ser	Ser	Leu	Ile	Cys	Gln	140
141	Ala	Thr	Asp	Phe	Ser	Pro	Lys	Gln	Ile	Ser	150	Leu	Ser	Trp	Phe	Arg	Asp	Gly	Lys	Arg	Ile	160
161	Val	Ser	Gly	Ile	Ser	Glu	Gly	Gln	Val	Glu	170	Thr	Val	Gln	Ser	Ser	Pro	Ile	Thr	Phe	Arg	180
181	Ala	Tyr	Ser	Met	Leu	Thr	Ile	Thr	Glu	Arg	190	Asp	Trp	Leu	Ser	Gln	Asn	Ala	Tyr	Thr	Cys	200
201	Gln	Val	Glu	His	Asn	Lys	Glu	Thr	Phe	Gln	210	Lys	Asn	Val	Ser	Ser	Ser	Cys	Asp	Val	Ala	220
221	Pro	Pro	Ser	Pro	Ile	Gly	Val	Phe	Thr	Ile	230	Pro	Pro	Ser	Phe	Ala	Asp	Ile	Phe	Leu	Thr	240
241	Lys	Ser	Ala	Lys	Leu	Ser	Cys	Leu	Val	Thr	250	Asn	Leu	Ala	Ser	Tyr	Asp	Gly	Leu	Asn	Ile	260
261	Ser	Trp	Ser	Arg	Gln	Asn	Ala	Lys	Ala	Leu	270	Glu	Thr	His	Thr	Tyr	Phe	Glu	Arg	His	Leu	280
281	Asn	Asp	Thr	Phe	Ser	Ala	Arg	Gly	Glu	Ala	290	Ser	Val	Cys	Ser	Pro	Glu	Asp	Trp	Glu	Ser	300
301	Glu	Glu	Phe	Thr	Cys	Thr	Val	Ala	His	Ser	310	Asp	Leu	Pro	Phe	Pro	Glu	Lys	Asn	Ala	Val	320
321	Ser	Lys	Pro	Lys	Asp	Val	Ala	Met	Lys	Pro	330	Pro	Ser	Val	Tyr	Leu	Leu	Pro	Pro	Thr	Arg	340
341	Glu	Gln	Leu	Ser	Leu	Arg	Glu	Ser	Ala	Ser	350	Val	Thr	Cys	Leu	Val	Lys	Ala	Phe	Ala	Pro	360
361	Ala	Asp	Val	Phe	Val	Gln	Trp	Leu	Gln	Arg	370	Gly	Glu	Pro	Val	Thr	Lys	Ser	Lys	Tyr	Val	380
381	Thr	Ser	Ala	Arg	Ala	Pro	Glu	Pro	Gln	Asp	390	Pro	Ser	Val	Val	Tyr	Phe	Val	His	Ser	Ile	400
401	Leu	Thr	Val	Ala	Glu	Glu	Asp	Trp	Ser	Lys	410	Gly	Glu	Thr	Tyr	Thr	Cys	Val	Val	His	Glu	420
421	Ala	Leu	Pro	His	Met	Val	Thr	Glu	Arg	Thr	430	Val	Asp	Lys	Ser	Thr	Gly	Lys	Pro	Thr	Leu	440
441	Tyr	Asn	Val	Ser	Leu	Val	Leu	Ser	Asp	Thr	450	Ala	Ser	Thr	Cys	Tyr						455

Figure 1.20 – Primary sequence of IgM1 constant heavy chain allele 1 (IMGT accession number U63637). The following residues are highlighted on the sequence: Asn = N-glycosylated asparagine; Cys = cysteine; Xxx = heavy constant 1 region; Xxx = heavy constant 2 region; Xxx = heavy constant 3 region; Xxx = heavy constant 4 region; numbering starts at the N-terminus of the constant region of the heavy chain.

Table 1.9 – Differences in the amino acid sequences of the alleles of IgM1 heavy chain compared to IGHM1*01

Allele	Amino acid position																				Insertion between 418-419			
	7	15	31	34	50	59	79	106	107	108	116	190	197	218	219	267	277	319	357	383		384	388	393
IGHM1*01	Lys	Val	Arg	Val	Ser	Glu	Ala	Ile	Ala	Thr	Val	Arg	Ala	Asp	Val	Ala	Glu	Ala	Ala	Ala	Arg	Pro	Val	
IGHM1*02	Arg	Met	Gln	Met	Gly	Ala		Val	Thr	Pro	Ile	Lys	Val	Asn		Gly	Gly	Thr	Gly	Ser	Pro	Ser	del.	Gly
IGHM1*03	Arg	Met	Gln	Met		Ala	Thr	Val	Thr	Pro	Ile	Lys	Val		Ala	Gly			Gly	Ser	Pro		del.	Gly
IGHM1*04	Arg				Gly	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

1	Xaa	Gly	Glu	Ser	His	Pro	Arg	Val	Phe	Pro	10	Leu	Val	Ser	Cys	Val	Ser	Ser	Pro	Ser	Asp	20
21	Glu	Ser	Thr	Val	Ala	Leu	Gly	Cys	Leu	Ala	30	Arg	Asp	Phe	Val	Pro	Asn	Ser	Val	Ser	Phe	40
41	Ser	Trp	Lys	Phe	Asn	Asn	Ser	Thr	Val	Ser	50	Ser	Glu	Arg	Phe	Trp	Thr	Phe	Pro	Glu	Val	60
61	Leu	Arg	Asp	Gly	Leu	Trp	Ser	Ala	Ser	Ser	70	Gln	Val	Val	Leu	Ala	Phe	Ala	Phe	Arg	80	
81	Gln	Gly	Pro	Asp	Asp	Tyr	Leu	Val	Cys	Glu	90	Val	Gln	His	Pro	Lys	Gly	Gly	Lys	Thr	Val	100
101	Gly	Thr	Val	Arg	Val	Ile	Ala	Thr	Lys	Ala	110	Glu	Val	Leu	Ser	Pro	Val	Val	Ser	Val	Phe	120
121	Val	Pro	Pro	Arg	Asn	Ser	Leu	Ser	Gly	Asp	130	Gly	Asn	Ser	Lys	Ser	Ser	Leu	Ile	Cys	Gln	140
141	Ala	Thr	Asp	Phe	Ser	Pro	Lys	Gln	Ile	Ser	150	Leu	Ser	Trp	Phe	Arg	Asp	Gly	Lys	Arg	Ile	160
161	Val	Ser	Gly	Ile	Ser	Glu	Gly	Gln	Val	Glu	170	Thr	Val	Gln	Ser	Ser	Pro	Ile	Thr	Phe	Arg	180
181	Ala	Tyr	Ser	Met	Leu	Thr	Ile	Thr	Glu	Arg	190	Asp	Trp	Leu	Ser	Gln	Asn	Val	Tyr	Thr	Cys	200
201	Gln	Val	Glu	His	Asn	Lys	Glu	Thr	Phe	Gln	210	Lys	Asn	Val	Ser	Ser	Ser	Cys	Asp	Val	Ala	220
221	Pro	Pro	Ser	Pro	Ile	Gly	Val	Phe	Thr	Ile	230	Pro	Pro	Ser	Phe	Ala	Asp	Ile	Phe	Leu	Thr	240
241	Lys	Ser	Ala	Lys	Leu	Ser	Cys	Leu	Val	Thr	250	Asn	Leu	Ala	Ser	Tyr	Asp	Gly	Leu	Asn	Ile	260
261	Ser	Trp	Ser	Arg	Gln	Asn	Gly	Lys	Ala	Leu	270	Glu	Thr	His	Thr	Phe	Glu	Arg	His	Leu	280	
281	Asn	Asp	Thr	Phe	Ser	Ala	Arg	Gly	Glu	Ala	290	Ser	Val	Cys	Ser	Glu	Asp	Trp	Glu	Ser	Gly	300
301	Glu	Glu	Phe	Thr	Cys	Thr	Val	Ala	His	Ser	310	Asp	Leu	Pro	Phe	Pro	Glu	Lys	Asn	Ala	Val	320
321	Ser	Lys	Pro	Lys	Asp	Val	Ala	Met	Lys	Pro	330	Pro	Ser	Val	Tyr	Leu	Leu	Pro	Pro	Thr	Arg	340
341	Glu	Gln	Leu	Ser	Leu	Arg	Glu	Ser	Ala	Ser	350	Val	Thr	Cys	Leu	Val	Lys	Gly	Phe	Ala	Pro	360
361	Ala	Asp	Val	Phe	Val	Gln	Trp	Leu	Gln	Arg	370	Gly	Glu	Pro	Val	Thr	Lys	Ser	Lys	Tyr	Val	380
381	Thr	Ser	Ser	Pro	Ala	Pro	Glu	Pro	Gln	Asp	390	Pro	Ser	Val	Tyr	Phe	Val	His	Ser	Ile	Leu	400
401	Thr	Val	Ala	Glu	Glu	Asp	Trp	Ser	Lys	Gly	410	Glu	Thr	Tyr	Thr	Cys	Val	Val	Gly	His	Glu	420
421	Ala	Leu	Pro	His	Met	Val	Thr	Glu	Arg	Thr	430	Val	Asp	Lys	Ser	Thr	Gly	Lys	Pro	Thr	Leu	440
441	Tyr	Asn	Val	Ser	Leu	Val	Leu	Ser	Asp	Thr	450	Ala	Ser	Thr	Cys	Tyr						455

Figure 1.21 – Primary sequence of IgM2 constant heavy chain allele 1 (IMGT accession number AY230207, 01.02.2021). The following residues are highlighted on the sequence: Asn = N-glycosylated asparagine; Cys = cysteine; Xxx = heavy constant 1 region; Xxx = heavy constant 2 region; Xxx = heavy constant 3 region; Xxx = heavy constant 4 region; numbering starts at the N-terminus of the constant region of the heavy chain.

the intestinal mucosa, or they can play a role in the agglutination and subsequent removal of pathogens, as is the case of sIgM (Arnold *et al.*, 2007). The presence of

multiple Fab domains as well as the richness of glycosylation in the secretory immunoglobulins make them powerful agents against pathogenic bacteria, viruses and toxins (Strugnell and Wijburg, 2010). The biological functionality of bovine milk immunoglobulins is interesting both from the perspective of animal health management, as well as bioactive components in human nutrition.

Table 1.10 – Differences in the amino acid sequences of alleles of IGHM2*01 and IGHM2*02

Allele	Amino acid position		
	106	177	319
IGHM2*01	Ile	Ile	Ala
IGHM2*02	Val	Val	Thr

Mass spectrometry for the analysis of milk proteins

While mass spectrometry in itself is over a century old, with the first crude mass spectrograph developed by Thomson (1897), its earliest application to the study of peptides, and implicitly proteins, is reported by Biemann *et al.* (1959). Proteomics work was first described by Wasinger *et al.* (1995), with the concept of the proteome being introduced by Wilkins *et al.* (1996) as the protein complement expressed by a genome.

Peptide-centric mass spectrometry

Initial proteomics studies relied on the two-dimensional electrophoretic (2-DE) separation of the proteins, followed by in-gel digestion and subsequent analysis of the released peptides by mass spectrometry (Tyers and Mann, 2003, Yates III, 2011). This technique became popular in the 1990s and has since been used also for the study of the milk proteome (Le *et al.*, 2017). The major downside of the technique comes from the fact that in systems with high dynamic range, such as bovine milk where over 90% of the proteome is made up by the six major proteins (α_{s1} -, α_{s2} -, β - and κ -casein, α -lactalbumin and β -lactoglobulin), there are still hundreds of minor proteins present in the sample, the majority of which will not be visualised on a gel. Liquid phase separation such as liquid chromatography and capillary electrophoresis have become the go-to techniques for protein separation in proteomics studies (Issaq *et al.*, 2002). The advantages of these techniques compared to gel electrophoresis include greater sensitivity, superior dynamic range, easier automation and shorter experimental times (O'Donnell *et al.*, 2004).

For peptide-centric mass spectrometry studies, trypsin has been the golden standard protease (Tsiatsiani and Heck, 2015). Trypsin cleaves exclusively at the C-terminal of Arg and Lys residues (Olsen *et al.*, 2004), both of which are basic amino

acid residues. Consequently, tryptic peptides carry at least two positive charges: the free amine at the N-terminus of the peptide, and the C-terminal basic amino acid residue. This charge load renders tryptic peptides a perfect match for analysis by tandem mass spectrometry in positive mode (Vandermarliere *et al.*, 2013). Furthermore, the high cleavage specificity of trypsin translates into low complexity of database searches.

As such, trypsin is a suitable protease for studies mapping the proteome of a sample. In this case, each individual protein is identified based on peptide mass fingerprinting, where one or more peptides are uniquely matched through database search to the sequence of a protein. This approach does not require full sequence coverage of the amino acid backbone, and only relies on the identification of short peptides, typically composed of 7-35 amino acid residues corresponding to a mass range of 600-4000 Da (Vandermarliere *et al.*, 2013). While the choice of trypsin as the protease for, *e.g.*, mapping of the bovine milk proteome may be appropriate, a challenge in obtaining representative results comes from the choice of protein database used for identification. Although the bovine genome was estimated to contain ~22,000 protein-encoding genes (Elsik *et al.*, 2009), the reviewed Swiss-Prot database for the *Bos taurus* species, which is one of the most commonly-used databases for protein identification, only contains ~6000 protein entries. The consequence of an incomplete database is obvious, in that proteins whose sequences are not in the database will also not be identified. The full UniProt database for the *Bos taurus* species contains ~37,500 sequences, which include redundancies, *i.e.*, several protein sequences for the same gene. The presence of redundant sequences in the database influences both the identification and the peptide intensity-based quantification of the proteins, and may lead to non-representative results. Furthermore, sequences such as those of the bovine immunoglobulins, which are core components of the bovine colostrum and are present in bovine milk throughout lactation (Gazi *et al.*, 2023), are theoretically present in UniProt, but they are not annotated and have neither a protein nor a gene name. Consequently, bovine immunoglobulins cannot be identified at all by searching proteomics data against the bovine UniProt database. In supplementary Figure S3.14 from **Chapter 3** we illustrate the impact of the protein database on the identification and label-free quantification of proteins from a colostrum sample. In **Chapter 3** we also describe the means by which we optimised the bovine protein database, thereby overcoming the limitations of incomplete, poorly annotated and highly redundant bovine protein databases.

While trypsin has many merits as the protease of choice for proteomics studies, it is not the suitable protease for every research question. For research questions

requiring coverage of specific regions of the protein backbone, such as regions that differ between genetic variants or those that harbour post-translational modifications, or for research questions requiring a high protein sequence coverage, trypsin on its own is often not suitable. The frequencies at which lysine and arginine residues occur in the proteome results in more than 50% of the tryptic peptides having lengths of 6 amino acid residues or shorter (Swaney *et al.*, 2010). Such peptides are too short to be uniquely assigned to any protein. On the other extreme, proteins that are highly hydrophobic contain regions devoid of lysine and arginine residues, which results in poor sequence coverage when analysed from a tryptic digest. This is the case with membrane proteins (Eichacker *et al.*, 2004), which include the milk fat globule membrane (MFGM) proteins, but also with caseins (Lucey and Horne, 2018), in which hydrophobic interactions play a crucial role in complex formation. In particular, as observed from Figure 1.5 and Figure 1.7, the amino acid sequences of β - and κ -casein contain tryptic peptides as long as 49-56 residues. As shown in Figure S2.4 in **Chapter 2**, peptide-centric mass spectrometry from a tryptic digest resulted in very poor or no coverage of these protein regions. The amino acid differences between most of the genetic variants of β -casein occur on the long tryptic peptides of Ile₄₉-Lys₉₇ and Tyr₁₁₄-Lys₁₆₉ (Table 1.2). In the case of κ -casein, the amino acid differences between most of the genetic variants (Table 1.3), as well as the post-translational modifications of phosphorylation and *O*-glycosylation mostly occur in the para- κ -casein region of the protein, also on a long tryptic peptide, *i.e.*, Thr₁₁₇-Val₁₆₉ (Bijl *et al.*, 2020). These long tryptic peptides are outside the range of sequence lengths that can be analysed in typical peptide-centric proteomics experiments. Therefore, for the investigation of these protein regions, trypsin (on its own) is not the suitable protease. One additional problem with tryptic digestion is when the tryptic cleavage sites are chemically modified, inhibiting cleavage by the protease, as is the case with glycation of the proteins. Glycation, particularly in the form of lactosylation, is a common chemical modification that occurs on the milk proteins primarily during thermal processing and subsequent storage (van Boekel, 1998). Glycation affects the free amino residues present on proteins and peptides, particularly the N-terminus, lysine and arginine residues. These then interact with reducing sugars, *e.g.*, the lactose that is abundantly present in milk. Due to lysine and arginine residues being both trypsin cleavage sites, as well as glycation targets, glycated proteins exhibit suboptimal digestibility by trypsin in particular (van Lieshout *et al.*, 2020). This results in increased numbers of missed cleavages, hindering downstream proteomics analysis (Deng *et al.*, 2017). Despite the decreased performance of trypsin towards the hydrophobic casein and glycated proteins, most published studies investigating milk protein glycation using peptide-centric

mass spectrometry still rely on tryptic digestion of the proteins. In studies where trypsin on its own is not a suitable protease, alternative proteases or combinations thereof may be considered. This particularly includes proteases with cleavage specificity other than or complementary to that of trypsin, such as GluC (cleaves C-terminal of Glu and Asp residues), AspN (cleaves N-terminal of Asp residues), α -lytic protease (cleaves C-terminal of Thr, Ala, Ser and Val residues) or chymotrypsin (cleaves C-terminal of Phe, Trp and Tyr residues). In **Chapter 2** we describe a multi-protease approach that we developed and applied to overcome the limitations of trypsin for the analysis of glycation development on the six most abundant milk proteins. Figure S2.4 in **Chapter 2** illustrates the high sequence coverage obtained with the multi-protease approach for both the non-glycated (thermised milk) and glycated proteins (aged milk powder).

Protein-centric mass spectrometry

The introduction of soft ionization techniques, electrospray ionization in particular, during the late 1980s (Fenn *et al.*, 1990, Wong *et al.*, 1988) opened the door for the exploration of not only small molecules by mass spectrometry, but also of macromolecular biomolecules, including intact proteins (Fenn *et al.*, 1989). In fact, the importance of this development for furthering scientific research was embodied by the Nobel Prize in Chemistry being partially awarded to John B. Fenn in 2002 (Cho and Normile, 2002). John Fenn and Koichi Tanaka shared half of the Nobel Prize for their developments of electrospray ionization and soft laser desorption ionization techniques for mass spectrometry, respectively. The other half of the 2002 Nobel Prize in Chemistry was awarded to Kurt Wüthrich for the development of nuclear magnetic resonance (Cho and Normile, 2002), another core technique used for the study of proteins.

Mass spectrometry using electrospray ionization was soon after applied for the analysis of the major bovine milk proteins in intact form. Leonil *et al.* (1995) described one of the first instances of intact milk protein analysis, in which the samples were first denatured and reduced, followed by separation by reverse-phase high-performance liquid chromatography (RP-HPLC) coupled with electrospray ionization mass spectrometry (ESI-MS). This technique is still widely used today for the analysis of intact milk proteins, specifically for identifying genetic variation and protein phenotypes (Huppertz *et al.*, 2021, Miranda *et al.*, 2020, Nguyen *et al.*, 2020, Nilsson *et al.*, 2020), but also for determining the extent of post-translational modifications, *e.g.*, casein phosphorylation or κ -casein O-glycosylation (Miranda *et al.*, 2020, Nilsson *et al.*, 2020, Thesbjerg *et al.*, 2022), or process- and/or storage-induced glycation (Fuerer *et al.*, 2020).

The identification of milk proteoforms by LC-ESI-MS relies on the matching of experimental masses to theoretical masses of the intact proteoforms. This introduces room for error in cases of different proteoforms with nearly identical molecular weights. The κ -CN proteoform of genetic variant E, harbouring one phosphorylated residue (1P) and one *O*-glycosylated residue occupied with the HexNAcHexNeuAc acidic trisaccharide (N1H1S1), *i.e.*, κ -CN E-1P N1H1S1, has monoisotopic and average masses of 19651.75 and 19663.71 Da, respectively. The β -lg proteoform of genetic variant A glycosylated with a total of 8 monohexose (H8) residues (which in milk and dairy products likely occurs in the form of 4 lactosyl residues), *i.e.*, β -lg A H8 has monoisotopic and average masses of 19651.88 and 19664.24 Da, respectively. These proteoforms are the products of different genes. Both of these proteoforms are likely to occur simultaneously in a thermally processed and/or aged product. Proteoform identification by intact mass matching only will not distinguish between these two, and additional factors should be considered for assigning the appropriate proteoform. One important factor to consider is the retention time of the proteins when separated by RP-HPLC: the κ -CNs tend to elute at the beginning of the chromatogram, while the β -lgs are typically seen towards the end of the chromatogram (Bijl *et al.*, 2014, Fuerer *et al.*, 2020, Leonil *et al.*, 1995, Visser *et al.*, 1991). Genetic polymorphism and post-translational modifications such as phosphorylation, glycosylation or glycation can cause shifts in the retention time (Fuerer *et al.*, 2020, Marx *et al.*, 2013, Otvos Jr *et al.*, 1992, Steen *et al.*, 2006). However, related proteoforms, *i.e.*, products of the same gene, cluster together in the chromatogram, and the retention time differences between related proteoforms are considerably smaller than the retention times between different proteins, *e.g.*, κ -CN and β -lg. Nevertheless, chromatographic retention time cannot always be used to differentiate between proteoforms of similar mass. For example, κ -CN B-1P N1H1 (monoisotopic mass 19358.71 Da, average mass 19370.52 Da) and κ -CN E-3P N1 (monoisotopic mass 19358.54 Da, average mass 19370.27 Da), are different forms of κ -CN. Both proteoforms can plausibly occur in milk, and will elute at comparable retention times on RP-HPLC. Caseins, including κ -CN, occur in milk in a distribution of proteoforms, as can be observed in Figure 2.2 in **Chapter 2**. Various proteoforms of the caseins occur in individual milk samples, however, the diversity and heterogeneity of proteoforms is highest in bulk milk. In the case of κ -CN, every genetic variant is dominated by the 1P form, followed by 2P, and to a lesser extent by the glycosylated forms (Bijl *et al.*, 2020, Holland *et al.*, 2004). Distinguishing between exemplified proteoforms κ -CN B-1P N1H1 and κ -CN E-3P N1 can be done by looking at the proteoform distribution of each of the B and E genetic variants. If, *e.g.*, the sample contains a distribution of the B variants, but no other E variant may

be detected, then κ -CN B-1P N1H1 is likely to be the correct assignment. Furthermore, proteoform identification by matching of their experimental masses to theoretical masses does not distinguish between, *e.g.*, positional isomers, in which the same modification may be present on different amino acid residues, or between changes in the order of the amino acids in the protein sequence, or isobaric amino acid substitutions in the sequence. In such cases, proteoform identification by intact mass matching must be verified by additional methods that result in the sequencing of the proteoform.

Intact protein analysis by mass spectrometry is not limited to the detection of masses of intact proteins by LC-MS, as classically performed in the case of milk proteins. As with peptide-centric analyses (*i.e.*, bottom-up proteomics), whole proteins can also be analysed by tandem mass spectrometry, a technique often referred to as *top-down proteomics* (Catherman *et al.*, 2014). In top-down proteomics, the mass of the intact protein, *i.e.*, the precursor ion, is measured, after which the intact protein is fragmented. Identification of the protein is done by database search, where the protein fragment ions are matched against *in silico* generated ions based on the protein database. While the principles of bottom-up and top-down proteomics analyses are comparable, top-down proteomics proves to be more challenging. The large size difference between intact proteins and small peptides results in charging of the proteins to far higher charge states than peptides in the gas phase. Furthermore, fragmentation of intact proteins tends to be localized to certain regions of the protein, rather than efficiently distributed across the entire protein backbone. The less efficient fragmentation and resulting poorer protein sequence coverage by fragment ions, as well as the increased computational requirements constitute some of the bottlenecks of top-down proteomics. Nevertheless, top-down proteomics provides an additional layer of confidence for protein identification, compared to the simple matching of intact protein masses determined by LC-MS.

In **Chapter 2** we describe a hybrid approach that we use for the identification and monitoring of glycosylated proteoform development during the processing and storage of a skim milk powder, with focus on the six most abundant milk proteins. At intact protein level, we analysed the proteins both using LC-MS/MS and LC-MS. Using this approach, we verified protein identification by database search of the top-down proteomics data, after which we assigned proteoforms based on their intact mass. Next, we also analysed the proteins by bottom-up proteomics from proteolytic digests. This combined approach allowed us to identify the glycosylated proteoforms and the extent of glycosylation on the proteins, and also to localise the glycosylation sites on the protein backbone.

Glycoproteomics

Protein glycosylation is a versatile post-translational modification that bears a heavy impact on a multitude of biological processes and functions (Bagdonaite *et al.*, 2022, Varki, 2017, Wandall *et al.*, 2021). Glycosylation dictates correct protein folding and structure, regulates signalling and protein interactions, and modulates protein activity. Changes in protein glycosylation occur between health and disease states (Ohtsubo and Marth, 2006, Reily *et al.*, 2019, Stowell *et al.*, 2015), and are also hallmarks of, *e.g.*, pregnancy (Bondt *et al.*, 2014, Jansen *et al.*, 2016, Ruhaak *et al.*, 2014) and lactation (Dingess *et al.*, 2021, Gazi *et al.*, 2023, Goonatilleke *et al.*, 2019, Takimori *et al.*, 2011). Unlike protein biosynthesis, protein glycosylation is a non-templated process, with no underlying blueprint existing to predict glycan composition and structure. Glycan composition refers to the monosaccharide makeup of the glycan, including the number of *N*-acetylhexosamine (HexNAc), hexose (Hex), fucose (Fuc) and sialic acid residues of which the glycan is built. At a deeper level of information, glycan structure refers to the nature and order of the monosaccharide residues, and linkage specificity between the monosaccharide residues. Structurally different glycan isoforms of identical mass and composition can simultaneously occur in biological samples, depending on the glycosylation machinery that is present. The thereby resulting complexity and diversity of glycosylation also raises challenges in its analysis.

Mass spectrometry methods for glycoprotein analysis can be classified into three categories, *i.e.*, 1) enzymatic or chemical release of the glycans from the glycoproteins and analysis thereof, 2) proteolytic digestion of the glycoproteins and analysis of the resulting glycopeptides, and 3) analysis of intact glycoproteins (Čaval, 2019, Yang *et al.*, 2017). The analysis of released glycans, also referred to as glycomics, can generate very detailed insights regarding glycan composition, structure and isomers. Downsides of this approach include that the identified glycans cannot be traced back to specific sites on the proteins, or even to which protein they originate from, thus providing no information on glycosylation macro-heterogeneity (glycosylation site occupancy), micro-heterogeneity (variation in glycan species) or meta-heterogeneity (variation in glycosylation across multiple sites of a given protein) (Čaval *et al.*, 2021). Glycoprotein analysis at glycopeptide level overcomes a great deal of the limitations of the glycomics approach, generating protein- and site-specific details on glycosylation, providing information on site occupancy and glycan diversity. However, the level of detail that can be gained on glycoforms from the analysis of glycopeptides is lower than that obtained from the glycomics approach. Nevertheless, mass spectrometry analysis at glycopeptide level is a fast-developing field that has been in the spotlight these last

few years. The third tier of mass spectrometric approaches in glycoproteomics is focused on the analysis at intact protein level. Next to the previously-mentioned two approaches, the analysis of intact glycoproteins can provide information on proteoform profiling and distribution, relative abundances, and stoichiometry, which is entirely lacking from the glycan- or glycopeptide-level approaches. However, very limited information on glycan site localization or glycan structures can be obtained from the intact protein-based approach. Overall, each of these different approaches presents a different set of advantages and disadvantages. To overcome the limitations of these different techniques, a hybrid approach, that combines two or even all three of these techniques, can be used to provide a holistic understanding of the glycoprotein(s) of interest.

With considerable developments in the field over the past decade, glycoproteomics has emerged as an important subdiscipline of proteomics, and a powerful tool in glycobiology (Bagdonaite *et al.*, 2022, Oliveira *et al.*, 2021, Yang *et al.*, 2017). Of the three techniques described above, the mass spectrometric analysis at glycopeptide level is currently the most widely-used one, as it can provide the greatest amount of information on its own when compared to the other two techniques.

While until recently, glycoproteomics studies of milk proteins still heavily relied on the analysis of released glycans (Cao *et al.*, 2019, Takimori *et al.*, 2011, Valk-Weeber *et al.*, 2020, Xiao *et al.*, 2022), mass spectrometric analysis at glycopeptide level is currently in its early stages and is becoming the preferred technique for the analysis of milk glycoproteins (Dingess *et al.*, 2021, Gazi *et al.*, 2023, Guan *et al.*, 2023, Kumar *et al.*, 2022, Qu *et al.*, 2021, Zhu *et al.*, 2020).

LC-MS/MS analyses at glycopeptide level imply chromatographic separation, followed by mass analysis of intact (glyco)peptides, fragmentation of selected precursor ions, and acquisition of the fragmentation patterns for peptide identification, and glycan composition and structure characterization (Reiding *et al.*, 2018). The commonly-used collision-induced dissociation causes fragmentation of peptide backbones into b- and y-ions, representing the N- and C-terminal fragments of the peptide, respectively. In the case of glycopeptides, preferential fragmentation of the glycan structures occurs, as a consequence of these linkages being more labile than the peptide bonds (Reiding *et al.*, 2018). The resulting glycan fragment ions are the B-ions of glycopeptide fragmentation, whereas the intact peptide backbone carrying the remaining glycan fragment represents the Y-ion. B-ions are also commonly referred to as oxonium ions. The fragmentation of glycans provides information on their composition and structure. Varying the collision energy levels in higher-energy collisional dissociation (HCD), a process called

stepping-collision energy HCD (stepped-HCD), leads to the fragmentation of not only the labile glycosidic bonds, but of also the peptide bonds, resulting in the formation of b and y ions that can be used for peptide sequencing and identification. Other fragmentation techniques, such as electron transfer dissociation (ETD), primarily fragment the peptide backbone resulting in c and z ions, while leaving the glycan intact. This type of fragmentation can be used for site localization of the glycan on the peptide sequence. Electron-transfer/higher-energy collisional dissociation (ETHcD) is a type of fragmentation that combines HCD and ETD, generating B and Y ions, and complementary b and y, and c and z ions, providing information that can be used both for glycan and peptide sequencing and identification, and for glycan site localization. During hybrid fragmentation, a longer time is spent in fragmenting a selected precursor, while conventional HCD takes place in a shorter time, thereby allowing for a higher number of precursors to be fragmented. To generate a holistic view of the sample and maximise both the numbers of glycosylated and non-glycosylated peptides identified, optimal MS/MS fragmentation methods would consist of conventional HCD fragmentation of non-glycosylated peptides, combined with hybrid fragmentation of glycosylated peptides. To achieve this, the lability of the glycosidic linkages to HCD can be taken advantage of and used for triggering hybrid types of dissociation that result in fragmentation of glycopeptides optimal for glycan and peptide identification, as well as glycan localization. Figure S3.15 from **Chapter 3** exemplifies oxonium ion traces that result from HCD fragmentation during the LC-MS/MS analysis of proteolytically digested bovine colostrum and mature milk samples. As we have described in our methods in **Chapters 3** and **4**, we used the detection of these oxonium ions in our methods to trigger hybrid fragmentation, such as stepped-energy HCD and ETHcD, which favoured glycopeptide analysis in our samples of interest. In **Chapter 3** we used a glycoproteomics workflow on tryptic digests of IgG captured from the mammary secretions of four individual cows during lactation starting from bovine colostrum at 0.5 d to mature milk at 28 d after calving. Using this approach, we characterised the lactational dynamics of the glycosylation, which proved to be diverse and heterogeneous, and identified NeuAc sialylation as the hallmark of colostrum IgG glycosylation. In **Chapter 4** we looked at the general changes in the *N*-glycoproteome of the same sample set from the four individual cows, identifying the main protein contributors to the glycoproteome at different time points during lactation. From the general glycoproteome, we further focused our study on α -lactalbumin, a critical component of milk and a protein abundantly present in the proteome at every investigated time point. As is the case of the previously-investigated IgG, α -lactalbumin glycosylation also showed lactational dynamic changes, albeit not to the same extent as what we found in the case of IgG.

Our glycoproteomics approach allowed us to identify structurally isomeric glycans, which despite the same composition, contained *N*-acetyllactosamine (LacNAc) antennae and bisecting *N*-acetylglucosamine (GlcNAc) in IgG, but *N*, *N'*-diacetyllactosamine (LacdiNAc) antennae and no bisection in α -lactalbumin.

In summary, no one method can answer all research questions. Mass spectrometry and proteomics are very broad fields of research. As specified in this introduction, every different approach leads to different insights, whereas hybrid approaches may provide a more holistic picture. However, the choices of analytical samples, materials, *e.g.*, protease for bottom-up digestion, separation technique, analytical tier, *i.e.*, released post-translational modification, peptides, intact proteins or native protein complexes, fragmentation method, *i.e.*, HCD, or stepped-HCD or EThcD, each and all have an impact on the results. Consequently, every study needs to be carefully designed to efficiently answer the research question, as opposed to data gathering that may or may not prove useful. Understanding and modelling of the nutritional and functional properties of the milk proteins in the context of a balanced, sustainable and healthy diet can greatly benefit from state-of-the-art instrumental techniques as the basis of research. It is therefore the aim of this thesis to demonstrate the application of state-of-the-art high-resolution mass spectrometry techniques to unravel elusive mysteries of bovine milk proteins. The characterisation of milk proteins at the molecular level in their native state in fresh milk, their lactational dynamic, and the further changes that occur to them during processing and storage of dairy products, is pre-requisite to studying and understanding the impact of the proteins in human nutrition and on health. Mass spectrometry proves in this sense to be a very powerful and versatile tool, the continuous developments of which show great potential for furthering milk protein research and dairy science.

References

- Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA. 2007. The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu Rev Immunol*, 25:21-50.
- Atyeo C, Alter G. 2021. The multifaceted roles of breast milk antibodies. *Cell*, 184:1486-1499.
- Bagdonaite I, Malaker SA, Polasky DA, Riley NM, Schjoldager K, Vakhrushev SY, Halim A, Aoki-Kinoshita KF, Nesvizhskii AI, Bertozzi CR. 2022. Glycoproteomics. *Nature Reviews Methods Primers*, 2:48.

Berzelius J. 1814. Über Thierische Chemie. *Schweiggers Journal für Chemie Physik*, 11:261-280.

Biemann K, Gapp G, Seibl J. 1959. Application of mass spectrometry to structure problems. I. Amino acid sequence in peptides. *Journal of the American Chemical Society*, 81:2274-2275.

Bijl E, de Vries R, van Valenberg H, Huppertz T, Van Hooijdonk T. 2014. Factors influencing casein micelle size in milk of individual cows: Genetic variants and glycosylation of κ -casein. *Int Dairy J*, 34:135-141.

Bijl E, Holland JW, Boland M. 2020. Posttranslational modifications of caseins. *Milk proteins*: Elsevier. p. 173-211.

Bondt A, Rombouts Y, Selman MH, Hensbergen PJ, Reiding KR, Hazes JM, Dolhain RJ, Wuhrer M. 2014. Immunoglobulin G (IgG) Fab glycosylation analysis using a new mass spectrometric high-throughput profiling method reveals pregnancy-associated changes. *Mol Cell Proteomics*, 13:3029-3039.

Bonfatti V, de Freitas DR, Lugo A, Vicario D, Carnier P. 2019. Effects of the detailed protein composition of milk on curd yield and composition measured by model micro-cheese curd making of individual milk samples. *J Dairy Sci*, 102:7863-7873.

Braconnot H. 1830. Ueber den Käsestoff und die Milch, und deren neue Nutzenwendungen. *Annalen der Physik*, 95:34-47.

Brignon G, Dumas BR, Mercier J-C, Pelissier J-P, Das B. 1977. Complete amino acid sequence of bovine α S2-casein. *FEBS Lett*, 76:274-279.

Brockhausen I, Wandall HH, Ten Hagen KG, Stanley P. 2022. O-GalNAc Glycans. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, Mohnen D, Kinoshita T, Packer NH, Prestegard JH, et al. editors. *Essentials of Glycobiology*. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press.

Butler JE. 1983. Bovine immunoglobulins: an augmented review. *Veterinary immunology and immunopathology*, 4:43-152.

Cakebread JA, Humphrey R, Hodgkinson AJ. 2015. Immunoglobulin A in bovine milk: a potential functional food? *J Agric Food Chem*, 63:7311-7316.

Cao X, Yang M, Yang N, Liang X, Tao D, Liu B, Wu J, Yue X. 2019. Characterization and comparison of whey N-glycoproteomes from human and bovine colostrum and mature milk. *Food Chem*, 276:266-273.

Catherman AD, Skinner OS, Kelleher NL. 2014. Top down proteomics: facts and perspectives. *Biochem Biophys Res Commun*, 445:683-693.

Čaval T. 2019. Mass Spectrometric Exploration of the GlycoUniverse. University Utrecht.

Čaval T, Heck AJ, Reiding KR. 2021. Meta-heterogeneity: evaluating and describing the diversity in glycosylation between sites on the same glycoprotein. *Mol Cell Proteomics*, 20.

Chi X, Li Y, Qiu X. 2020. V (D) J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology*, 160:233-247.

Cho A, Normile D. 2002. Mastering macromolecules. American Association for the Advancement of Science.

Chucri TM, Monteiro J, Lima A, Salvadori M, Junior JK, Miglino MA. 2010. A review of immune transfer by the placenta. *J Reprod Immunol*, 87:14-20.

de Haan N, Falck D, Wuhrer M. 2020. Monitoring of immunoglobulin N-and O-glycosylation in health and disease. *Glycobiology*, 30:226-240.

Deng Y, Wierenga PA, Schols HA, Sforza S, Gruppen H. 2017. Effect of Maillard induced glycation on protein hydrolysis by lysine/arginine and non-lysine/arginine specific proteases. *Food Hydrocolloids*, 69:210-219.

Dingess KA, Gazi I, van den Toorn HW, Mank M, Stahl B, Reiding KR, Heck AJ. 2021. Monitoring human milk β -casein phosphorylation and O-glycosylation over lactation reveals distinct differences between the proteome and endogenous peptidome. *Int J Mol Sci*, 22:8140.

Dingess KA, Hoek M, van Rijswijk DM, Tamara S, den Boer MA, Veth T, Damen MJ, Barendregt A, Romijn M, Juncker HG. 2022. Identification of common and distinct origins of human serum and breastmilk IgA1 by mass spectrometry-based clonal profiling. *Cell Mol Immunol*:1-12.

Dumas BR, Brignon G, Grosclaude F, Mercier JC. 1972. Structure primaire de la caséine β bovine: séquence complète. *Eur J Biochem*, 25:505-514.

Eichacker LA, Granvogl B, Mirus O, Muller BC, Miess C, Schleiff E. 2004. Hiding behind hydrophobicity: transmembrane segments in mass spectrometry. *J Biol Chem*, 279:50915-50922.

Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, Guigó R, *et al.* 2009. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science*, 324:522-528.

Evershed RP, Payne S, Sherratt AG, Copley MS, Coolidge J, Urem-Kotsu D, Kotsakis K, Özdoğan M, Özdoğan AE, Nieuwenhuysse O. 2008. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature*, 455:528-531.

Fang Z-H, Bovenhuis H, Delacroix-Buchet A, Miranda G, Boichard D, Visker M, Martin P. 2017. Genetic and nongenetic factors contributing to differences in α S-casein phosphorylation isoforms and other major milk proteins. *J Dairy Sci*, 100:5564-5577.

Fang Z-H, Visker M, Miranda G, Delacroix-Buchet A, Bovenhuis H, Martin P. 2016. The relationships among bovine α S-casein phosphorylation isoforms suggest different phosphorylation pathways. *J Dairy Sci*, 99:8168-8177.

Farrell Jr H, Jimenez-Flores R, Bleck G, Brown E, Butler J, Creamer L, Hicks C, Hollar C, Ng-Kwai-Hang K, Swaisgood H. 2004. Nomenclature of the proteins of cows' milk—Sixth revision. *J Dairy Sci*, 87:1641-1674.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246:64-71.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1990. Electrospray ionization—principles and practice. *Mass Spectrom Rev*, 9:37-70.

Fuerer C, Jenni R, Cardinaux L, Andetsion F, Wagnière S, Moulin J, Affolter M. 2020. Protein fingerprinting and quantification of β -casein variants by ultra-performance liquid chromatography—high-resolution mass spectrometry. *J Dairy Sci*, 103:1193-1207.

Gazi I, Reiding KR, Groeneveld A, Bastiaans J, Huppertz T, Heck AJ. 2023. Key changes in bovine milk immunoglobulin G during lactation: NeuAc sialylation is a hallmark of colostrum immunoglobulin G N-glycosylation. *Glycobiology*, 33:115-125.

Godden SM, Lombard JE, Woolums AR. 2019. Colostrum management for dairy calves. *Vet Clin North Am Food Anim Pract*, 35:535-556.

Goonatilleke E, Huang J, Xu G, Wu L, Smilowitz JT, German JB, Lebrilla CB. 2019. Human milk proteins and their glycosylation exhibit quantitative dynamic variations during lactation. *J Nutr*, 149:1317-1325.

Goulding D, Fox P, O'Mahony J. 2020. Milk proteins: An overview. *Milk proteins*:21-98.

Grosclaude F, Mahé MF, Ribadeau-Dumas B. 1973. Structure primaire de la caseine α 1 et de la caseine β bovines: Correctif. *Eur J Biochem*, 40:323-324.

Grosclaude F, Marie-Françoise M, Mercier J-C, Ribadeau-Dumas B. 1972. LOCALISATION DES SUBSTITUTIONS D'ACIDES AMINÉS DIFFÉRENCIANT LES VARIANTES A ET B DE LA CASÉINE x BOVINE. *Ann Genet Sel Anim: EDP Sciences*. p. 515-521.

Guan B, Chai Y, Amantai X, Liu X, Chen X, Cao X, Yue X, Liu B. 2023. Glycoproteomics analysis reveals differential site-specific N-glycosylation of donkey milk fat globule membrane protein during lactation. *Food Chem*, 402:134266.

Holland JW, Deeth HC, Alewood PF. 2004. Proteomic analysis of κ -casein micro-heterogeneity. *Proteomics*, 4:743-752.

Holland JW, Deeth HC, Alewood PF. 2005. Analysis of O-glycosylation site occupancy in bovine κ -casein glycoforms separated by two-dimensional gel electrophoresis. *Proteomics*, 5:990-1002.

Holland JW, Deeth HC, Alewood PF. 2006. Resolution and characterisation of multiple isoforms of bovine κ -casein by 2-DE following a reversible cysteine-tagging enrichment strategy. *Proteomics*, 6:3087-3095.

Horstman AM, Huppertz T. 2022. Milk proteins: Processing, gastric coagulation, amino acid availability and muscle protein synthesis. *Crit Rev Food Sci Nutr*:1-16.

Huppertz T. 2013. Chemistry of the caseins. *Advanced dairy chemistry*: Springer. p. 135-160.

Huppertz T, Heck J, Bijl E, Poulsen NA, Larsen LB. 2021. Variation in casein distribution and mineralisation in the milk from Holstein-Friesian cows. *Int Dairy J*, 119:105064.

Hurley M, Larsen L, Kelly A, McSweeney P. 2000. The milk acid proteinase cathepsin D: a review. *Int Dairy J*, 10:673-681.

Issaq HJ, Conrads TP, Janini GM, Veenstra TD. 2002. Methods for fractionation, separation and profiling of proteins and peptides. *Electrophoresis*, 23:3048-3061.

Jansen BC, Bondt A, Reiding KR, Lonardi E, De Jong CJ, Falck D, Kammeijer GS, Dolhain RJ, Rombouts Y, Wuhrer M. 2016. Pregnancy-associated serum N-glycome changes studied by high-throughput MALDI-TOF-MS. *Sci Rep*, 6:1-10.

Jennewein MF, Alter G. 2017. The Immunoregulatory Roles of Antibody Glycosylation. *Trends Immunol*, 38:358-372.

Kumar BG, Lijina P, Akshata S. 2022. N-glycoprofiling of lactoferrin with site-specificity from buffalo colostrum. *Int Dairy J*, 127:105215.

Kumar N, Arthur CP, Ciferri C, Matsumoto ML. 2020. Structure of the secretory immunoglobulin A core. *Science*, 367:1008-1014.

Kumar N, Arthur CP, Ciferri C, Matsumoto ML. 2021. Structure of the human secretory immunoglobulin M core. *Structure*, 29:564-571. e563.

Le TT, Deeth HC, Larsen LB. 2017. Proteomics of major bovine milk proteins: Novel insights. *Int Dairy J*, 67:2-15.

Leonil J, Mollé D, Gaucheron F, Arpino P, Guénot P, Maubois J. 1995. Analysis of major bovine milk proteins by on-line high-performance liquid chromatography and electrospray ionization-mass spectrometry. *Le Lait*, 75:193-210.

Lucey JA, Horne DS. 2018. Perspectives on casein interactions. *Int Dairy J*, 85:56-65.

Madsen BD, Rasmussen MD, Nielsen MO, Wiking L, Larsen LB. 2004. Physical properties of mammary secretions in relation to chemical changes during transition from colostrum to milk. *J Dairy Res*, 71:263-272.

Marnila P, Korhonen H. 2011. Milk Proteins | Immunoglobulins. In: Fuquay JW editor. *Encyclopedia of Dairy Sciences*: Academic Press. p. 807-815.

Marx H, Lemeer S, Schliep JE, Matheron L, Mohammed S, Cox J, Mann M, Heck AJ, Kuster B. 2013. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat Biotechnol*, 31:557-564.

McGrath BA, Fox PF, McSweeney PL, Kelly AL. 2016. Composition and properties of bovine colostrum: a review. *Dairy Sci Technol*, 96:133-158.

Mercier JC, Brignon G, Ribadeau-dumas B. 1973. Structure primaire de la caséine κ B bovine: Séquence complète. *Eur J Biochem*, 35:222-235.

Mercier JC, Grosclaude F, Ribadeau-Dumas B. 1971. Structure primaire de la caséine α sl-bovine: Séquence complète. *Eur J Biochem*, 23:41-51.

Miranda G, Bianchi L, Krupova Z, Trossat P, Martin P. 2020. An improved LC–MS method to profile molecular diversity and quantify the six main bovine milk proteins, including genetic and splicing variants as well as post-translationally modified isoforms. *Food chemistry: X*, 5:100080.

Muehlhoff E, Bennett A, McMahon D. 2013. *Milk and dairy products in human nutrition* Food and Agriculture Organization of the United Nations (FAO).

Nagumo K, Tanaka M, Chuang VTG, Setoyama H, Watanabe H, Yamada N, Kubota K, Tanaka M, Matsushita K, Yoshida A, *et al.* 2014. Cys34-Cysteinylated Human Serum Albumin Is a Sensitive Plasma Marker in Oxidative Stress-Related Chronic Diseases. *PLoS One*, 9:e85216.

Nguyen DD, Solah VA, Busetti F, Smolenski G, Cooney T. 2020. Application of ultra-high performance liquid chromatography coupled to high-resolution mass spectrometry (Orbitrap™) for the determination of beta-casein phenotypes in cow milk. *Food Chem*, 307:125532.

Nilsson K, Johansen LB, de Koning D, Duchemin S, Hansen MS, Ståhlhammar H, Lindmark-Månsson H, Paulsson M, Fikse W, Glantz M. 2020. Effects of milk proteins and posttranslational modifications on noncoagulating milk from Swedish Red dairy cattle. *J Dairy Sci*, 103:6858–6868.

Nwosu CC, Strum JS, An HJ, Lebrilla CB. 2010. Enhanced detection and identification of glycopeptides in negative ion mode mass spectrometry. *Anal Chem*, 82:9654-9662.

O'Riordan N, Kane M, Joshi L, Hickey RM. 2014. Structural and functional characteristics of bovine milk protein glycosylation. *Glycobiology*, 24:220–236.

O'Donnell R, Holland J, Deeth H, Alewood P. 2004. Milk proteomics. *Int Dairy J*, 14:1013-1023.

Ohtsubo K, Marth JD. 2006. Glycosylation in cellular mechanisms of health and disease. *Cell*, 126:855-867.

Oliveira T, Thaysen-Andersen M, Packer NH, Kolarich D. 2021. The Hitchhiker's guide to glycoproteomics. *Biochem Soc Trans*, 49:1643-1662.

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

Otvos Jr L, Urge L, Thurin J. 1992. Influence of different N-and O-linked carbohydrates on the retention times of synthetic peptides in reversed-phase high-performance liquid chromatography. *Journal of Chromatography A*, 599:43-49.

Plomp R, Dekkers G, Rombouts Y, Visser R, Koeleman CA, Kammeijer GS, Jansen BC, Rispens T, Hensbergen PJ, Vidarsson G, *et al.* 2015. Hinge-Region O-Glycosylation of Human Immunoglobulin G3 (IgG3)*[S]. *Mol Cell Proteomics*, 14:1373-1384.

Qu Y, Kim B-J, Koh J, Dallas DC. 2021. Analysis of bovine kappa-casein glycomacropeptide by liquid chromatography–tandem mass spectrometry. *Foods*, 10:2028.

Rauh VM, Johansen LB, Ipsen R, Paulsson M, Larsen LB, Hammershøj M. 2014. Plasmin activity in UHT milk: Relationship between proteolysis, age gelation, and bitterness. *J Agric Food Chem*, 62:6852-6860.

Reiding KR, Bondt A, Franc V, Heck AJ. 2018. The benefits of hybrid fragmentation methods for glycoproteomics. *TrAC, Trends Anal Chem*, 108:260-268.

Reily C, Stewart TJ, Renfrow MB, Novak J. 2019. Glycosylation in health and disease. *Nat Rev Nephrol*, 15:346-366.

Roffet-Salque M, Gillis RE, Evershed RP, Vigne J-D. 2018. Milk as a pivotal medium in the domestication of cattle, sheep and goats. *Hybrid Communities*: Routledge. p. 127-143.

Ruhaak LR, Uh H-W, Deelder AM, Dolhain RE, Wuhrer M. 2014. Total plasma N-glycome changes during pregnancy. *J Proteome Res*, 13:1657-1668.

Saito T, Itoh T. 1992. Variations and distributions of O-glycosidically linked sugar chains in bovine κ -casein. *J Dairy Sci*, 75:1768-1774.

Stanfield RL, Haakenson J, Deiss TC, Criscitiello MF, Wilson IA, Smider VV. 2018. The unusual genetics and biochemistry of bovine immunoglobulins. *Adv Immunol*, 137:135-164.

Stanley P, Moremen KW, Lewis NE, Taniguchi N, Aebi M. 2022. N-Glycans. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, Mohnen D, Kinoshita T, Packer NH, Prestegard JH, et al. editors. *Essentials of Glycobiology*. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press.

Steen H, Jebanathirajah JA, Rush J, Morrice N, Kirschner MW. 2006. Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. *Mol Cell Proteomics*, 5:172-181.

Stelwagen K, Carpenter E, Haigh B, Hodgkinson A, Wheeler T. 2009. Immune components of bovine colostrum and milk. *J Anim Sci*, 87:3-9.

Stowell SR, Ju T, Cummings RD. 2015. Protein glycosylation in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 10:473-510.

Strugnell RA, Wijburg OL. 2010. The role of secretory antibodies in infection immunity. *Nature Reviews Microbiology*, 8:656-667.

Sunds AV, Poulsen NA, Larsen LB. 2019. Application of proteomics for characterization of caseinomacropeptide isoforms before and after desialidation. *J Dairy Sci*, 102:8696-8703.

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.

Takimori S, Shimaoka H, Furukawa JI, Yamashita T, Amano M, Fujitani N, Takegawa Y, Hammarström L, Kacsokovics I, Shinohara Y. 2011. Alteration of the N-glycome of bovine milk glycoproteins during early lactation. *The FEBS journal*, 278:3769-3781.

Thesbjerg MN, Johansen M, Larsen LB, Poulsen NA. 2022. Differences in post-translational modifications of proteins in milk from early and mid-lactation dairy cows as studied using total ion chromatograms from LC-ESI/MS. *Int Dairy J*, 130:105262.

Thomson JJ. 1897. XL. Cathode rays. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 44:293-316.

Tsiatsiani L, Heck AJR. 2015. Proteomics beyond trypsin. *The Febs Journal*, 282:2612-2626.

Tyers M, Mann M. 2003. From genomics to proteomics. *Nature*, 422:193-197.

Valk-Weeber RL, Deelman-Driessen C, Dijkhuizen L, Eshuis-de Ruiter T, van Leeuwen SS. 2020. In depth analysis of the contribution of specific glycoproteins to the overall bovine whey N-linked glycoprofile. *J Agric Food Chem*, 68:6544-6553.

van Boekel M. 1998. Effect of heating on Maillard reactions in milk. *Food Chem*, 62:403-414.

van Lieshout GAA, Lambers TT, Bragt MCE, Hettinga KA. 2020. How processing may affect milk protein digestion and overall physiological outcomes: A systematic review. *Crit Rev Food Sci Nutr*, 60:2422-2445.

Van Rossum C, Buurma-Rethans E, Dinnissen C, Beukers M, Brants H, Ocké M. 2020. The diet of the Dutch: Results of the Dutch National Food Consumption Survey 2012-2016.

Vandermarliere E, Mueller M, Martens L. 2013. Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom Rev*, 32:453-465.

Varki A. 2017. Biological roles of glycans. *Glycobiology*, 27:3-49.

Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Marth JD, Bertozzi CR, Hart GW, Etzler ME. 2009. Symbol nomenclature for glycan representation. *Proteomics*, 9:5398-5399.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.* 2001. The Sequence of the Human Genome. *Science*, 291:1304-1351.

Visser S, Slangen CJ, Rollema HS. 1991. Phenotyping of bovine milk proteins by reversed-phase high-performance liquid chromatography. *Journal of Chromatography A*, 548:361-370.

Vreeman H, Visser S, Slangen CJ, Van Riel J. 1986. Characterization of bovine κ -casein fractions and the kinetics of chymosin-induced macropeptide release from carbohydrate-free and carbohydrate-containing fractions determined by high-performance gel-permeation chromatography. *Biochemical Journal*, 240:87-97.

Wandall HH, Nielsen MA, King-Smith S, de Haan N, Bagdonaite I. 2021. Global functions of O-glycosylation: promises and challenges in O-glycobiology. *The FEBS Journal*, 288:7183-7212.

Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL, Humphery-Smith I. 1995. Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis*, 16:1090-1094.

Weström B, Arévalo Sureda E, Pierzynowska K, Pierzynowski SG, Pérez-Cano F-J. 2020. The immature gut barrier and its importance in establishing immunity in newborn mammals. *Front Immunol*, 11:1153.

Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez J-C, Yan JX, Gooley AA, Hughes G, Humphery-Smith I. 1996. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)*, 14:61-65.

Wong S, Meng C, Fenn J. 1988. Multiple charging in electrospray ionization of poly (ethylene glycols). *The Journal of Physical Chemistry*, 92:546-550.

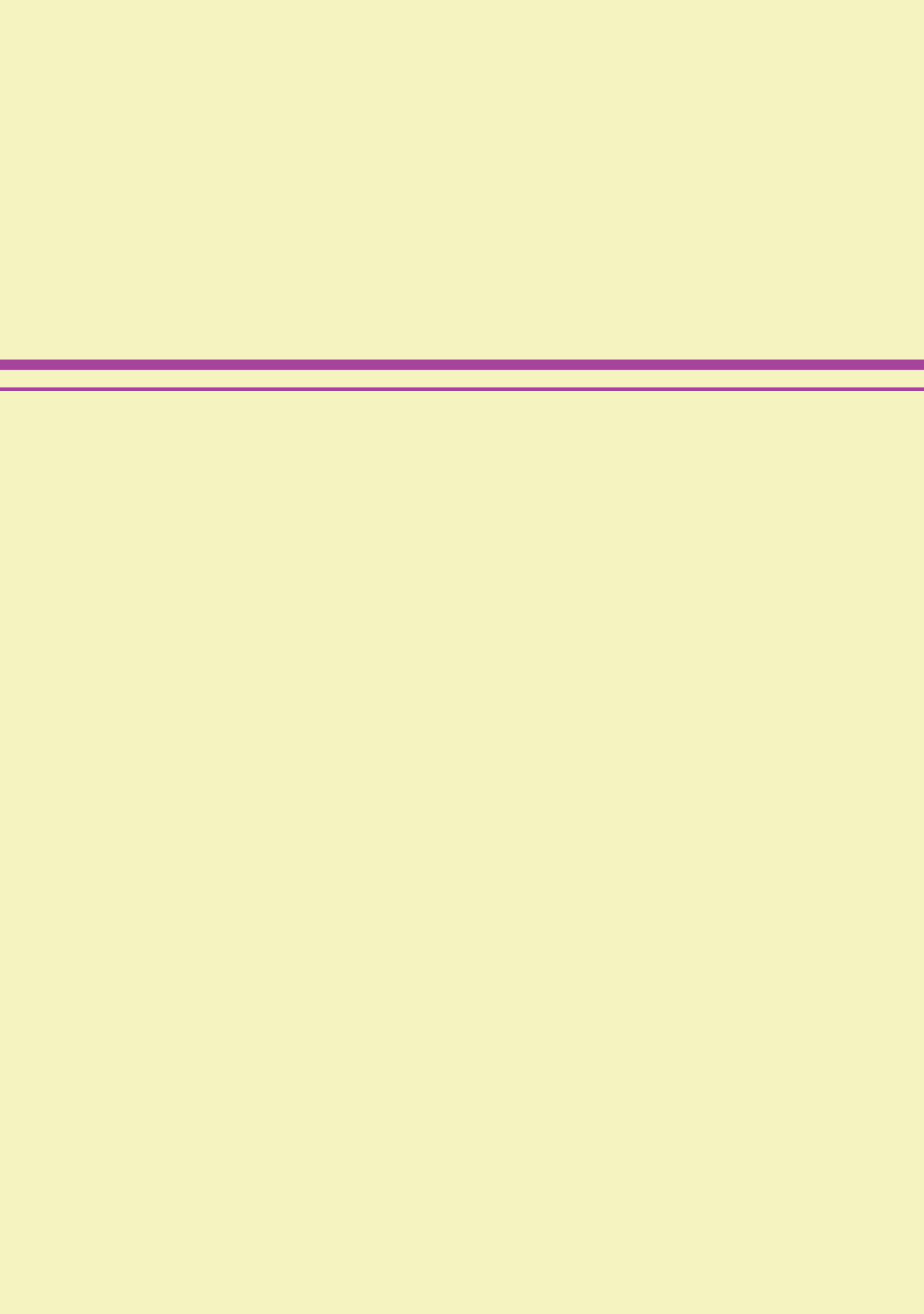
Xiao J, Wang J, Gan R, Wu D, Xu Y, Peng L, Geng F. 2022. Quantitative N-glycoproteome analysis of bovine milk and yogurt. *Current Research in Food Science*, 5:182-190.

Yang Y, Franc V, Heck AJ. 2017. Glycoproteomics: a balance between high-throughput and in-depth analysis. *Trends Biotechnol*, 35:598-609.

Yates III JR. 2011. A century of mass spectrometry: from atoms to proteomes. *Nat Methods*, 8:633-637.

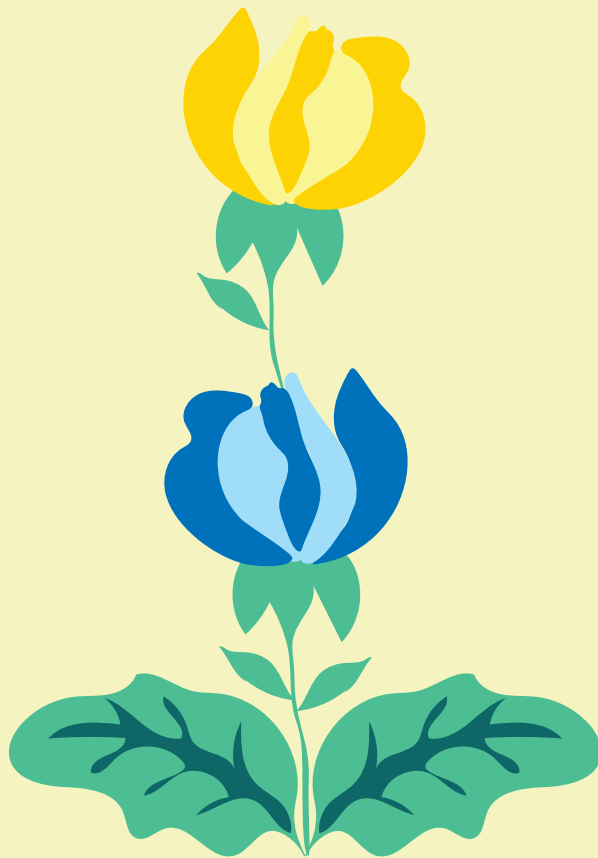
Zhang J, Chen Q, Zhou Y, Zhang Y, Ren Y, Liu L. 2022. Characterization and determination of bovine immunoglobulin G subtypes in milk and dairy products by UPLC-MS. *Food Chem*, 390:133170.

Zhu J, Lin Y-H, Dingess KA, Mank M, Stahl B, Heck AJ. 2020. Quantitative longitudinal inventory of the N-glycoproteome of human milk from a single donor reveals the highly variable repertoire and dynamic site-specific changes. *J Proteome Res*, 19:1941-1952.



CHAPTER 2

Identifying glycation hot-spots in bovine milk proteins
during production and storage of skim milk powder



Chapter 2. Identifying glycation hot-spots in bovine milk proteins during production and storage of skim milk powder

Published as:

Gazi, I. *et al.* 2022. *International Dairy Journal*, 129:105340

<https://doi.org/10.1016/j.idairyj.2022.105340>

Abstract

We investigated protein glycation in a complex milk system under controlled conditions representative of real-life consumer products, analysing intermediate and final products from skim milk powder production, and aged powder samples. We combined protein-centric LC-MS(/MS) with peptide-centric multi-protease LC-MS/MS focusing on the six most abundant bovine milk proteins. This strategy resulted in the identification of glycated proteoforms and of the extent of glycation *per* protein, high protein sequence coverage, and identification and relative occupancy of the glycation sites. We identified new glycation hot-spots additionally to the ones already described in literature. Primary sequence motif analysis revealed that glycation hot-spots were preceded N-terminally by a stretch rich in basic amino acids, and followed C-terminally by a stretch enriched in aliphatic and hydrophobic amino acids. Our study considerably extends the current understanding of milk protein glycation, discussing glycation hot-spots and their localization in relation to the primary sequences and higher-order protein structures.

Introduction

Heat treatment of milk is imperative for rendering it safe for consumption, specifically from the perspective of microbiological safety (Boor *et al.*, 2017). However, an unintended consequence of heat treatment is the series of chemical reactions, collectively known as the Maillard reaction (MR). This process starts with protein glycation and eventually leads to advanced glycation end products (van Boekel, 1998). Glycated proteins exhibit altered digestibility and reduced bioavailability of the essential amino acid lysine (van Lieshout *et al.*, 2020), whereas advanced MR products have been linked with increased immunogenicity, allergenicity, inflammation, and oxidative stress (Baye *et al.*, 2017, Sharma and Barone, 2019, Teodorowicz *et al.*, 2017, Toda *et al.*, 2019). Detailed investigation of glycation in milk systems is thus important to understand the reaction mechanism and to prevent or mitigate it.

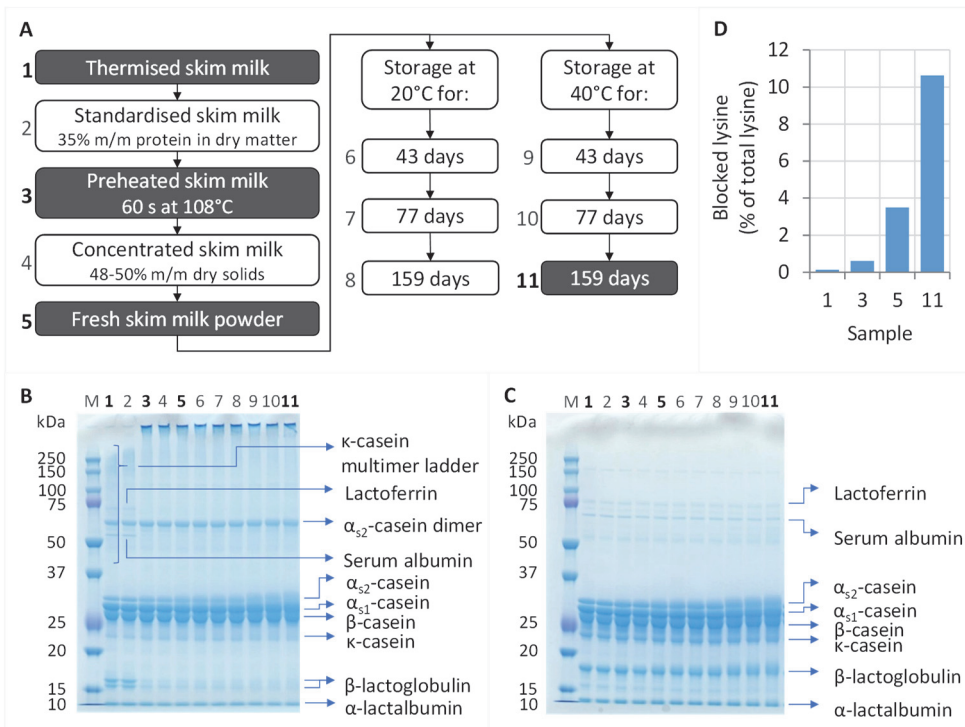
Glycation of milk proteins and the advancement of the MR is most frequently analysed by using MR indicators such as available lysine or furosine in the early stages of the reaction, 5-hydroxymethyl-2-furfuraldehyde (HMF) or 2-furfuraldehyde (furfural) in the intermediate stages, and browning index in the advanced stages of the reaction (Aalaei *et al.*, 2019, Gómez-Narváez *et al.*, 2019). Alternatively, mass spectrometry (MS) of the six most abundant intact milk proteins, *i.e.*, α_{s1} -, α_{s2} -, β - and κ -casein, α -lactalbumin and β -lactoglobulin (Arena *et al.*, 2017, Oliver, 2011, Siciliano *et al.*, 2013), can be employed to analyse glycated

proteins. This latter approach monitors mass shifts induced by adding glycans on proteins, revealing the overall extent of glycation, without identifying glycation sites and their occupancy. Glycation site localization in the protein structure is important for identifying newly-formed immunogenic and allergenic epitopes. A peptide-centric mass spectrometry approach allows site-specific characterization and complements MS on intact proteins (Franc *et al.*, 2018, Tamara *et al.*, 2020, van de Waterbeemd *et al.*, 2018, Yang *et al.*, 2016). Trypsin is the gold standard protease in peptide-centric MS (Tsiatsiani and Heck, 2015), cleaving exclusively C-terminal of arginine and lysine residues (Olsen *et al.*, 2004). Glycation of these two residues hampers cleavage by trypsin, resulting in suboptimal digestion and hindering downstream LC-MS/MS analysis (Deng *et al.*, 2017). Nevertheless, the greater part of published peptide-centric MS research on milk protein glycation still relies on tryptic digestion, despite this decreased performance of the enzyme towards glycated proteins. To gain a holistic picture of the extent of glycation, localization and site occupancy during the processing and storage of skim milk powder, we combined here protein- and multi-protease peptide-centric MS approaches. This strategy resulted in high protein sequence coverage irrespective of the extent of glycation, overcoming the limitations of using solely trypsin.

Apart from the choice of protease in the peptide-centric approach, the selected dairy system to investigate and the experimental design may also lead to biased results. Studies focusing on heat- and storage-induced changes occurring in model systems of isolated milk proteins in combination with reducing sugars (Cardoso *et al.*, 2018, Carulli *et al.*, 2011, Cattaneo *et al.*, 2017, Chevalier *et al.*, 2002, Gómez-Narváez *et al.*, 2019, Leiva *et al.*, 2017) only consider a few isolated factors, *e.g.*, protein and carbohydrate concentration, and type of carbohydrate. Due to differences in sample complexity and processing conditions, findings from model systems do not necessarily translate well to consumer dairy systems.

Glycation studies conducted on off-the-shelf commercial products (Birlouez-Aragon *et al.*, 2004, Delatour *et al.*, 2009, Hegele *et al.*, 2008, Milkovska-Stamenova and Hoffmann, 2016a, Milkovska-Stamenova and Hoffmann, 2016b, Renzone *et al.*, 2015) generate results directly representative of the consumer dairy products. However, the difficulty in interpreting results from these studies comes from the fact that the starting ingredients, and processing and storage conditions are all unknown, and may differ between the same types of products from different brands.

Our study was designed for the in-depth investigation of protein glycation during the production and storage of skim milk powder. The production process was



2

Figure 2.1 – Physical-chemical changes in the major milk proteins during production and storage of skim milk powder analysed by SDS-PAGE. (A) Flowchart depicting the eleven samples analysed, namely the intermediate and final products of the skim milk powder production process, and the aged powders obtained by storage at 20 and 40 °C, respectively. (B) and (C) SDS-polyacrylamide gel electrophoretograms of all samples shown in the flowchart (A) analysed under non-reducing (B) and under reducing (C) conditions; M = molecular weight marker. (D) Total percentage of blocked lysine residues determined as lactosylated lysine expressed as a percentage of total lysine in samples 1, 3, 5 and 11 as annotated in flowchart (A).

carried out on pilot scale representative of industrial processing to overcome the previously described limitations. All processing and storage steps were performed under controlled conditions. All intermediate, final and aged samples belonged to the same batch. The extent of glycation was monitored in a site-specific manner for the six most abundant milk proteins. We observed the gradual increase of glycation and substantial residue-specific differences, and identified glycation hot-spots in these six proteins. We revealed a generic preferential sequence motif of glycated lysine residues, thus expanding the current understanding of milk protein glycation with new insights.

Results

Our study aimed to investigate protein glycation in consumer products, analysing samples taken during the processing of skim milk into skim milk powder on pilot

scale, and during subsequent aging of the powder. To study protein glycation under these conditions, we analysed a batch of 11 different samples in parallel, as schematically summarized in Figure 2.1A.

Gel electrophoresis and determination of the percentage of blocked lysine residues

Samples were analysed by SDS-PAGE to obtain a comprehensive overview of the steps where the proteins undergo changes, and of the types of changes that occur. Images of the SDS-polyacrylamide gels of the samples summarized in Figure 2.1A are shown in Figure 2.1B and C. High molecular weight disulphide-linked protein aggregates were formed due to preheating the milk at 105 °C for 60 s; these aggregates were too large to enter the gel and therefore were visible as intense bands at the top of the non-reducing gel (Figure 2.1B). The aggregates were not observed in the reducing gel (Figure 2.1C) where the individual protein bands became visible, further confirming that the nature of the aggregates was indeed disulphide-linked and not that of other covalent linkages.

Another noticeable change observed in the gels is the broadening of the protein bands (Figure 2.1B and C). This started to occur in the preheated milk and increased with subsequent processing, and particularly during storage of the powder. The broadening of the protein bands likely indicates that the original proteins have differentiated into a series of more diverse proteoforms with altered migration times due to process- and storage-induced chemical modifications.

Based on the SDS-PAGE results (Figure 2.1B and C), four key samples were chosen for further analyses, *i.e.*, the thermised milk (sample 1), the preheated milk (sample 3), the fresh skim milk powder (sample 5), and the aged skim milk powder (sample 11; 159 d at 40 °C).

The broadening of the protein bands on SDS-PAGE was presumed to be primarily caused by glycation of the proteins with lactose, *i.e.*, lactosylation. The four samples specified above were analysed for the percentage of blocked lysine residues to confirm whether protein glycation was indeed responsible for the observed changes. The results from the blocked lysine analyses are shown in Figure 2.1D, confirming that lactosylation occurred during processing and storage. Virtually no blocked lysine was detected in the thermised milk. A low level of blocked lysine (<1% of total lysine) was detected in the preheated milk following heat treatment at 108 °C for 60 s. Further heat load during concentration of the milk and spray-drying led to the increase of blocked lysine to 3.5% in the fresh skim milk powder. Our experimental storage conditions (water activity <0.2 and temperature of 20° or

40 °C) ensured that the lactose in the milk powders remained below its glass transition temperature (T_g ; Schuck et al., 2005). Despite the limited molecular mobility below T_g and thus slower glycation reaction kinetics (Gonzales *et al.*, 2010), the level of blocked lysine further increased to 10.6% of total lysine during storage of the powder at 40 °C for 159 d.

Protein-centric mass spectrometry

We further analysed the thermised and preheated milk, and the fresh and aged (159 d at 40 °C) skim milk powders by protein-centric mass spectrometry. Each sample was analysed by a triplet of one LC-MS and two LC-MS/MS methods, amounting to a total of 12 runs (see Materials and Methods and Supplementary information for further details). The combined results of this approach ensured identification and relative quantitation of the different proteoforms of the six most abundant milk proteins in the samples, *i.e.*, α_{s1} -, α_{s2} -, β - and κ -casein, α -lactalbumin and β -lactoglobulin, as presented in Figure 2.2. Cumulatively, these proteins make up >90% of total protein in bovine milk (Brown *et al.*, 2020, Dupont *et al.*, 2013), and thus also constitute the major source of protein glycation in bovine milk. Furthermore, within the current experimental setup, glycation was found to be the most abundant modification occurring during processing and storage of skim milk powder, with native and glycated proteoforms accounting for the greater part of the intact mass spectra (supplementary Figure S2.1).

The thermised milk was the sample with the lowest intensity of processing in the current study. As expected, and in line with results shown in in Figure 2.1D, little to no glycation could be detected in this sample when analysed with the protein-centric approach (Figure 2.2). Consequently, the thermised milk is representative of the endogenous repertoire of proteoforms for each of the six most abundant bovine milk proteins, including heterogeneity resulting from genetic polymorphism and varying levels of phosphorylation (Figure 2.2). For simplicity, the *O*-glycosylated forms of κ -casein were not included in this analysis and the focus was kept on the more abundant non-glycosylated κ -casein proteoforms.

Preheating milk led to an increase in protein heterogeneity primarily due to lactosylation, detected as dihexose residue mass increments (+324 Da) at the intact protein level. Further increases in protein heterogeneity occurred particularly during skim milk powder storage, with increasing abundances of glycated proteoforms. Next to lactosylation, lower levels of glycation with monohexoses (+162 Da) were also observed on intact proteins, particularly in the aged powder.

The two proteins most susceptible to glycation were found to be α_{s2} -casein and β -lactoglobulin. The entire proteoform profile of the α_{s2} -casein shifted and diversified during production due to up to 2 glycation sites *per* protein molecule being occupied in the fresh powder (Figure 2.2D). The non-glycated proteoforms still dominated the α_{s2} -casein proteoform profile of the fresh skim milk powder. This changed in the case of the aged powder, where up to 5 lactose residues *per* α_{s2} -casein molecule were detected. The α_{s2} -casein fraction in this sample became dominated by the proteoforms with 2 and 3 lactose residues *per* molecule.

In the case of β -lactoglobulin glycation could already be detected in the thermised milk, indicating that this protein is particularly susceptible to glycation even under very mild processing conditions (Figure 2.2B). The degree of β -lactoglobulin glycation further increased during processing resulting in up to 3 lactose residues *per* molecule detected in the fresh powder. The monolactosylated forms became dominant in the proteoform profiles along with the non-glycated forms. Considerable glycation occurred further during storage leading to up to 5 lactose residues *per* β -lactoglobulin molecule in the aged powder, with the proteoform profile shifting in favour of proteoforms carrying 0-2 lactose residues *per* molecule (Figure 2.2B).

In the case of α_{s1} -casein, no glycation was detected in the thermised milk and only limited glycation could be detected during further processing with up to 1 lactose residue *per* molecule in both the preheated milk and in the fresh powder (Figure 2.2C). In contrast, considerable glycation of the α_{s1} -casein occurred during storage of the powder, with up to 5 lactose residues *per* molecule being detected, whereby mono- and dilactosylated forms dominated the proteoform profile. The glycation of β -casein occurred gradually during processing, with monolactosylation of the proteins during the preheating of milk and dilactosylation occurring in the fresh powder (Figure 2.2E). At the same time, the samples remained dominated by the non-glycated proteoforms. As also seen in the other proteins, considerable glycation of β -casein occurred during storage, yielding a proteoform profile dominated by the forms with 0-2 lactose residues, and a total of up to 4 lactose residues detected *per* molecule.

Unlike β -lactoglobulin, the other major whey protein, α -lactalbumin, proved to be far less susceptible to glycation (Figure 2.2A). Limited glycation occurred during production with only monolactosylation detectable in the preheated milk and dilactosylation detectable in the fresh powder, but the samples remained dominated by the non-glycated proteoform of α -lactalbumin. The glycation

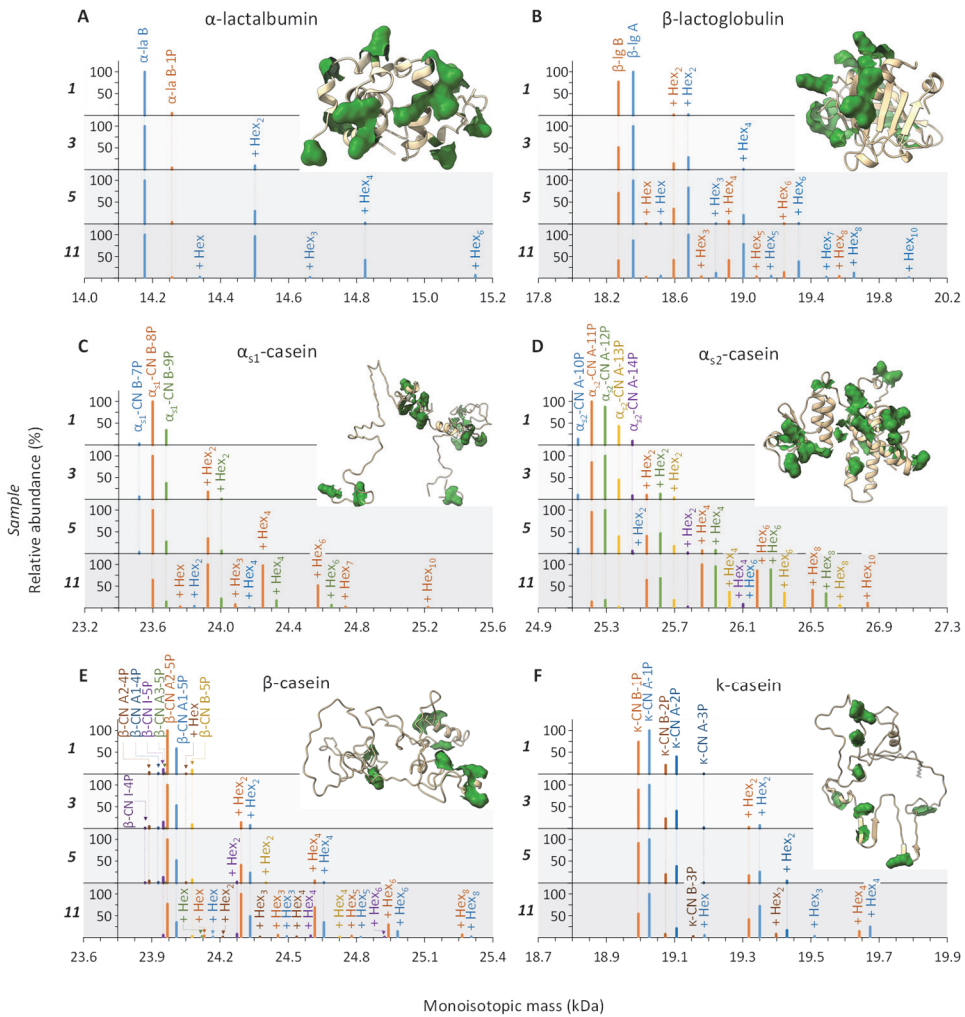


Figure 2.2 – Changes in the proteoform profiles of the six most abundant bovine milk proteins during production and storage of skim milk powder: α -lactalbumin (A), β -lactoglobulin (B), α_{s1} -casein (C), α_{s2} -casein (D), β -casein (E) and κ -casein (F). Shown are deconvoluted mass spectra depicting all proteoforms that were identified in thermised milk (1), preheated milk (3), fresh skim milk powder (5) and aged (159 d at 40 °C) skim milk powder (11). The nomenclature used for the different proteoforms is as proposed by the American Dairy Science Association Committee (Farrell Jr et al., 2004). The most intense proteoform in each spectrum was set to 100% abundance, relative to which the intensities of the other proteoforms were also expressed. The monomeric protein structures depicted in each panel were generated with ChimeraX 1.1; the lysine residues are highlighted in green. The whey protein structures are α -lactalbumin PDB 1F6S, chain B, and β -lactoglobulin PDB 1BEB, chain B. The casein structures are the energy-minimized monomers that were modelled into the submicellar casein particles as described by Farrell et al. (2013).

advanced during storage with up to 3 lactose residues detectable *per* molecule and with the proteoform profile becoming dominated by the non-glycated and monolactosylated forms of α -lactalbumin (Figure 2.2A). κ -Casein showed even lower susceptibility to glycation than α -lactalbumin (Figure 2.2F). Monolactosylation could be detected in the preheated milk and fresh powder, and dilactosylation was detected in the aged powder. Irrespective of the sample, the proteoform profile of κ -casein remained dominated by the non-glycated proteoforms.

Irrespective of protein, no preference of glycation was observed towards specific genetic variants or phosphoproteoforms; instead, all forms of the same protein appeared to become glycated in a similar way (supplementary Table S2.2 and Figure S2.2).

The percentage of glycated lysine in the sample calculated from the protein-centric mass spectrometry results was found to be in good accordance with the results of the blocked lysine analysis (supplementary Figure S2.3). The percentage of glycated lysine estimated by mass spectrometry was 0.9 times that determined from the blocked lysine analysis. This difference could be caused by the assumptions made for calculation (*i.e.*, glycation occurred exclusively on the six most abundant bovine milk proteins in the form of lysine lactosylation, excluding all other proteins and peptides, and other forms of glycation) and/or by the intensity suppression of glycation in the mass spectrometer due to reduction of the positive charges on the glycated residues (Fuerer *et al.*, 2020).

Peptide-centric mass spectrometry

We complemented the data obtained by protein-centric mass spectrometry with peptide-centric mass spectrometry analyses to obtain a detailed site-specific characterization of protein glycation. The four samples were digested with a series of proteases of different specificities (see Table 2.1), amounting to 28 LC-MS/MS runs, the cumulative results of which ensured high sequence coverage irrespective of glycation status (supplementary Figure S2.4). Glycation was searched in the form of dihexose (*i.e.*, lactosylation) and monohexose additions on lysine or arginine residues, or protein N-termini of the six most abundant milk proteins. Coverage of all putative glycation sites was achieved, except for the protein N-terminus of α_{s2} -casein. From this data it was possible to extract the major glycation sites and to quantify glycation site occupancy (see Materials and methods and Supplementary information for further details).

Table 2.1 – Protease combinations and digestion conditions used for hydrolysing the milk proteins

Digestion	Step 1		Step 2	
	Enzyme ¹	Incubation	Enzyme	Incubation
1	Trypsin	37 °C, overnight	-	-
2	Chymotrypsin	RT ² , overnight	-	-
3	GluC	RT, overnight	-	-
4	AspN	37 °C, overnight	-	-
5	Chymotrypsin	RT, 4 h	Trypsin	37 °C, overnight
6	GluC	RT, 4 h	Trypsin	37 °C, overnight
7	AspN	37 °C, 4 h	Trypsin	37 °C, overnight

Figure 2.3 provides site-specific information on glycation in the six most abundant milk proteins during production and storage of the skim milk powder. Data is shown only for the sites where glycation was detected at $\geq 1\%$ of site occupancy. The vast majority of the detected glycation sites are lysine residues. To a lower extent, glycation was also detected at the protein N-terminus of α -lactalbumin (Figure 2.3A). As also observed from the analysis at the protein level (Figure 2.2), the peptide-centric approach revealed that glycation predominantly occurred in the form of lactosylation and, to a much lower extent, also in the form of glycation with monohexoses. Glycation with monohexoses was primarily detected in the aged powder. The results of the peptide-centric approach, and also in line with the results at the protein level (Figure 2.2), show that β -lactoglobulin and α_{s2} -casein exhibited the highest susceptibility to glycation (Figure 2.3B and D, respectively). For each of these proteins, up to 9 glycation sites were identified in the analysed samples, with one of them showing nearly 100% occupancy in the aged powder. Particularly residues Lys₁₄₁ in β -lactoglobulin and Lys₁₇₃ in α_{s2} -casein were detected at 37% and 100% occupancy already in the preheated milk. The occupancy of β -lactoglobulin Lys₁₄₁ gradually increased to 95% glycation in the aged powder. As also observed from the protein-centric results, α_{s1} -casein showed lower susceptibility to glycation during milk powder production than α_{s2} -casein and β -lactoglobulin (<5% glycation site occupancy for all sites), but it developed considerable glycation during storage (Figure 2.3C). Particularly α_{s1} -casein residues Lys₇, Lys₄₂, Lys₈₃, Lys₁₀₂ and Lys₁₀₃ had site occupancies in the range of 14-69% glycation in the aged powder. β -Casein showed lower susceptibility to glycation than the α_{s1} -caseins (Figure 2.3E), and the least susceptible proteins were found to be α -lactalbumin (Figure 2.3A) and κ -casein (Figure 2.3F), with only 4 and 3 glycation sites detected *per* protein molecule, respectively. For β -casein, κ -casein

and α -lactalbumin, glycation site occupancies did not exceed 5% during the processing of skim milk into skim milk powder. Considerable increases in glycation site occupancy occurred during storage for β -casein residues Lys₂₉, Lys₃₂, Lys₁₀₅, and Lys₁₆₉, α -lactalbumin residues Glu₁, Lys₆₂ and Lys₉₈, and κ -casein residue Lys₂₄, all increasing to the range of 10-27% occupancy. These findings are also supported by our data obtained by using the protein-centric approach. The overall picture

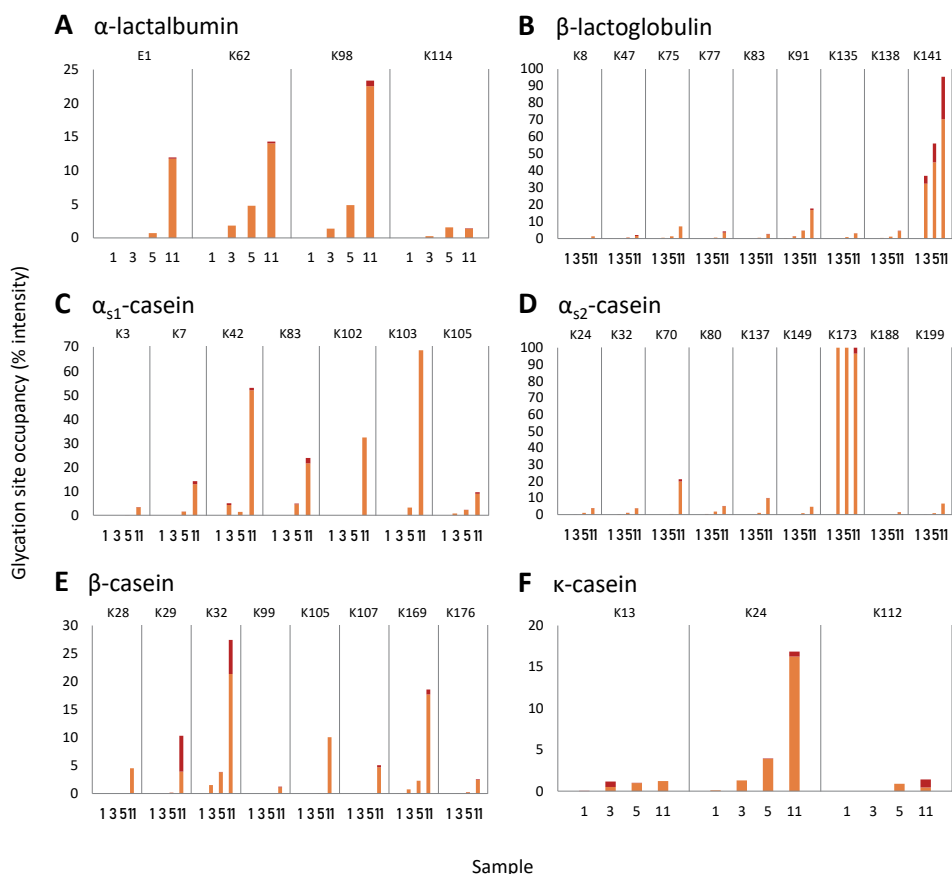


Figure 2.3 – Glycation hot-spots detected in the six most abundant bovine milk proteins. Glycation site occupancy (sum of intensities of peptides where a site has been identified as being glycated expressed as a percentage of total sum of intensities of all peptides containing the respective site) was determined by using a peptide-centric mass spectrometry approach, combining data obtained by using 7 protease combinations and a total of 28 LC-MS/MS runs. The glycation occupancy is shown for each of the four samples; thermised milk (sample 1), preheated milk (sample 3), fresh skim milk powder (sample 5) and aged (159 d at 40 °C) skim milk powder (sample 11). ■ indicates glycation through addition of a dihexose; ■ indicates glycation by a monohexose.

emerging from the data shown in Figures 2.2 and 2.3 is that little to no glycation was detected in the thermised milk. Low levels of glycation were induced by the heat treatment (60 s at 108 °C). Glycation increased in the skim milk powder as a result of the heat load during spray-drying. Considerable increases in glycation at existing and new sites occurred during the accelerated storage (159 d at 40 °C) of the skim milk powder. Glycation primarily occurred with dihexoses, *i.e.*, lactose. Glycation with monohexoses was also detected to a lower extent, particularly in the aged skim milk powder. Moreover, specific lysine residues were substantially more prone to glycation than others, as further detailed below.

Glycated lysine motif analysis

The site-specific information obtained from the peptide-centric mass spectrometry approach (Figure 2.3) indicated that not all putative glycation sites are equally susceptible to modification, with only a few selected sites in each protein representing so-called glycation hot-spots. Literature to date does not thoroughly describe the origin of these differences between the various lysine residues in milk proteins. In trying to address this question, we performed a sequence motif analysis of all the lysine residues in the six most abundant milk proteins to investigate whether the amino acid sequence containing the glycated or non-glycated residues correlates with their susceptibility towards glycation. The motif of the non-glycated lysine residues (63 residues) was subtracted from the motif of the glycated lysine residues with >5% site occupancy (22 residues) from the aged skim milk powder (159 d at 40 °C) to better highlight preferential sequence composition around the observed major glycation sites. This analysis, illustrated in Figure 2.4 reveals for the first time that: 1) majorly glycated lysine residues in the most abundant bovine milk proteins were often preceded by a stretch of positively charged amino acids (blue), *i.e.*, primarily lysine residues and to a smaller extent histidine residues, and 2) at the C-terminal end of the glycated lysine a stretch of mostly aliphatic and hydrophobic amino acids, *i.e.*, valine, leucine and isoleucine, proline and methionine was found to be enriched (Figure 2.4, black).

Discussion

We carefully designed our study to fill knowledge gaps surrounding milk protein glycation in a complex and realistic dairy system that arise from sample selection, protease selection for peptide-centric mass spectrometry, general experimental design, and data analysis and interpretation. With this aim, we analysed a series of samples belonging to a single batch of bulk milk processed into skim milk powder and subsequently aged under controlled conditions. Assessment based on SDS-PAGE and the percentage of blocked lysine residues revealed glycation occurring

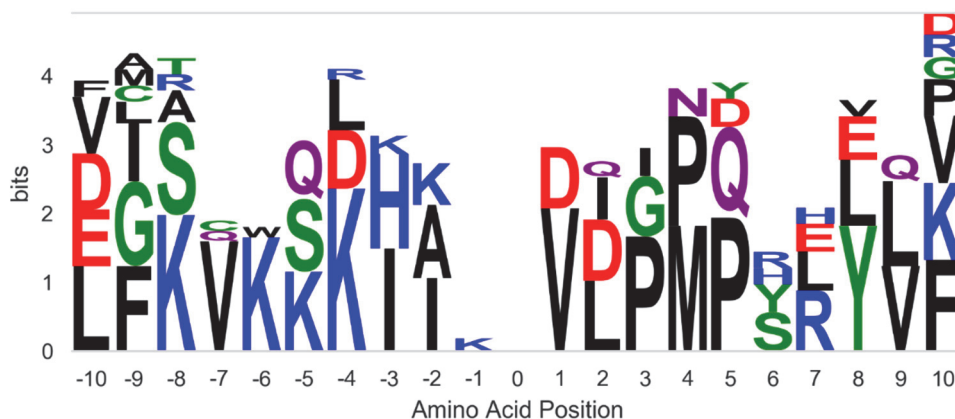


Figure 2.4 – Enriched amino acid sequence motif surrounding glycosylated lysine residues in bovine milk proteins. Position probability matrix calculated based on the subtraction of position frequency matrices of the 63 non-glycosylated lysine residues (<5% glycosylation site occupancy) from the 22 majorly glycosylated lysine residues ($\geq 5\%$ glycosylation site occupancy) in the six most abundant milk proteins in the aged (159 d at 40 °C) skim milk powder. Position 0 indicates the glycosylated lysine residues; positions -10 to -1 and 1 to 10 indicate the 10 amino acids N-terminal and C-terminal of the glycosylated lysine in the amino acid sequence, respectively.

during thermal processing and considerably increasing during storage of the powders (Figure 2.1). Analysis of the four key samples, *i.e.*, the thermised milk, the preheated milk, the fresh skim milk powder, and the aged skim milk powder (159 d at 40 °C), using a hybrid mass spectrometry approach combining protein-centric LC-MS and MS/MS (Figure 2.2) with peptide-centric multi-protease LC-MS/MS (Figure 2.3), yielded an in-depth view of glycation on the main bovine milk proteins. The protein-centric mass spectrometry results gave a quantitative overview of the different proteoforms and of their extent of glycation, while the peptide-centric approach complemented the aforementioned results with qualitative information on the localization of the glycation hot-spots.

We clearly observed that not all lysine residues are equally targeted by glycation. Among the glycation hot-spots we identified, several were already known, but novel ones gain visibility thanks to our experimental design. Analysing milk protein glycation solely in a tryptic digest is biased against identifying glycation occurring in lysine and arginine-rich regions, such as residues Lys₁₃₅, Lys₁₃₈ and Lys₁₄₁ in β -lactoglobulin, Lys₁₀₂, Lys₁₀₃ and Lys₁₀₅ in α_{s1} -casein, Lys₂₁ and Lys₂₄ in α_{s2} -casein, Lys₂₈, Lys₂₉ and Lys₃₂ in β -casein or Lys₁₁₁, Lys₁₁₂ and Lys₁₁₆ in κ -casein. While only poor coverage of these sites is typically achieved relying on trypsin hydrolysis (Milkovska-Stamenova and Hoffmann, 2016a, Milkovska-Stamenova and Hoffmann, 2016b), our multi-protease approach revealed these protein regions to

harbour several glycation hot-spots, as depicted in Figure 2.3. Conversely, glycation hot-spots located on tryptic macropeptides that are outside the range of conventional peptide-centric LC-MS/MS methods, are often missed as well. This is the case of residues Lys₇₀ and Lys₁₆₉ that we here reveal to be two of the main glycation sites in α_{s2} -casein and β -casein, respectively. Lys₇₀ is located on the α_{s2} -casein tryptic 46-70; its glycated form results in a missed cleavage at Lys₇₀ and ends up on the longer tryptic peptide 46-76 of 31 residues in length. Lys₁₆₉ of β -casein is located on an even longer tryptic macropeptide of residues 114-169 (56 amino acids); glycation and missed cleavage at Lys₁₆₉ puts it on the longer tryptic peptide 114-176 of 63 amino acid residues in length. While these glycation hot-spots were also missed in trypsin-based peptide-centric approaches (Milkovska-Stamenova and Hoffmann, 2016a, Milkovska-Stamenova and Hoffmann, 2016b), we have shed light on them here on account of the high sequence coverage (supplementary Figure S2.4) obtained with our multi-protease approach.

Identification and relative quantitation of the glycation hot-spots that develop during production and storage of skim milk powder still leaves the question unanswered as to why specific sites are favoured over others. At the primary sequence level, we reveal in this study, for the first time, a preferred motif where the glycation hot-spot in milk proteins is preceded to the N-terminal by positively charged amino-acids (lysine and histidine residues), and to the C-terminal it is followed mostly by aliphatic and hydrophobic amino acids (valine, leucine and isoleucine, proline and methionine residues; see Figure 2.4). The study of Johansen *et al.* (2006) was one of the first to analyse a putative glycation motif in a combined dataset of 20 proteins, comprising 89 glycated and 126 non-glycated lysine residues, originating from *in vivo* and *in vitro* mammalian studies. Despite considerable differences in the study systems and conditions, their findings indicate the occurrence of basic residues to the N-terminal and of acidic residues to the C-terminal of the glycated lysine, in agreement with our data. Furthermore, Johansen *et al.* (2006) argued that basic and acidic residues located in the proximity of the glycation hot-spot, either on the primary sequence or on the 3D structure, act as catalysts for the Amadori rearrangement. A much broader study in a similar direction from Zhang *et al.* (2011) investigated the glycation motif occurring in over 3700 blood plasma and erythrocyte proteins due to diabetes. Although less pronounced, their primary sequence motif also supports the enrichment of basic residues N-terminal, and of aliphatic and acidic residues C-terminal of glycation hot-spots, as also observed in our study (Figure 2.4).

Next to revealing the preferential motif in the primary sequence of the milk proteins, we considered whether higher-order structural features may also

correlate to the glycation hot-spots revealed in our study. The hollow β -barrel structure with a hydrophobic pocket of β -lactoglobulin (PDB entry 1BEB; Brownlow, 1997) places most of the hydrophilic residues at the surface of the molecule, rendering the majority of the lysine residues accessible for glycation in the globular protein. However, as depicted in Figure 2.1B, most of the β -lactoglobulin no longer entered the non-reducing gel following preheating, indicating that the protein unfolded and likely formed disulphide-linked aggregates. Denaturation and unfolding of the molecule can further expose reactive lysine residues, increasing their accessibility for lactosylation, explaining the high degree of glycation that was observed for β -lactoglobulin (Figure 2.2B, Figure 2.3B). While also a globular protein, the structure of α -lactalbumin is more compact and stabilized by a Ca^{2+} ion and 4 disulphide bridges (Permyakov and Berliner, 2000), which renders it resistant to both glycation (Figure 2.2A and Figure 2.3A) and proteolytic degradation. As observed from the SDS-polyacrylamide gels in Figure 2.1B and C, most of the α -lactalbumin remained presumably still compactly folded following preheating of the milk, possibly explaining its observed low level of glycation. The glycation hot-spots detected in α -lactalbumin (Figure 2.3A), *i.e.*, residues Glu₁, Lys₆₂ and Lys₉₈, and to a lesser extent Lys₁₁₄, all have surface-exposed side-chains that are not involved in the secondary structure of the protein (PDB entry 1F6S; Chrysin, Brew & Acharya, 2000). The α _s-caseins underwent the highest levels of glycation (Figure 2.2C and D, and Figure 2.3C and D) despite their localization in the core of the casein micelles (Huppertz *et al.*, 2017). This can possibly be attributed to the scarcity of casein secondary structure (Farrell *et al.*, 2013) combined with the open and porous structure of the casein micelles (Bouchoux *et al.*, 2010), allowing accessibility of lactose into the micelle.

Extensive research into casein structure has resulted in energy-minimized modelled structures for homo-oligomers and submicellar casein particles (Farrell *et al.*, 2003), relative to which the glycation hot-spots can be considered. Glycation of α _{s1}-casein was only detected in the N-terminal part of the protein, up until Lys₁₀₅ (Figure 2.3C). The C-terminal part of the protein is for the most part involved in intermolecular interactions (Farrell Jr *et al.*, 2009, Kumosinski *et al.*, 1994a, Kumosinski *et al.*, 1994b). The resulting limited solvent and lactose accessibility may explain the absence of glycation observed in this region of the protein. The lysine residues in κ -casein are found for the most part in the para- κ -casein region, which is also the region of interaction with the casein micelle, limiting putative glycation of the protein. Lys₂₄ has high solvent exposure (Kumosinski *et al.*, 1994b) and it was detected as one of the sites to undergo the highest level of glycation (Figure 3F). β -Casein is the most proline-rich of the caseins (35 proline residues out of 209 amino

acids), with therefore the least secondary structure, the most open conformation and the highest lactose accessibility. Glycation sites Lys₂₈, Lys₂₉ and Lys₃₂, and Lys₉₉, Lys₁₀₅ and Lys₁₀₇ (Figure 2.3E) are clustered close to each other and are found in 2 lobes of the protein localized polar opposite from the dimerization region (Kumosinski *et al.*, 1993). All of the sites mentioned above are solvent-exposed. Furthermore, Lys₂₈-Lys₂₉, Lys₁₀₅-His₁₀₆ and Lys₁₀₇-Glu₁₀₈ are the major plasmin hydrolysis sites in β -casein (Eigel *et al.*, 1984), they are easily accessible by this large enzyme, and consequently also accessible by lactose, and prone to glycation. Glycation sites Lys₁₆₉ and Lys₁₇₆ (Figure 2.3E) are in close proximity of the C-terminal tail of the protein involved in intermolecular interactions (Kumosinski *et al.*, 1994a). Residues Lys₁₆₉ and Lys₁₇₆ are solvent-exposed within a large open cavity in the β -casein molecule, allowing accessibility of lactose molecules.

In light of these findings, rather than an absolute affinity of the reducing carbohydrate molecules to the amino acid sequence, we believe the sequence motif to be correlated with protein conformation. Thereby, lysine residues found in this motif are likely to be surface-exposed in the structure of the proteins, and consequently accessible for glycation.

Materials and Methods

All chemical and biochemical reagents were purchased from Sigma-Aldrich (Germany), unless otherwise specified.

Milk samples

Skim milk powder was produced on pilot scale at the FrieslandCampina Innovation Centre in Wageningen, The Netherlands, following the flowchart shown in Figure 2.1A. The batch of samples was prepared from mature bulk milk from Dutch Friesian Holstein cows. Since these samples are representative of the average Dutch bovine milk, no additional biological replicates were analysed. Briefly, skim milk was thermised and standardized to 35% (*m/m*) protein in dry matter. The standardized milk was preheated for 60 s at 108 °C. The preheated milk was further concentrated to 48-50% (*m/m*) dry matter, followed by spray-drying (inlet air temperature = 205 °C; outlet air temperature = 90 °C) resulting in <0.2 water activity in the powder. The skim milk powder was sealed in tin cans that were further subjected to a storage trial at either 20 or 40 °C for 43, 77, and 159 d. Aliquots of fresh and aged powders were stored at -20 °C until further analysis.

Gel electrophoresis and analysis of blocked lysine residues

All intermediate, final and aged products shown in Figure 2.1A were subjected to sodium dodecyl sulphate – polyacrylamide gel electrophoresis (SDS-PAGE) under

both reducing (with dithiothreitol, DTT) and non-reducing (without DTT) conditions. The samples were treated with XT sample buffer, loaded onto 26-well 12% Criterion™ XT Bis-Tris Precast Gels, and electrophoresis was run in XT MOPS running buffer, all purchased from Bio-Rad (Veenendaal, The Netherlands). Precision Plus Protein™ Dual Color Standards (Bio-Rad, The Netherlands) were run on the gels in parallel with the samples for protein size reference. The gels were stained with GelCode Blue (Thermo Fisher Scientific, Landsmeer, The Netherlands). The detailed experimental procedure is described in the Supplementary material.

Blocked (*i.e.*, glycated) lysine was determined for the thermised and preheated milk and for the fresh and aged (159 d at 40 °C) skim milk powders following previously described procedures by Nyakayiru *et al.* (2020). Acid hydrolysis of Amadori products leads to the formation of furosine, which is not endogenous to the dairy products. Lysine and furosine were quantified using ion-pair RP-HPLC following acid hydrolysis (Delgado *et al.*, 1992). Krause *et al.* (2003) determined a molar yield of 34% lactulosyllysine to furosine, which was used for the conversion of the measured furosine contents. Blocked lysine was calculated as lactulosyllysine expressed as a percentage of the total lysine contents of the sample.

Protein-centric mass spectrometry

The thermised and preheated milk and the fresh and aged (159 d at 40 °C) milk powders were selected for analysis by mass spectrometry. The powders were reconstituted in MilliQ at 10% (*m/m*) powder in solution. The samples were denatured and reduced according to a method modified from Bondt *et al.* (2021). LC-MS and MS/MS analyses were performed using a Thermo Scientific Vanquish Flex UHPLC system (Thermo Fisher Scientific, Germering, Germany) connected online to an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) operated in intact protein mode and at low pressure. Reversed-phase separation was achieved with a MAbPac RP 2.1 mm × 50 mm column (Thermo Fisher Scientific, Germany). Each sample was analysed with a triplet of complementary methods comprising full MS, MS/MS with electron transfer dissociation (ETD) and MS/MS with electron transfer dissociation supplemented by higher-energy collisional dissociation (ETHcD). No additional technical replicates were performed. LC-MS raw data were deconvoluted with the BioPharma Finder 3.2 Software using the Xtract algorithm (Thermo Fisher Scientific, CA, USA). The database for searching the LC-MS/MS results was generated based on the six most abundant bovine milk proteins downloaded in XML format from UniProtKB (accession numbers P00711, P02754, P02662, P02663, P02666 and P02668) and imported from the XML format file into ProSightPC. The LC-MS/MS results were automatically searched in Thermo Proteome Discoverer (v 2.4.0.305)

using the ProSightPD nodes for High-High experimental workflows. A library of proteoforms and corresponding monoisotopic masses was generated to include the most commonly occurring genetic variants and expected phosphorylation states of the major bovine milk proteins in Dutch Holstein-Friesian cows. For each basic proteoform, the masses of glycated proteoforms were generated in a range of 1-12 hexose residues at increments of 1 hexose residue. Proteoforms were annotated based on LC-MS/MS database search identification and matching between the experimental and theoretical masses. Sample preparation, LC-MS and MS/MS analyses, database generation, database search, library generation and data analysis are described in detail in the Supplementary material.

Peptide-centric mass spectrometry

The four samples mentioned above were also analysed by a peptide-centric mass spectrometry approach. The samples were denatured, reduced and alkylated, followed by proteolytic digestion with trypsin (Promega, Madison, WI, USA), chymotrypsin, GluC, AspN (latter three purchased from Roche, Mannheim, Germany), and combinations thereof as described in Table 2.1. Since each sample was therefore analysed from a total of seven different and complementary proteolytic digests, no additional technical replicates were performed. Following proteolytic digestion, the resulting peptides were extracted by solid-phase extraction using Oasis HLB 96-well plates (Waters, Etten-Leur, The Netherlands). The samples were analysed using an Agilent 1290 Infinity HPLC system (Agilent Technologies, Middelburg, The Netherlands) coupled online to a Q Exactive HF hybrid quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) according to a method modified from Zhu *et al.* (2020). The samples were separated by reversed-phase chromatography using in-house packed trapping (2 cm length, 100 µm inner diameter; ReproSil-Pur C18-AQ, 3 µm; Dr. Maisch GmbH, Ammerbuch, Germany) and analytical columns (50 cm length, 50 µm inner diameter; Poroshell 120 EC-C18, 2.7 µm; Agilent Technologies, The Netherlands). LC-MS/MS raw data was searched with Thermo Proteome Discoverer V2.4 (Thermo Fisher Scientific, Germany) using the Mascot V2.6 (Matrix Science, Boston, MA, USA) search engine. To ensure high identification rate, the modifications taken into consideration in the database search included glycation with mono- and dihexose of protein N-termini, lysine and arginine residues, phosphorylation of serine and threonine residues, oxidation of methionine residues, cyclization of N-terminal glutamine residues to pyroglutamic acid, and fixed carbamidomethylation of cysteine residues. The data was carefully post-processed, including filtering of unreliable identifications and normalization across digests and injections, prior to combining into a comprehensive picture. Glycation site occupancy was calculated

as the sum of abundances of all peptide isoforms covering a glycated site expressed as a percentage of the total sum of abundances of all peptide isoforms covering the respective site, whether glycated or not. Sample preparation, LC-MS/MS analysis, database search and data processing are described in detail in the Supplementary material.

Glycated lysine motif analysis

Glycation motif analysis was performed for the six most abundant milk proteins on the reference sequences from UniProtKB. The signal peptides were excluded to focus the analysis on the mature proteins present in the milk. The glycated sites chosen for the analysis were lysine residues, whereby the glycation site occupancy was detected at >5% of relative abundance in the aged milk powder. The sequence logos were constructed for the glycated and non-glycated lysine residues in the six most abundant milk proteins, using standard position weighted matrices (Hertz and Stormo, 1999). For both the glycated and non-glycated lysine residues, their sequence motifs were extracted for the ten amino acids preceding and following them. The sequence motif of the non-glycated lysine residues was subsequently subtracted from that of the glycated ones to highlight the differences better.

References

- Aalaei K, Rayner M, Sjöholm I. 2019. Chemical methods and techniques to monitor early Maillard reaction in milk products; A review. *Crit Rev Food Sci Nutr*, 59:1829-1839.
- Arena S, Renzone G, D'Ambrosio C, Salzano AM, Scaloni A. 2017. Dairy products and the Maillard reaction: A promising future for extensive food characterization by integrated proteomics studies. *Food Chem*, 219:477-489.
- Baye E, de Courten MPJ, Walker K, Ranasinha S, Earnest A, Forbes JM, de Courten B. 2017. Effect of dietary advanced glycation end products on inflammation and cardiovascular risks in healthy overweight adults: a randomised crossover trial. *Sci Rep*, 7:1-6.
- Birlouez-Aragon I, Pischetsrieder M, Leclere J, Morales FJ, Hasenkopf K, Kientsch-Engel R, Ducauze CJ, Rutledge D. 2004. Assessment of protein glycation markers in infant formulas. *Food Chem*, 87:253-259.
- Bondt A, Hoek M, Tamara S, de Graaf B, Peng W, Schulte D, van Rijswijck DMH, den Boer MA, Greisch J-F, Varkila MRJ, *et al.* 2021. Human plasma IgG1 repertoires are simple, unique, and dynamic. *Cell Systems*.

- Boor KJ, Wiedmann M, Murphy S, Alcaine S. 2017. A 100-year review: microbiology and safety of milk handling. *J Dairy Sci*, 100:9933-9951.
- Bouchoux A, Gesan-Guiziou G, Pérez J, Cabane B. 2010. How to squeeze a sponge: casein micelles under osmotic stress, a SAXS study. *Biophys J*, 99:3754-3762.
- Brown BA, Zeng X, Todd AR, Barnes LF, Winstone JMA, Trinidad JC, Novotny MV, Jarrold MF, Clemmer DE. 2020. Charge detection mass spectrometry measurements of exosomes and other extracellular particles enriched from bovine milk. *Anal Chem*, 92:3285-3292.
- Brownlow S, Cabral JHM, Cooper R, Flower DR, Yewdall SJ, Polikarpov I, North ACT, Sawyer L. 1997. Bovine β -lactoglobulin at 1.8 Å resolution—still an enigmatic lipocalin. *Structure*, 5:481-495.
- Cardoso HB, Wierenga PA, Gruppen H, Schols HA. 2018. Maillard induced glycation behaviour of individual milk proteins. *Food Chem*, 252:311-317.
- Carulli S, Calvano CD, Palmisano F, Pischetsrieder M. 2011. MALDI-TOF MS characterization of glycation products of whey proteins in a glucose/galactose model system and lactose-free milk. *J Agric Food Chem*, 59:1793-1803.
- Cattaneo S, Stuknytė M, Masotti F, De Noni I. 2017. Protein breakdown and release of β -casomorphins during in vitro gastro-intestinal digestion of sterilised model systems of liquid infant formula. *Food Chem*, 217:476-482.
- Chevalier F, Chobert JM, Dalgalarondo M, Choiset Y, Haertlé T. 2002. Maillard glycation of β -lactoglobulin induces conformation changes. *Nahrung*, 46:58-63.
- Chrysina ED, Brew K, Acharya KR. 2000. Crystal structures of apo- and holo-bovine α -lactalbumin at 2.2-Å resolution reveal an effect of calcium on inter-lobe interactions. *J Biol Chem*, 275:37021-37029.
- Delatour T, Hegele J, Parisod V, Richoz J, Maurer S, Steven M, Buetler T. 2009. Analysis of advanced glycation endproducts in dairy products by isotope dilution liquid chromatography–electrospray tandem mass spectrometry. The particular case of carboxymethyllysine. *Journal of Chromatography A*, 1216:2371-2381.
- Delgado T, Corzo N, Santa-María G, Jimeno M, Olano A. 1992. Determination of furosine in milk samples by ion-pair reversed phase liquid chromatography. *Chromatographia*, 33:374-376.

Deng Y, Wierenga PA, Schols HA, Sforza S, Gruppen H. 2017. Effect of Maillard induced glycation on protein hydrolysis by lysine/arginine and non-lysine/arginine specific proteases. *Food Hydrocolloids*, 69:210-219.

Dupont D, Croguennec T, Brodkorb A, Kouaouci R. 2013. Quantitation of proteins in milk and milk products. In: McSweeney PLH, Fox PF editors. *Advanced Dairy Chemistry: Springer*. p. 87-134.

Eigel WN, Butler JE, Ernstrom CA, Farrell Jr HM, Harwalkar VR, Jenness R, Whitney RM. 1984. Nomenclature of proteins of cow's milk: fifth revision. *J Dairy Sci*, 67:1599-1631.

Farrell HM, Brown EM, Hoagland PD, Malin EL. 2003. Higher order structures of the caseins: a paradox? In: Fox PF, McSweeney PLH editors. *Advanced Dairy Chemistry—1 Proteins*. Boston, MA, USA: Springer. p. 203-231.

Farrell HM, Brown EM, Malin EL. 2013. Higher order structures of the caseins: A paradox? In: McSweeney PLH, Fox PF editors. *Advanced Dairy Chemistry: Springer*. p. 161-184.

Farrell Jr HM, Jimenez-Flores R, Bleck GT, Brown EM, Butler JE, Creamer LK, Hicks CL, Hollar CM, Ng-Kwai-Hang KF, Swaisgood HE. 2004. Nomenclature of the proteins of cows' milk—Sixth revision. *J Dairy Sci*, 87:1641-1674.

Farrell Jr HM, Malin EL, Brown EM, Mora-Gutierrez A. 2009. Review of the chemistry of α S2-casein and the generation of a homologous molecular model to explain its properties. *J Dairy Sci*, 92:1338-1353.

Franc V, Zhu J, Heck AJR. 2018. Comprehensive proteoform characterization of plasma complement component C8 α β γ by hybrid mass spectrometry approaches. *J Am Soc Mass Spectrom*, 29:1099-1110.

Fuerer C, Jenni R, Cardinaux L, Andetsion F, Wagnière S, Moulin J, Affolter M. 2020. Protein fingerprinting and quantification of β -casein variants by ultra-performance liquid chromatography–high-resolution mass spectrometry. *J Dairy Sci*, 103:1193-1207.

Gómez-Narváez F, Perez-Martinez L, Contreras-Calderon J. 2019. Usefulness of some Maillard reaction indicators for monitoring the heat damage of whey powder under conditions applicable to spray drying. *Int Dairy J*, 99:104553.

Gonzales AP, Naranjo G, Leiva G, Malec L. 2010. Maillard reaction kinetics in milk powder: Effect of water activity at mild temperatures. *Int Dairy J*, 20:40-45.

Hegele J, Buetler T, Delatour T. 2008. Comparative LC–MS/MS profiling of free and protein-bound early and advanced glycation-induced lysine modifications in dairy products. *Anal Chim Acta*, 617:85-96.

Hertz GZ, Stormo GD. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563-577.

Huppertz T, Gazi I, Luyten H, Nieuwenhuijse H, Alting A, Schokker E. 2017. Hydration of casein micelles and caseinates: Implications for casein micelle structure. *Int Dairy J*, 74:1-11.

Johansen MB, Kiemer L, Brunak S. 2006. Analysis and prediction of mammalian protein glycation. *Glycobiology*, 16:844-853.

Krause R, Knoll K, Henle T. 2003. Studies on the formation of furosine and pyridosine during acid hydrolysis of different Amadori products of lysine. *European Food Research and Technology*, 216:277-283.

Kumosinski TF, Brown EM, Farrell Jr HM. 1993. Three-dimensional molecular modeling of bovine caseins: an energy-minimized β -casein structure. *J Dairy Sci*, 76:931-945.

Kumosinski TF, Brown EM, Farrell Jr HM. 1994a. Predicted energy-minimized α 1-casein working model. In: Kumosinski TF, Liebman MN editors. *Molecular Modeling, from Virtual Tools to Real Problems*. Washington, DC, USA: ACS Symposium Series. p. 368-391.

Kumosinski TF, King G, Farrell HM. 1994b. An energy-minimized casein submicelle working model. *J Protein Chem*, 13:681-700.

Leiva GE, Naranjo GB, Malec LS. 2017. A study of different indicators of Maillard reaction with whey proteins and different carbohydrates under adverse storage conditions. *Food Chem*, 215:410-416.

Milkovska-Stamenova S, Hoffmann R. 2016a. Hexose-derived glycation sites in processed bovine milk. *J Proteomics*, 134:102-111.

Milkovska-Stamenova S, Hoffmann R. 2016b. Identification and quantification of bovine protein lactosylation sites in different milk products. *J Proteomics*, 134:112-126.

Nyakayiru J, van Lieshout GAA, Trommelen J, van Kranenburg J, Verdijk LB, Bragt MCE, van Loon LJC. 2020. The glycation level of milk protein strongly modulates post-prandial lysine availability in humans. *Br J Nutr*, 123:545-552.

Oliver CM. 2011. Insight into the glycation of milk proteins: an ESI-and MALDI-MS perspective. *Crit Rev Food Sci Nutr*, 51:410-431.

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

Permyakov EA, Berliner LJ. 2000. α -Lactalbumin: structure and function. *FEBS Lett*, 473:269-274.

Renzone G, Arena S, Scaloni A. 2015. Proteomic characterization of intermediate and advanced glycation end-products in commercial milk samples. *J Proteomics*, 117:12-23.

Schuck P, Blanchard E, Dolivet A, Méjean S, Onillon E, Jeantet R. 2005. Water activity and glass transition in dairy ingredients. *Le Lait*, 85:295-304.

Sharma SD, Barone M. 2019. Deleterious Consequences of Dietary Advanced Glycation End Products on Human Health Due to Oxidative Stress and Inflammation. *Dietary Patterns, Food Chemistry and Human Health*: Springer. p. 1-13.

Siciliano RA, Mazzeo MF, Arena S, Renzone G, Scaloni A. 2013. Mass spectrometry for the analysis of protein lactosylation in milk products. *Food Res Int*, 54:988-1000.

Tamara S, Franc V, Heck AJR. 2020. A wealth of genotype-specific proteoforms fine-tunes hemoglobin scavenging by haptoglobin. *Proceedings of the National Academy of Sciences*, 117:15554-15564.

Teodorowicz M, Van Neerven J, Savelkoul H. 2017. Food processing: The influence of the maillard reaction on immunogenicity and allergenicity of food proteins. *Nutrients*, 9:835.

Toda M, Hellwig M, Henle T, Vieths S. 2019. Influence of the Maillard reaction on the allergenicity of food proteins and the development of allergic inflammation. *Curr Allergy Asthma Rep*, 19:4.

Tsiatsiani L, Heck AJR. 2015. Proteomics beyond trypsin. *The Febs Journal*, 282:2612-2626.

van Boekel M. 1998. Effect of heating on Maillard reactions in milk. *Food Chem*, 62:403-414.

van de Waterbeemd M, Tamara S, Fort KL, Damoc E, Franc V, Bieri P, Itten M, Makarov A, Ban N, Heck AJR. 2018. Dissecting ribosomal particles throughout the kingdoms of life using advanced hybrid mass spectrometry methods. *Nature Communications*, 9:1-12.

van Lieshout GAA, Lambers TT, Bragt MCE, Hettinga KA. 2020. How processing may affect milk protein digestion and overall physiological outcomes: A systematic review. *Crit Rev Food Sci Nutr*, 60:2422-2445.

Yang Y, Liu F, Franc V, Halim LA, Schellekens H, Heck AJR. 2016. Hybrid mass spectrometry approaches in glycoprotein analysis and their usage in scoring biosimilarity. *Nature Communications*, 7:1-10.

Zhang Q, Monroe ME, Schepmoes AA, Clauss TRW, Gritsenko MA, Meng D, Petyuk VA, Smith RD, Metz TO. 2011. Comprehensive identification of glycosylated peptides and their glycation motifs in plasma and erythrocytes of control and diabetic subjects. *J Proteome Res*, 10:3076-3088.

Zhu J, Lin Y-H, Dingess KA, Mank M, Stahl B, Heck AJR. 2020. Quantitative longitudinal inventory of the N-glycoproteome of human milk from a single donor reveals the highly variable repertoire and dynamic site-specific changes. *J Proteome Res*, 19:1941-1952.

Overview of supplementary information for this chapter

Supplementary materials and methods

Gel electrophoresis

Protein-centric mass spectrometry

Sample preparation

LC-MS and LC-MS/MS analysis of samples

Data analysis

Peptide-centric mass spectrometry

In solution proteolytic digestion of milk proteins

LC-MS/MS analysis of samples

Database search and analysis of results

References

Supplementary tables

Supplementary figures

Supplementary materials and methods

Gel electrophoresis

The thermised and standardized milk samples were diluted 13 times with ultrapure MilliQ water (Merck Millipore, Darmstadt, Germany). The preheated milk and the concentrated milk were diluted 19 and 66 times with MilliQ, respectively. The powder samples were reconstituted in MilliQ at 10% *m/m* powder in solution, followed by 13 times dilution with MilliQ. The diluted samples were mixed in a ratio of 1:3 v/v with the XT sample buffer (Bio-Rad, Veenendaal, The Netherlands). For the samples run under reducing conditions, dithiothreitol (DTT) was added to the XT sample buffer to a final concentration of 100 mM. The samples run under non-reducing and reducing conditions were incubated at 50 and 95 °C, respectively, for 10 min under mixing. Following incubation, the samples were centrifuged for 10 min at 20,000 × *g* and room temperature, and the cream layer and/or sediment were avoided during sampling. No protein sediment formed during centrifugation; the purpose of the centrifugation step was to sediment potential impurities (*e.g.*, dust particles) and to avoid their injection onto the LC column. A volume of 5 µL of each sample was loaded onto a 26-well 12% Criterion™ XT Bis-Tris Precast Gel (Bio-Rad, The Netherlands). Precision Plus Protein™ Dual Color Standards (Bio-Rad, The Netherlands) were run on the gels in parallel with the samples for protein size reference. The cell was filled with 500 mL of XT MOPS running buffer (Bio-Rad, The Netherlands) pre-diluted 20 times with MilliQ. Electrophoresis was carried out for 10 min at 80 V followed by approximately 1 h at 200 V until the dye front reached the bottom of the gel. The gels were placed in fixing solution (40% ethanol + 10% acetic acid in MilliQ) for 30 min immediately after stopping electrophoresis. The fixed gels were stained for 1 h in GelCode Blue (Thermo Fisher Scientific, Landsmeer, The Netherlands). The stained gels were rinsed overnight in MilliQ followed by scanning using an Amersham Imager 600 (GE Healthcare, North Richland Hills, TX, USA).

Protein-centric mass spectrometry

Sample preparation

The samples were denatured and reduced according to a method modified from Bondt *et al.* (2021). The skim milk powders were reconstituted in MilliQ water at 10% *m/m* powder in solution. For all liquid samples, a volume of 12.5 µL of milk was diluted with 67.5 µL of MilliQ, except for the preheated milk where 8.75 µL of milk were diluted with 71.25 µL of MilliQ. A volume of 10 µL of 100 mM TCEP was added to each of the diluted milk samples. Following the addition of TCEP the samples were acidified with 10 µL of 10% FA. The samples were vortexed in-between and after the addition of each reagent and then incubated for 40 min at 60 °C under

constant mixing to ensure denaturation and reduction of the proteins. Following incubation, the samples were cooled in an ice bath and stored at -20 °C until analysis. Prior to analysis, the samples were centrifuged at 20.000 × g and 4 °C for 15 min. The sediment and/or cream layer were avoided during sampling.

LC-MS and LC-MS/MS analysis of samples

The samples were analysed using a Thermo Scientific Vanquish Flex UHPLC system (Thermo Fisher Scientific, Germering, Germany) connected online to an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) operated in intact protein mode and at low pressure. Reversed-phase separation was performed with a MAbPac RP 2.1 mm × 50 mm column (Thermo Fisher Scientific, Germany) equilibrated at 40 °C. A volume of 4 µL of the denatured and reduced sample was injected *per* single LC-MS(/MS) run. Mobile-phase solvent A consisted of 0.1% FA in water, and mobile-phase solvent B consisted of 0.1% FA in ACN. Total method runtime was 27 min, with the gradient as follows: 5 min at 10% B, 1 min increase from 10 to 31% B, 14 min ramp from 31 to 41% B, 1 min increase from 41 to 95% B, 1 min washing the column with 95% B, 1 min decrease from 95 to 10% B and 4 min column equilibration at 10% B. The electrospray voltage was set to 0 V for the first 5 min to prevent introducing salts and other small compounds into the mass spectrometer. Each sample was analysed with three methods: full MS, MS/MS with electron transfer dissociation (ETD) and MS/MS with electron transfer dissociation activated by higher energy collision (ETHcD). All MS scans were acquired at 120,000 resolution in an m/z range of 400-3,000 Th with an AGC target of 1e6. Maximum IT was defined at 250 ms with 5 microscans recorded in the full MS method and 2 microscans at MS1 level in the MS/MS methods. Source-induced dissociation (SID) voltage was set to 15 V. All data-dependent MS/MS scans were acquired at 120,000 resolution with an AGC target of 5e6, maximum IT of 246 ms and 5 microscans. In both MS/MS methods ETD reaction time was defined at 16 ms, ETD reagent target was set to 1e6 and maximum ETD reagent IT was set to 200 ms. In the ETHcD MS/MS method a supplemental activation (SA) collision energy of 15% was used.

Data analysis

Isotopically resolved spectra from the full MS experiments were deconvoluted with the BioPharma Finder 3.2 Software (Thermo Fisher Scientific, CA, USA) using the Xtract algorithm. The Xtract parameters used are as follows: signal-to-noise (S/N) threshold 3, mass range m/z 400-3,000, charge range 2-50 and minimum number of detected charge states of 3. The source spectra were generated using the sliding windows algorithm with the following parameters: merge tolerance of 30 ppm,

maximum retention time gap of 1 min and a minimum number of detected intervals of 3.

The database for searching the LC-MS/MS results was generated based on the six most abundant bovine milk proteins downloaded in XML format from UniProtKB on 09/22/2020 (accession numbers P00711, P02754, P02662, P02663, P02666 and P02668). The database imported from the XML format file into ProSightPC was treated as follows: initiator methionine removal, N-terminal acetylation and other modifications contained in the XML file were allowed; up to 13 features or 70 kDa maximum mass were allowed *per* proteoform. Automatic searches were conducted in Thermo Proteome Discoverer (v 2.4.0.305) using the ProSightPD nodes for High-High experimental workflows. Two ProSightPD Annotated Proteoform nodes and the ProSightPD Subsequence Search node were run in parallel at 20, 500 and 20 ppm precursor mass tolerances, respectively. Fragment mass tolerance was in all three cases set to 20 ppm.

A library of proteoforms and corresponding monoisotopic masses was generated containing the most commonly occurring genetic variants and expected phosphorylation states of the six most abundant bovine milk proteins in Dutch Holstein-Friesian cows. For each native proteoform the masses of glycated proteoforms were generated in a range of 1-12 hexose residues with increments of 1 hexose. The parameters used for generating this library are shown in Table B1. Monoisotopic masses were calculated based on the amino acid sequence of the mature protein (amino acid residue masses obtained from http://www.matrixscience.com/help/aa_help.html) with the additional masses of the putative post-translational modifications (PTMs; PTM masses obtained from: <https://www.Sigma-Aldrich.com/life-science/proteomics/post-translational-analysis/phosphorylation/mass-changes.html>). The differences in primary sequence of the mature proteins as a function of genetic variant were adjusted as described in Farrell Jr *et al.* (2004).

The deconvoluted masses from the LC-MS files were matched within +/- 1 Da window to the precursor masses from the PrSM (proteoform spectrum matches) table of the LC-MS/MS database search results. Accession number was assigned to the proteoforms with matching masses and retention times. Additionally, the deconvoluted masses from LC-MS were matched within a window of +/- 1 Da against the aforementioned customized library of proteoforms. Where possible proteoforms were annotated based on both LC-MS/MS database search identification and on matching between experimental and theoretical masses. In cases where no identification was made by database search, proteoforms were

annotated by matching of their monoisotopic masses to theoretical proteoforms of proteins eluting in a similar retention time range. Proteoform intensity was normalized on the sum of proteoform intensities for each protein.

Peptide-centric mass spectrometry

In solution proteolytic digestion of milk proteins

A denaturation buffer was prepared consisting of 10 mM TCEP (tris(2-carboxyethyl)phosphine), 100 mM Tris, 40 mM CAA (chloroacetamide) and 1% (m/v) SDC (sodium deoxycholate). The pH was adjusted to 8.5 as necessary with Tris prior to the addition of SDC. The thermised milk and reconstituted powders were diluted 70 times with the denaturation buffer, and the preheated milk was diluted 100 times. The samples were incubated for a minimum of 10 minutes at room temperature following dilution with the denaturation buffer.

The denatured samples were digested with trypsin (Promega, Madison, WI, USA), chymotrypsin, GluC, AspN (latter 3 purchased from Roche, Mannheim, Germany), and combinations thereof with trypsin as indicated in Table 1 of the main text. Each enzyme was added at a concentration of 1 μg enzyme/100 μL of denatured milk sample. The samples digested with a single enzyme were incubated overnight at either room temperature or 37 °C, depending on the enzyme. The samples digested with two enzymes were incubated for 4 h with the first enzyme at either room temperature or 37 °C depending on the enzyme, followed by the addition of trypsin and further overnight incubation at 37 °C.

Following enzymatic digestion the samples were acidified with 10% TFA (trifluoroacetic acid; Fisher Scientific, Landsmeer, The Netherlands) until a pH value in the range of 1.5-2.0 was reached. The acidification step was done in order to precipitate SDC and to stop the enzymatic reaction. The precipitated SDC was sedimented by centrifugation for 20 min at 20,000 $\times g$ and room temperature. The peptides were extracted from the supernatants by solid phase extraction using Oasis HLB 96-well plates (Waters, Etten-Leur, The Netherlands) according to the instructions of the manufacturer. Following extraction, the samples were dried in a SpeedVac and stored at -20 °C until further analysis.

LC-MS/MS analysis of samples

The dried samples were reconstituted in 100 μL of 2% FA. A volume of 5 μL from this stock solution was further diluted with 182.5 μL of 2% FA and 7.5 μL of the diluted solution were injected into the LC-MS/MS. The samples were analysed using an Agilent 1290 Infinity HPLC system (Agilent Technologies, Middelburg, The Netherlands) coupled online to a Q Exactive HF hybrid quadrupole-Orbitrap mass

spectrometer (Thermo Fisher Scientific, Bremen, Germany) according to a method modified from Zhu *et al.* (2020). The samples were separated by reversed-phase chromatography using in-house packed trapping (2 cm length, 100 μm inner diameter; ReproSil-Pur C18-AQ, 3 μm ; Dr. Maisch GmbH, Ammerbuch, Germany) and analytical columns (50 cm length, 50 μm inner diameter; Poroshell 120 EC-C18, 2.7 μm ; Agilent Technologies, The Netherlands). Mobile-phase solvent A consisted of 0.1% FA in water, and mobile-phase solvent B consisted of 0.1% FA in 80% ACN. Total method runtime was 115 min, where trapping was performed for 5 min with 0% B at 5 $\mu\text{L}/\text{min}$, after which the peptides were eluted with a passively split flow of 300 nL/min as follows: 0.1 min ramp from 0 to 10% B, 94.9 min elution from 10 to 40% B, 3 min ramp from 40 to 100% B, 1 min washing of the column with 100% B, 1 min decrease from 100 to 0% B and 10 min equilibration to 0% B. MS scans were recorded at a resolution of 60,000, with an automatic gain control (AGC) target of $3e6$, 20 ms maximum injection time (IT) and a scan range of m/z 215-2,500. Data-dependent MS2 scans were recorded at a resolution of 30,000, AGC target of $1e5$, 54 ms maximum IT and a scan range of m/z 200-2000. Fragmentation was induced for the top 15 peaks using a dynamic exclusion of 16 s. Target peaks were isolated in a m/z 1.4 window and subjected to higher-energy collision-induced dissociation (HCD) at a normalized collision energy value of 27. Charge state screening was enabled, and precursors with unknown charge state or a charge state of 1 or above 6 were excluded.

Database search and analysis of results

LC-MS/MS raw data was searched with Thermo Proteome Discoverer V2.4 (Thermo Fisher Scientific, CA, USA) using the Mascot V2.6 (Matrix Science, Boston, MA, USA) search engine. In order to considerably reduce the search space, the protein database was built to consist of only the top 6 milk proteins, *i.e.*, α -lactalbumin (P00711), β -lactoglobulin (P02754), α_{s1} -casein (P02662), α_{s2} -casein (P02663), β -casein (P02666) and κ -casein (P02668). The sequences of the mature proteins, excluding the signal peptides, were downloaded from UniProtKB (<https://www.uniprot.org/>) on 03.09.2020. All samples were searched allowing for unspecific proteolytic cleavage, fragment mass tolerance of 0.05 Da and precursor mass tolerance of 10 ppm. Cysteine carbamidomethylation was set as a fixed modification. The following dynamic modifications were allowed: phosphorylation of Ser or Thr residues; oxidation of Met residues; glycation with either monohexoses or dihexoses on Lys or Arg or protein N-terminus; cyclization of N-terminal Gln to pyroglutamic acid. The results were validated using the Target Decoy PSM Validator node reducing the false discovery rate (FDR) to 1%.

Only peptide isoforms scoring 20 or higher on Mascot ion score were allowed. The peptide isoform abundances were normalized *per* protein across all samples (*i.e.*, the different milk samples and the different proteolytic digests) in order for the respective protein abundance to be identical across all samples. The protein abundance was calculated as the sum of abundances of all its peptide isoforms divided by the total number of its peptide isoforms *per* sample. All results were merged *per* milk sample, eliminating the effect of protease and setting the focus on the protein and on the milk sample. The following were calculated for all protein N-terminal, lysine and arginine sites:

- The sum of normalized abundances of all isoforms containing the respective site, *i.e.*, total abundance *per* site;
- The sum of normalized abundances of all isoforms containing the respective site glycated with a dihexose, *i.e.*, abundance of glycation with dihexose *per* site;
- The sum of normalized abundances of all isoforms containing the respective site glycated with a monohexose, *i.e.*, abundance of glycation with monohexose *per* site.

This allowed for identification of the glycation sites and for calculation of glycation site occupancy. Glycation site occupancy was calculated as the sum of abundances of all peptide isoforms covering a glycated site expressed as a percentage of the total sum of abundances of all peptide isoforms covering the respective site, whether glycated or not.

References

Bondt A, Hoek M, Tamara S, de Graaf B, Peng W, Schulte D, van Rijswijck DMH, den Boer MA, Greisch J-F, Varkila MRJ, *et al.* 2021. Human plasma IgG1 repertoires are simple, unique, and dynamic. *Cell Systems*.

Farrell Jr HM, Jimenez-Flores R, Bleck GT, Brown EM, Butler JE, Creamer LK, Hicks CL, Hollar CM, Ng-Kwai-Hang KF, Swaisgood HE. 2004. Nomenclature of the proteins of cows' milk—Sixth revision. *J Dairy Sci*, 87:1641-1674.

Zhu J, Lin Y-H, Dingess KA, Mank M, Stahl B, Heck AJR. 2020. Quantitative longitudinal inventory of the N-glycoproteome of human milk from a single donor reveals the highly variable repertoire and dynamic site-specific changes. *J Proteome Res*, 19:1941-1952.

Supplementary tables

Table S2.1 – Parameters used to generate the library of proteoforms and theoretical masses

Protein	UniProt accession/ genetic variant	Genetic variants	Number of phosphorylated amino acids	Number of hexose residues
α -lactalbumin	P00711/B	B	0–1	0–12
β -lactoglobulin	P02754/B	A, B	0	0–12
α_{S1} -casein	P02662/B	B	7–9	0–12
α_{S2} -casein	P02663/A	A	10–15	0–12
β -casein	P02666/A2	A1, A2, A3, B, I	4–5	0–12
κ -casein	P02668/A	A, B	0–3	0–12

2

Table S2.2 – Average number of hexose residues per molecule ^a

Sample	Proteoform										
	α -la B	β -lg A	β -lg B	α_{S1} -CN B-8P	α_{S1} -CN B-9P	α_{S2} -CN B-11P	α_{S2} -CN B-12P	α_{S2} -CN B-13P	β -CN A1-5P	β -CN A2-5P	κ -CN A-1P
Thermised milk	0.00	0.05	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Preheated milk	0.17	0.48	0.44	0.30	0.06	0.21	0.22	0.15	0.26	0.25	0.14
Fresh powder	0.53	1.27	0.91	0.52	0.39	0.77	0.81	0.61	0.72	0.72	0.41
Aged powder (159 d, 40 °C)	1.64	2.77	2.64	2.94	2.55	4.69	4.35	4.51	2.54	2.53	1.26

^aCalculated based on the proteoforms distributions in Fig. 2 for the most abundant genetic variants and phosphoproteoforms consistently detected in all samples.

Supplementary figures

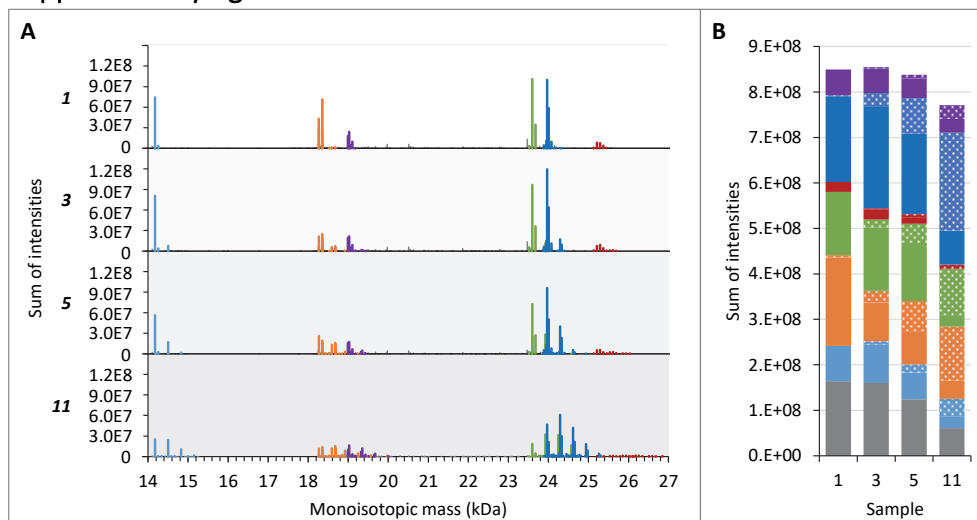


Figure S2.1 – Panel (A): Deconvoluted mass spectra from intact protein LC-MS of thermised milk (1), preheated milk (3), fresh powder (5) and aged powder (159 d at 40 °C; 11) in the mass range 14–27 kDa. — α -lactalbumin; — β -lactoglobulin; — α_{S1} -casein; — α_{S2} -casein; — β -casein; — κ -casein; — other. Panel (B): Sums of intensities of proteoforms depicted in panel A. The colour codes are the same as those used in panel A. Native and glycosylated proteoforms are indicated in solid fill (■) and pattern fill (▨), respectively. The sums of glycosylated and native proteoforms explain the greater part of the intact mass spectra, indicating that glycation was the most abundant modification occurring during production and storage of skim milk powder; other modifications, if present, occurred to a reduced extent.

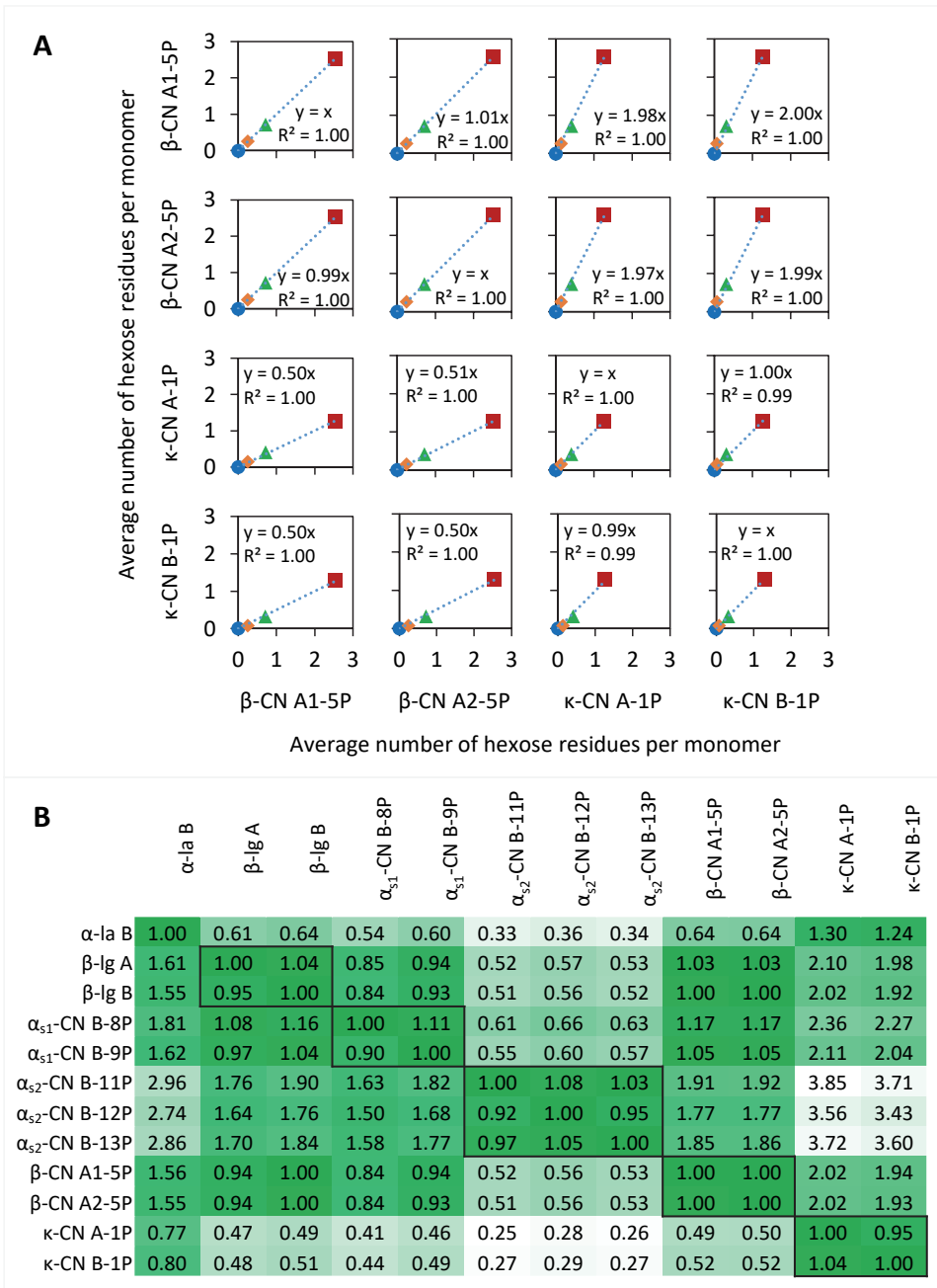


Figure S2.2 – Panel (A) Correlation plots exemplified for the average number of hexose residues per molecule of β -CN A1-5P, β -CN A2-5P, κ -CN A-1P and κ -CN A-2P in thermised milk (●), preheated milk (◆), fresh powder (▲) and aged powder (159 d at 40 °C; ■). Panel (B): Correlation matrix showing the linear regression slopes for the average number of hexose residues per molecule, as exemplified in panel A. The matrix was built on the data shown in Table S2, comprising the most abundant genetic variants and phosphoproteoforms detected consistently in all samples. Outlined in thick black borders are sets of different genetic variants and phosphoforms of the same protein.

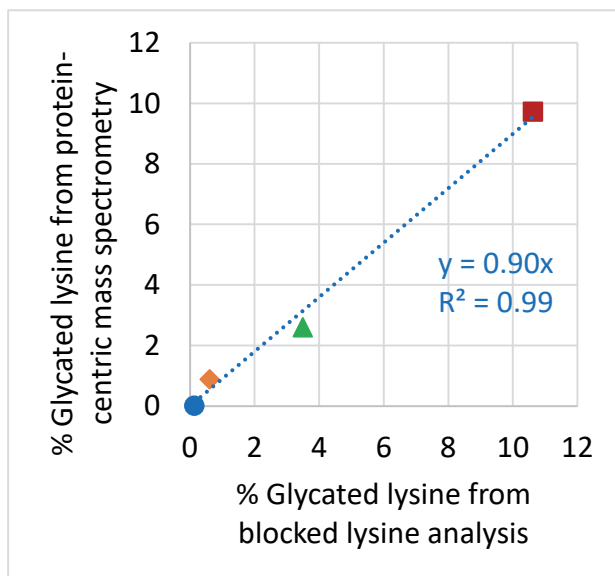


Figure S2.3 – Correlation plot between the percentages of glycated lysine determined from the blocked lysine analysis and calculated from the protein-centric mass spectrometry results, respectively, in thermised milk (●), preheated milk (◆), fresh powder (▲) and aged powder (159 d at 40 °C; ■). The percentage of glycated lysine per sample was calculated from the mass spectrometry results based only on the six most abundant milk proteins, taking into account their intensity-based relative abundances, the average number of hexose residues per protein molecule, and the total number of lysine residues per protein; for simplicity it was assumed that all glycation occurred in the form of lysine lactosylation. Indicates the best fitting linear trendline.

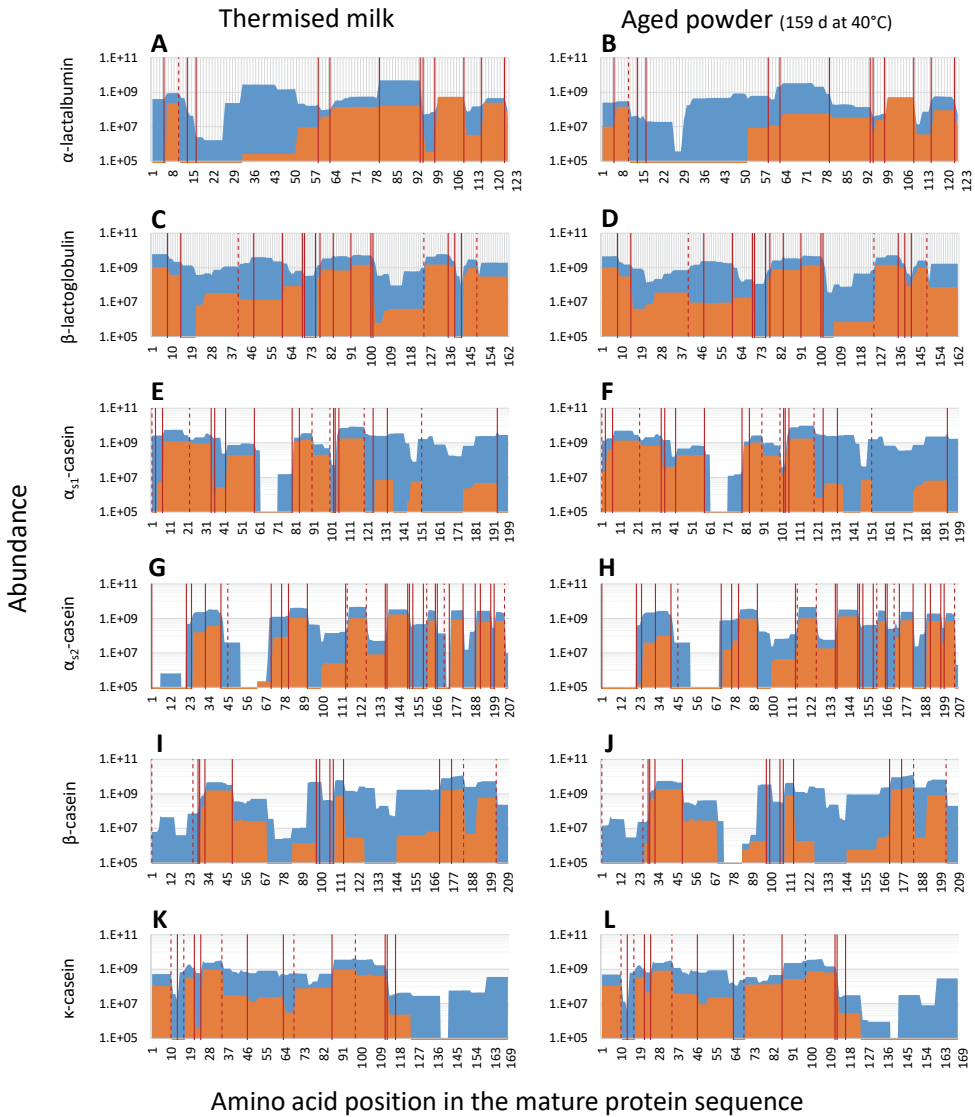
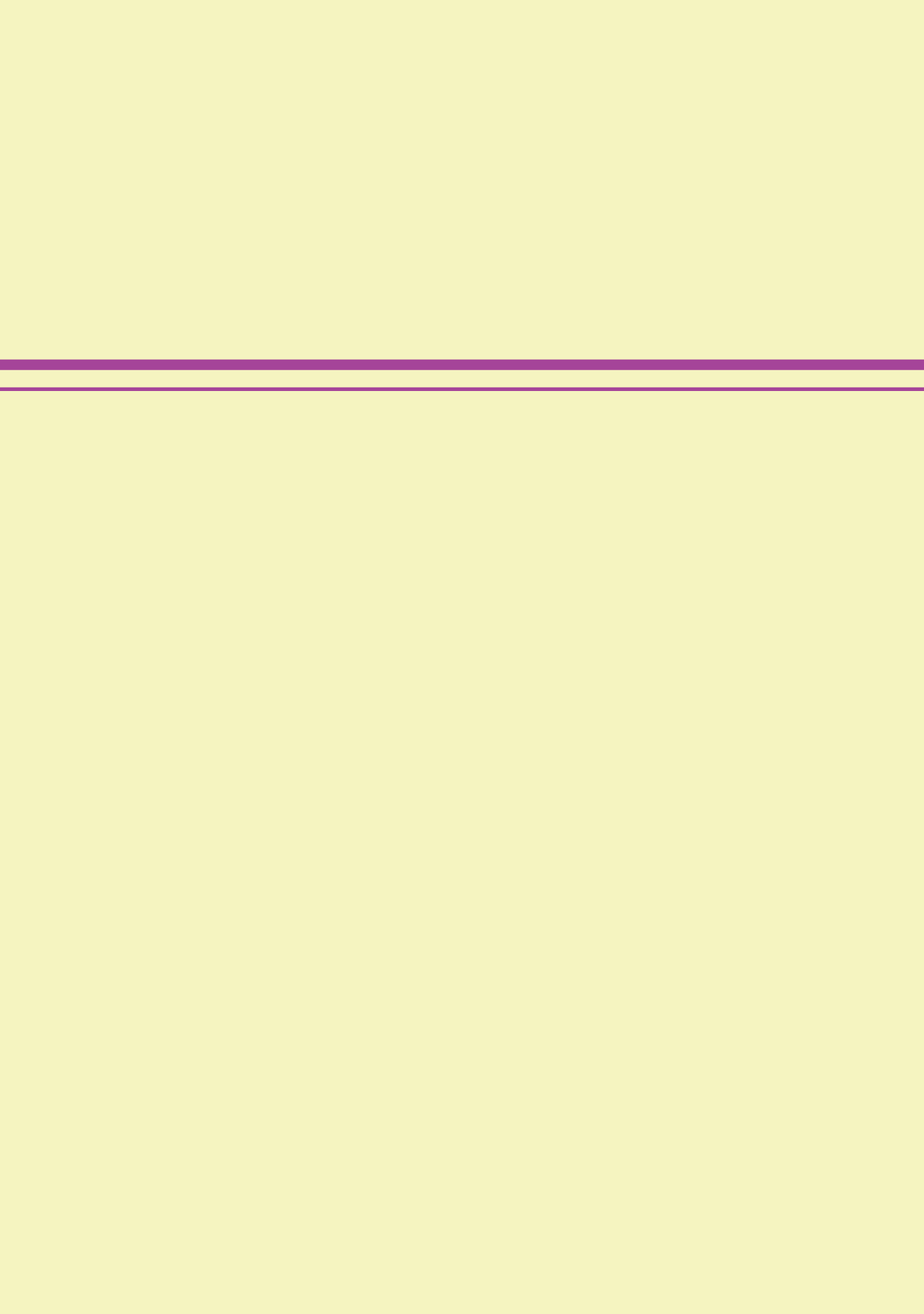


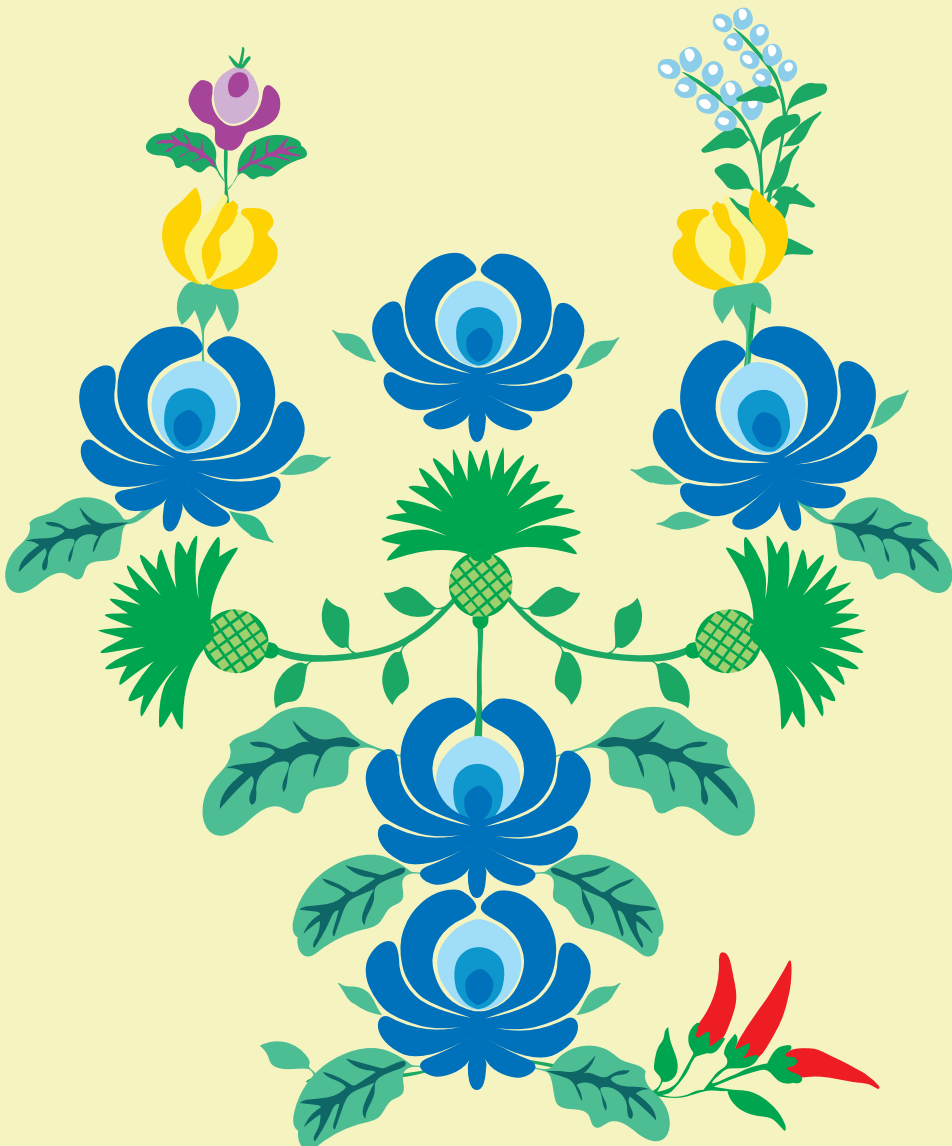
Figure S2.4 – Amino acid sequence coverage from peptide-centric LC-MS/MS of the six most abundant milk proteins: α -lactalbumin (A and B), β -lactoglobulin (C and D), α_{S1} -casein (E and F), α_{S2} -casein (G and H), β -casein (I and J) and κ -casein (K and L) in the thermised milk (A, C, E, G, I and K) and in the powder aged for 159 d at 40 °C (B, D, F, H, J and L). The differences in sequence coverage are shown between the tryptic digest (orange) and the multi-protease approach (blue). Tryptic cleavage sites (lysine and arginine residues) are indicated with solid and dashed lines, respectively.

2



CHAPTER 3

Key changes in bovine milk immunoglobulin G during lactation: NeuAc sialylation is a hallmark of colostrum immunoglobulin G *N*-glycosylation



Chapter 3. Key changes in bovine milk immunoglobulin G during lactation: NeuAc sialylation is a hallmark of colostrum immunoglobulin G *N*-glycosylation

Published as:

Gazi, I. *et al.* 2023. ***Glycobiology***, 33:115-125

<https://doi.org/10.1093/glycob/cwad001>

Abstract

We monitored longitudinal changes in bovine milk IgG in samples from four cows at 9 time points in between 0.5-28 d following calving. We used peptide-centric LC-MS/MS on proteolytic digests of whole bovine milk, resulting in the combined identification of 212 individual bovine milk protein sequences, with IgG making up >50% of the protein content of every 0.5-d colostrum sample, which reduced to $\leq 3\%$ in mature milk. In parallel, we analysed IgG captured from the bovine milk samples to characterise its *N*-glycosylation, using dedicated methods for bottom-up glycoproteomics employing product ion-triggered hybrid fragmentation; data are available via ProteomeXchange with identifier PXD037755. The bovine milk IgG *N*-glycosylation profile was revealed to be very heterogeneous, consisting of >40 glycoforms. Furthermore, these *N*-glycosylation profiles changed substantially over the period of lactation, but consistently across the four individual cows. We identified NeuAc sialylation as the key abundant characteristic of bovine colostrum IgG, significantly decreasing in the first days of lactation, and barely detectable in mature bovine milk IgG. We also report, for the first time to our knowledge, the identification of subtype IgG3 in bovine milk, alongside the better-documented IgG1 and IgG2. The detailed molecular characteristics we describe of the bovine milk IgG, and their dynamic changes during lactation, are important not only for the fundamental understanding of the calf's immune development, but also for understanding bovine milk and its bioactive components in the context of human nutrition.

Introduction

In mammals, the mammary gland secretion produced up to 3 d following birth is considered to be colostrum (McGrath *et al.*, 2016). Colostrum is compositionally and functionally very different from mature milk (McGrath *et al.*, 2016). One of the main functions of colostrum, next to providing essential nutrition, is to transfer passive immunity from the mother to the offspring (Weström *et al.*, 2020). Following the production of colostrum, the mammary secretion gradually transitions into mature milk. The composition and functionality of colostrum and milk are species-specific. With human milk constituting the golden standard to which, *e.g.*, bovine milk-based infant formula is designed, it is important to study and understand bovine milk at the molecular level, to investigate how it differs from human milk, and how it can be used to as closely as possible match its nutritional and functional qualities. While the protein composition of human colostrum is made up by up to 20% by immunoglobulins, the majority of which consist of secretory IgA (sIgA) (Atyeo and Alter, 2021), bovine colostrum has much higher proportions of immunoglobulins, *i.e.*, up to 80% of total protein, consisting

primarily of IgG (Marnila and Korhonen, 2011). In the last 2 to 3 weeks before calving, plasma levels of IgG, particularly IgG1, decrease in the cow; at the same time, IgG1 is thought to be actively transported and accumulates into the mammary gland, with maximum transfer occurring 1-3 d before calving (Delouis, 1978). The active transfer of IgG and serum proteins into bovine milk greatly diminishes and ultimately ceases over the first days after calving (Delouis, 1978), but IgG remains present in the bovine milk throughout the entire lactation period, presumably through passive transfer or local production in the mammary gland.

IgG contains an *N*-glycosylation motif in the CH2 domain of the heavy chain, which is conserved between the different mammalian species (Raju *et al.*, 2000). Protein glycosylation is a common and complex post-translational modification that controls often many biological pathways (Schjoldager *et al.*, 2020). Glycosylation can determine protein folding, stability, function, localization, trafficking and half-life (Schjoldager *et al.*, 2020). Furthermore, the diversity and complexity of glycosylation can be a hallmark of health or disease (Reily *et al.*, 2019), pregnancy (Jansen *et al.*, 2016) and lactation (Goonatilleke *et al.*, 2019).

Common methodologies for studying bovine IgG glycosylation include lectin microarray profiling (Yu *et al.*, 2020), the hydrolysis of glycans and quantification of monosaccharides (Feeney *et al.*, 2019), and the release and analysis of intact glycans by mass spectrometry (van Leeuwen *et al.*, 2012). Each of these strategies, however, has its limitations. Lectin microarray profiling can recognise and quantify glycan features (Hirabayashi *et al.*, 2013), such as glycan type (*i.e.*, complex, hybrid or high-mannose), glycan substructures or types of linkages in glycan substructures, without providing information on glycan composition, complete structure and localization in the protein. Quantification of the monosaccharides gives an average picture of the sample glycome (Rohrer *et al.*, 2013), without specific information on the individual glycoforms and their localization on the proteins. Analysis of the released glycans does provide information on the composition, heterogeneity and structure of the glycoforms (Rojas-Macias *et al.*, 2019), but does not indicate which protein or where on the protein they originate from, and what the occupancy of the glycosylation sites on the proteins is. Emerging technologies in the constantly developing field of glycoproteomics have made possible the analysis of glycopeptides by mass spectrometry, providing compositional and structural information on both the peptide and the glycan, site localization of the glycan, and occupancy of the glycation site (Hinneburg *et al.*, 2016, Reiding *et al.*, 2018, Ruhaak *et al.*, 2018).

Here, we designed a study to investigate how bovine milk IgG changes

longitudinally during lactation in relation to the rest of the mammary secretion proteome. While the decrease in abundance of IgG during the first days of lactation is well documented, we combined careful experimental design and state-of-the-art glycoproteomics technology for the in-depth investigation of bovine milk IgG *N*-glycosylation. Next to the dynamic changes during lactation, we studied the samples of several individual cows to assess how variable the findings are between animals. We observed a considerable decrease of IgG abundance during the first days following calving, and a further gradual decrease towards the end of the first month post-partum. Next to the better-documented IgG1 and IgG2, we report here, for the first time to our knowledge, the identification of subtype IgG3 in the bovine milk of all four cows and at every investigated time point. We found bovine milk IgG glycosylation to be highly diverse and complex, seemingly much more so than human IgG glycosylation, and we identified biological features of the bovine milk IgG *N*-glycosylation, such as NeuAc sialylation, that are a hallmark of bovine colostrum, greatly diminishing in mature bovine milk.

3

Results

To study the longitudinal changes in the bovine milk proteome, IgG relative abundances and *N*-glycosylation, we analysed a total of 39 individual bovine milk samples collected from 4 individual Dutch Friesian-Holstein cows referred to as Cow 1, Cow 2, Cow 3 and Cow 4. The time scale of the bovine milk sampling is schematically illustrated in Figure 3.1A.

Protein content and 1D gel electrophoresis of the bovine milk samples

The protein content of all collected samples was determined by the BCA method, and the samples were further analysed by SDS-PAGE to gain a first overview of changes occurring during lactation in the bovine milk proteome. Figure 3.1B illustrates the longitudinal changes in protein content between the four individual cows. Considerable decreases in protein content occurred during the first few days of lactation in all four cows, with the most notable change occurring between the first two sampling time points. Substantial differences in the protein content between the individual cows were only observed in the first two days of lactation, after which the protein content of all samples stabilised to comparable levels for the rest of the sampled time points. The highest protein content was observed in bovine colostrum in the first sampled time point, *i.e.*, 0.5 d, at (mean \pm standard deviation) 228.0 ± 3.8 mg protein/mL for Cow 1, followed by 178.5 ± 4.8 and 169.4 ± 7.4 mg/mL for Cow 2 and Cow 4, respectively, and 141.5 ± 9.9 mg/mL for Cow 3. For all four cows, the protein content of the mature bovine milk stabilised to 28.7-30.2 mg/mL at 28 d after calving.

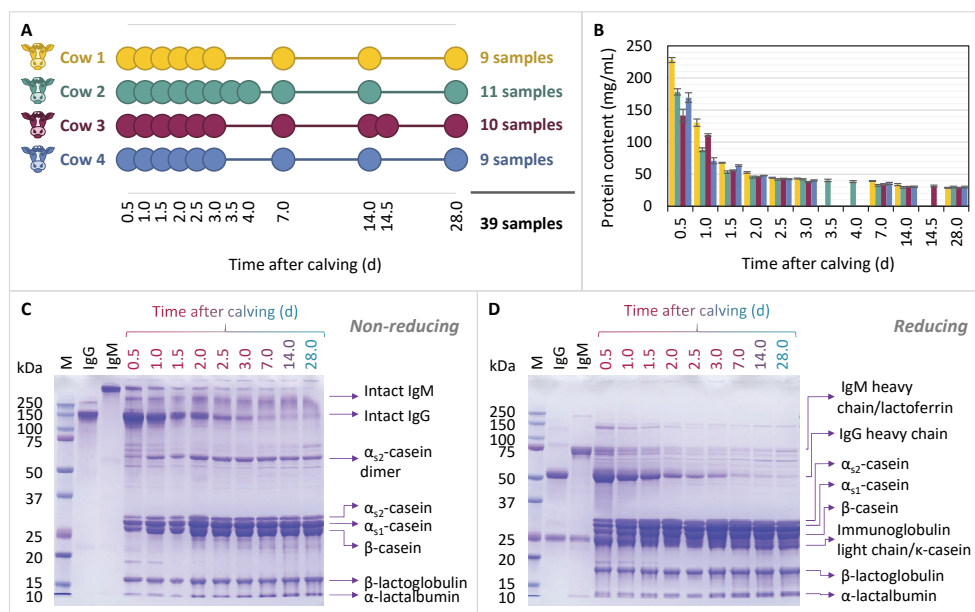


Figure 3.1 - Insights into the protein content and composition of the bovine milk samples collected during lactation from the four individual cows, i.e., Cow 1, Cow 2, Cow 3 and Cow 4. (A) Chart depicting the sampling time points for each individual cow following calving. (B) Total protein content of the bovine milk samples, determined using a bicinchoninic acid (BCA) assay. The error bars represent the standard deviation of triplicate measurements. The colour codes are consistent between panels (A) and (B), and are specific for each individual cow. (C) and (D) SDS-PAGE gel images of the samples collected from Cow 4 analysed under non-reducing (C) and reducing (D) conditions. All bovine milk samples were loaded on the gel at 20 μ g protein/well. The first 3 gel lanes display a protein marker panel: M = Precision Plus Protein™ Dual Color Standards, IgG = bovine serum IgG, IgM = bovine serum IgM, the latter two loaded on the gel at 4 μ g protein/well.

Next to the marked decrease in protein concentration during lactation, we also found considerable longitudinal changes in the protein composition of the samples. Bovine milk samples were separated on SDS-PAGE gels, from which a gradual shift in composition could be observed, as illustrated in Figure 3.1C and 1D for the samples collected from Cow 4. At 0.5 d, the first sampled time point, the proteome was dominated by immunoglobulins, IgG making up a considerable proportion thereof, but this changed to a casein-dominated proteome over the course of approximately a week. While Figure 3.1 shows the changes in protein composition for one cow (Cow 4), similar changes were seen in the bovine milk of the other three cows (supplementary Figure S3.1).

Bovine milk proteome composition

We next analysed each individual sample by bottom-up proteomics, performing for each sample triplicate LC-MS/MS methods, accumulating to a total of 117 runs. This

approach resulted in the identification of 212 individual bovine milk protein sequences across all samples analysed. The abundances of the identified proteins were determined using label-free intensity Based Absolute Quantification (iBAQ) value. Some of the most notable differences between the bovine colostrum proteome at 0.5 d and the mature bovine milk proteome at 28 d, averaged over all cows, are depicted in Figure 3.2A and 3.2B. The source data behind these graphs can be found in supplementary Table S3.1. The volcano plots and bar charts *per* individual cows are provided in supplementary Figures S3.2 and S3.3, respectively.

The major findings of the dynamic bovine milk proteome determined by proteomics

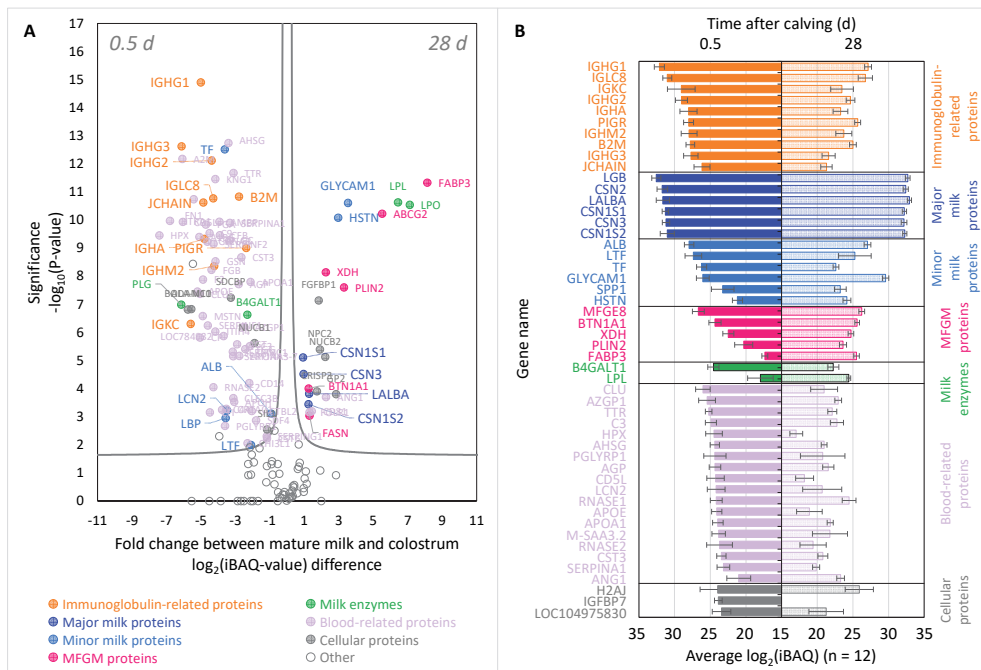


Figure 3.2 - Comparison between the proteomes of the samples collected at 0.5 and 28 d after calving averaged across all four cows, as determined by bottom-up mass spectrometry. (A) Volcano plot depicting the proteins with significantly different abundances in the proteomes of samples collected at 0.5 and 28 d after calving. Similar volcano plots are shown for each individual cow in supplementary Figure S2. (B) Differences in intensity Based Absolute Quantification abundance ($\log_2(\text{iBAQ})$) between the 50 most abundant protein sequences present in the proteomes of samples collected at 0.5 (left) and 28 (right) d after calving. For the selection of the 50 most abundant protein sequences, firstly the highest iBAQ-value was selected for each individual protein sequence between the iBAQ-values at 0.5 and 28 d. Next, the protein sequences with the 50 highest iBAQ-values were established as the 50 most abundant proteins of bovine colostrum and mature milk proteomes combined. The error bars indicate the standard deviation of each protein sequence iBAQ-value measured in 12 samples (i.e., triplicate measurements of samples from four cows). The colour codes highlighting the different protein sequence categories are consistent between panels (A) and (B). Similar bar charts are shown for each individual cow in supplementary Figure S3.2.

proved to be in line with the results from SDS-PAGE (Figure 3.1C and D, and Figure S3.4). All immunoglobulin-related proteins were found to be very abundant in the bovine colostrum, and significantly decreased in abundance during lactation. Next to the immunoglobulins, a majority of blood serum-related proteins were found to be significantly more abundant in the bovine colostrum than in the mature bovine milk. This included minor bovine milk proteins such as serotransferrin (TF), lactoferrin (LTF) and serum albumin (ALB), and endogenous bovine milk enzymes, such as plasmin(ogen) (PLG) and β -1,4-galactosyltransferase 1 (B4GALT1). Unlike the IgG1-rich bovine colostrum, the proteome of the mature bovine milk was dominated by the six major bovine milk proteins, *i.e.*, the four caseins, α -lactalbumin (LALBA) and β -lactoglobulin (BLG). Alongside the six major bovine milk proteins, minor proteins such as lactophorin (GLYCAM1) and histatherin (HSTN), the majority of milk fat globule membrane (MFGM) proteins, and milk enzymes lipoprotein lipase (LPL) and lactoperoxidase (LPO), all showed significant increases in abundance in mature bovine milk. For a complete picture of the proteome changes at each time point during lactation and for each individual cow, please consult the heatmap in the supplementary Figure S3.4 and the source data provided in supplementary Table S3.1.

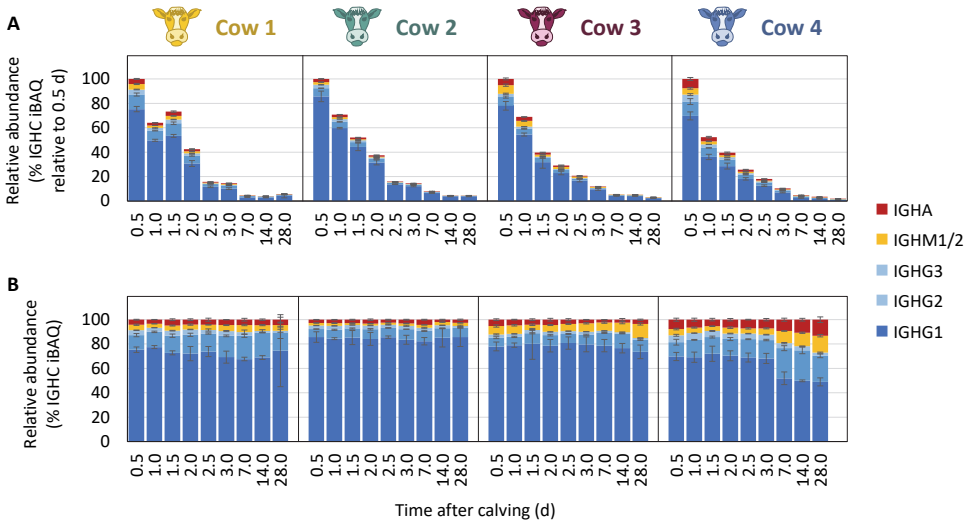


Figure 3.3 - Changes in the relative abundances (% iBAQ) of IGHC1 (dark blue), IGHC2 (medium blue), IGHC3 (light blue), IGHM1/2 (yellow) and IGHA (red) in the bovine milk proteome of Cow 1, Cow 2, Cow 3 and Cow 4, between 0.5 and 28 d after calving. (A) The decrease of immunoglobulin concentration during lactation is shown by expressing heavy chain constant abundances during lactation relative to those in the first sampled time point. (B) Depicted are the relative abundances of the heavy chain constant regions to each other, irrespective of their abundance in the bovine milk proteome. The error bars represent the standard deviations of triplicate measurements.

Having identified immunoglobulin classes IgG, IgM, IgA as key components of the bovine colostrum proteome, Figure 3.3A and B provide more detailed insight into their abundances during lactation. The abundance of the immunoglobulins relative to the rest of the bovine milk proteome was found to decrease up to ~50 times during lactation from 0.5 to 28 d after calving (Figure 3.3A). Despite the decrease in concentration of the immunoglobulins during lactation, the relative ratios of the different immunoglobulins (*i.e.*, IgG, IgM, IgA) to each other remained alike, showing higher similarity within each cow than between the four individual cows (Figure 3.3B). While longitudinal and individual differences could be observed in the relative ratios of the different immunoglobulins to each other, the bovine milk immunoglobulin repertoire shared many similarities between all samples: IgG was found to be the dominant isotype, followed by IgM and IgA. All three bovine IgG subtypes (*i.e.*, IgG1, IgG2 and IgG3) were detected in all of the analysed samples, with IgG1 always showing the highest abundance, followed by IgG2, and to a much smaller extent also by IgG3, a protein which has not been previously reported to be present in bovine milk. The data shown in Figure 3.3 also display that not all cows have a similar milk immunoglobulin profile, whereby in particular the Cow 4 appears to represent an outlier, with relatively high levels of IgG2, IgM and IgA in the mature bovine milk.

***N*-Glycosylation of IgG captured from bovine milk**

The data above show that the immunoglobulins, and particularly IgG, represented to a large extent the major differences between the proteome of bovine colostrum and mature milk. To further investigate whether not only the total concentration, but also the molecular characteristics, of IgG varied over lactation, we next affinity-captured IgG from all the bovine milk samples by making use of Protein G-based enrichment. Isolated IgG was subsequently analysed by bottom-up mass spectrometry techniques employing product ion-dependent hybrid fragmentation methods optimized for the analysis of glycopeptides (Dingess *et al.*, 2021, Reiding *et al.*, 2018). The analysis of captured IgG from 39 samples and one bovine serum pool IgG using triplicate LC-MS/MS measurements amounted to a total of 120 runs, resulting in the identification of a repertoire of 44 different *N*-glycoforms occurring on the CH2 domain of IgG, identified across all samples, time points and cows. The heatmap in Figure 3.4A depicts the combined information of glycan macro- and micro-heterogeneity, determined based on spectral counts for *N*-glycosylation site in the CH2 domain of IgG from each individual sample.

Very high similarity was observed between the corresponding samples of the individual cows. The IgG *N*-glycosylation site was found to be occupied to an average proportion of 98%, mostly with complex *N*-glycans, although a very small

proportion of non-glycosylated peptides occurred in all samples as well. The *N*-glycans were clustered in Figure 3.4A based on their biochemical features, with Figure 3.4B showing the changes thereof across lactation. Representative spectra of the non-glycosylated peptide, and glycopeptides exhibiting each selected biochemical feature, were manually inspected and are shown in supplementary Figures S3.5-S3.12. Of all IgG *N*-glycans in bovine colostrum at 0.5 d after calving, 48% (based on spectral count) were found to be sialylated, with the remaining 52% corresponding to neutral glycans. Preliminary screening of the raw files revealed strong oxonium ion evidence for the occurrence of sialylation with *N*-acetylneuraminic acid (NeuAc) and *N*-glycolylneuraminic acid (NeuGc), but not for the occurrence of acetylated sialic acid residues, or of deaminated neuraminic acid (Kdn) (Supplementary Figure S3.15). Interestingly, sialylation with *N*-acetylneuraminic acid (NeuAc), *N*-glycolylneuraminic acid (NeuGc), and also with the combination of NeuAc and NeuGc within the same glycan, were all detected in the bovine milk samples, particularly in the first days following parturition. NeuAc sialylation was found to be an exclusive characteristic of bovine colostrum, decreasing to nearly undetectable levels in mature bovine milk (Figure 3.4B, 3rd panel from the top). Comparably to the mature bovine milk IgG, no NeuAc sialylation was detected in the IgG sample acquired from pooled bovine serum. The level of NeuGc sialylation, on the other hand, remained fairly constant throughout lactation (Figure 3.4B, 4th panel from the top). As a potential consequence of the decrease in NeuAc sialylation, the proportion of neutral glycans increased during lactation, reaching an average level of 83% (based on spectral count) in mature bovine milk samples. Fucosylation was present among both neutral and sialylated glycans, and remained relatively constant across time (Figure 3.4B, 2nd panel from the top). Branching of the complex *N*-glycans occurred with LacNAc antennae. The average number of LacNAc antennae *per* glycan gradually decreased from 1.60 ± 0.14 in bovine colostrum to 1.16 ± 0.07 in the mature bovine milk (Figure 3.4B, 5th panel from the top). This appeared to be the result of the decrease in complexity and branching of the glycans during lactation. A small proportion of glycans with bisecting GlcNAc was also observed, whose abundance increased during lactation (Figure 3.4B, 6th panel from the top). The *N*-glycosylation of the pooled bovine serum IgG was found to most closely resemble that of the mature bovine milk, and exhibited less resemblance to that of bovine colostrum or transitional bovine milk.

Discussion

In this study, we monitored the changes in bovine milk proteome composition, with emphasis on IgG, following its content and biochemical properties during lactation. Changes occurred most dramatically within the first days after calving, followed by

NeuAc sialylation is a hallmark of bovine colostrum IgG glycosylation

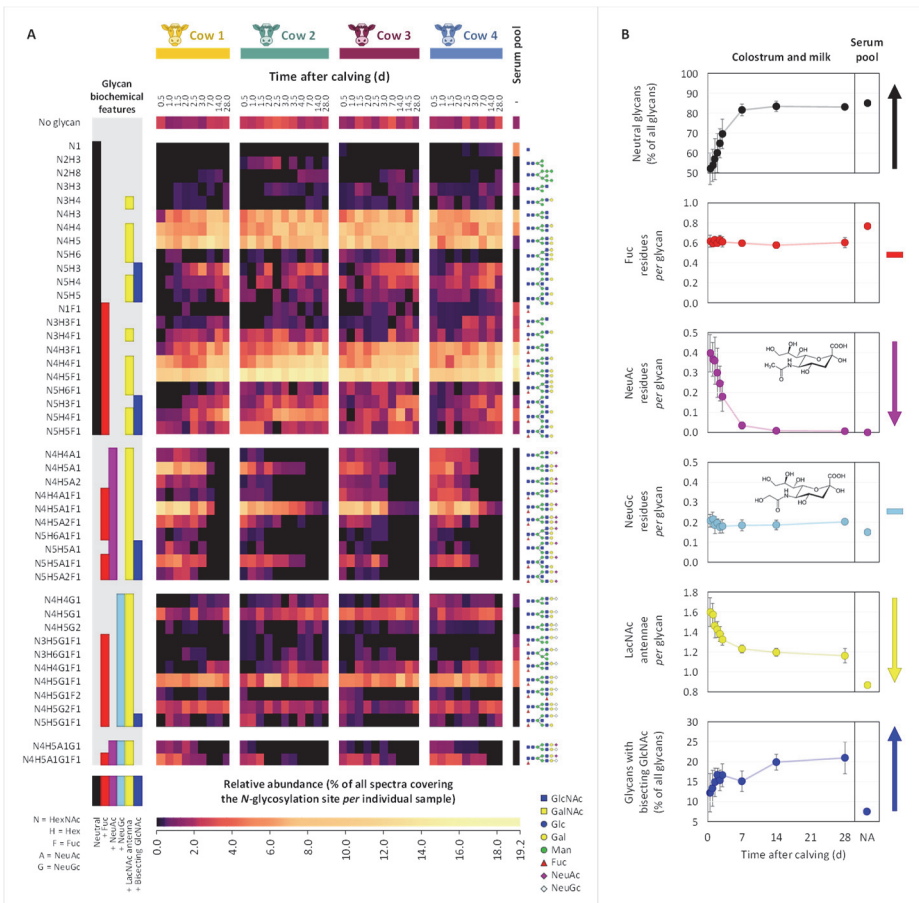


Figure 3.4 - Changes in the N-glycosylation profiles of bovine IgG in the interval 0.5-28 d after calving in the bovine milk of the four cows, i.e., Cow 1, Cow 2, Cow 3 and Cow 4. Additionally, in the last column, for comparison, the N-glycosylation profile of bovine serum IgG is depicted (pooled sample). (A) Heatmap depicting the macro- and micro-heterogeneity of the IgG CH2 domain N-glycosylation determined based on spectral counts. Normalisation was performed relative to all spectra covering the glycosylation site, whereby the sum of all glycoforms and the non-glycosylated site amounts to 100%, and the relative quantitation of the non-glycosylated site indirectly indicates site occupancy. Clustering was performed based on the biochemical features of the N-glycans: neutral (black), fucosylated (red), sialylated with NeuAc (magenta), sialylated with NeuGc (light blue), containing LacNAc antennae (yellow) or bisecting GlcNAc (blue), respectively. The glycan composition is indicated to the left of the heatmap. To the right of the heatmap are proposed glycan structures corresponding to each glycan composition. (B) Illustrative lactation dynamics of the N-glycan biochemical features averaged across the four cows and compared to their corresponding level in the bovine serum IgG, with the panels from top to bottom describing the changes in neutral glycans, fucosylation, sialylation with NeuAc or NeuGc, containing LacNAc antennae or bisecting GlcNAc, respectively. The error bars represent the standard deviation of values from the four cows. The colour codes of the biochemical features are consistent between panels (A) and (B). Abbreviations: Gal = galactose; Glc = glucose; Man = mannose; Hex = hexose; GalNAc = N-acetylgalactosamine; GlcNAc = N-acetylglucosamine; HexNAc = N-acetylhexosamine; Fuc = fucose; NeuAc = N-acetylneuraminic acid; NeuGc = N-glycolylneuraminic acid.

a more gradual transition from bovine colostrum to mature bovine milk.

Changes in the bovine milk proteome during lactation converge around the immunoglobulins

The six main bovine milk proteins, *i.e.*, the four caseins, α -lactalbumin and β -lactoglobulin (Gazi *et al.*, 2022), were found to be abundant at all analysed time points (Figure 3.2, Figure S3.4). Next to these well-known constituents, and in line with the literature (McGrath *et al.*, 2016), we found several immunoglobulins and related proteins, making up a great proportion of in particular the bovine colostrum proteome. Amongst the immunoglobulin-related proteins, we highlighted in our results (Figure 3.2, Figure S3.4) the heavy and light chains of immunoglobulins IgG, IgA and IgM, the immunoglobulin joining chain (JCHAIN), the polymeric immunoglobulin receptor (PIGR), and β_2 -microglobulin (B2M). We found all these proteins to be very abundant in bovine colostrum and significantly decreased during lactation. The JCHAIN plays a role in the oligomerisation of IgM and IgA, leading to the formation of IgM pentamers and IgA dimers, trimers and tetramers (Dingess *et al.*, 2022, Johansen *et al.*, 2000). PIGR is the precursor of the secretory component, the protein whose role it is to transport polymeric immunoglobulins across epithelial barriers (Mostov, 1994), and in this case, into the bovine milk. B2M, initially isolated from bovine milk as the minor whey protein lactollin (Groves and Greenberg, 1977, Groves *et al.*, 1963), is the light chain component of the neonatal Fc receptor (FcRn) heterodimer that binds, transports and protects IgG and serum albumin from degradation (Baumrucker *et al.*, 2022). The purpose of its presence in bovine milk, particularly at higher concentration in the bovine colostrum, is as of yet uncertain. It can be speculated that B2M is shed from the FcRn heterodimer into bovine milk following transcytosis of IgG *via* the mammary secretory epithelial cells. It is unlikely that the B2M from bovine milk plays a role in the further transcytosis of bovine milk IgG from the calf's intestine into its blood circulation, considering that endogenous B2M is necessary for the correct folding and surface expression of the FcRn heterodimer in the intestinal epithelial cells (Praetor and Hunziker, 2002). Next to the FcRn/B2M complex prolonging the half-life of circulatory IgG, it has also been shown that B2M alone also exhibits a protective effect on IgG (Kim *et al.*, 2008). The role of the B2M in bovine colostrum and bovine milk may therefore be to protect the passive immunity the calf receives from the cow, until such time that the calf's own immune system is sufficiently developed.

The bovine milk immunoglobulome was dominated by IgG, which we found to make up 51-73% of the total protein content of bovine colostrum in the first bovine

milking (Figure S3.13B), in line with previous reports (Marnila and Korhonen, 2011). Conversely to bovine milk, the immunoglobulin repertoire of human milk is dominated by secretory IgA (sIgA), with IgG constituting only a minor component human milk (Atyeo and Alter, 2021, Dingess *et al.*, 2022). The placental barrier is species-specific and determines whether immune transfer to the foetus can occur *in utero* through the placenta (Chucrí *et al.*, 2010). In humans, prenatal transfer of IgG from the mother *via* the placenta ensures that neonates have systemic immunity at birth (Renegar, 2005). Due to the high number of placental barrier layers in ruminants, including bovine, prenatal transfer of IgG is not possible (Chucrí *et al.*, 2010). Consequently, calves are born agammaglobulinemic, and rely on the IgG from bovine colostrum for survival (Godden *et al.*, 2019). The IgG from bovine colostrum is therefore meant to be taken up into circulation and provide the calf with passive systemic immunity, explaining the dominant position of IgG at the expense of sIgA in the bovine milk immunoglobulome that we described in our results (Figure 3.3). Unlike bovines, humans do not need to transfer passive systemic immunity to the neonate through milk. This leads to the human milk immunoglobulome to be dominated by sIgA, an immunoglobulin active in the neonate's intestines, providing passive mucosal immunity (Renegar, 2005).

The transfer of milk immunoglobulins from the offspring's gut into its circulation ceases at the time of gut closure. This takes place 2-3 d after birth, after the calf has consumed the IgG-rich bovine colostrum essential for survival in its first days of life (Weaver *et al.*, 2000). In the case of humans, gut closure occurs before birth (Weström *et al.*, 2020). Consequently, neonates lack the ability to absorb milk IgG into circulation, thereby also explaining the low IgG levels present in human milk (Atyeo and Alter, 2021).

While bovine milk IgG and human milk sIgA are designed for different purposes, as described above, the consumption of bovine milk IgG by infants may still provide health benefits comparable to the human milk sIgA (Ulfman *et al.*, 2018). Considering that bovine milk IgG can also not be absorbed into circulation due to the pre-natal infant gut closure (Weström *et al.*, 2020), its function in human infant nutrition may instead be that of passive mucosal immunity in the intestine, as in the case of the human milk sIgA.

The secretory component in sIgA (*i.e.*, derived from the PIGR precursor shown in Figure 3.2, Figure S3.4), next to ensuring transport of the IgA across the epithelial barrier into milk, also protects the IgA from proteolytic degradation in the gastrointestinal tract (Lindh, 1975). While IgG lacks the secretory component, bovine IgG is less susceptible to proteolytic degradation by digestive proteases than

its human counterpart (Burton *et al.*, 2020). Furthermore, considering the infant's immature and developing digestive system (Bourlieu *et al.*, 2014), recovery of bovine IgG from infant faeces has been shown to be high (Jasion and Burnett, 2015), confirming its high resistance to proteolytic degradation.

Changes in bovine IgG *N*-glycosylation during lactation converge around loss of NeuAc sialylation

The IgG *N*-glycosylation repertoire detected here reveals that bovine milk IgG glycosylation is more diverse and heterogeneous than that of its human milk IgG counterpart, although both *N*-glycosylation sites are located in conserved regions of their respective CH2 domains. While we revealed here a diverse repertoire of 44 glycoforms co-occurring on bovine milk IgG between 0.5 and 28 d post-partum (Figure 3.4), Zhu *et al.* (2020), using comparable LC-MS/MS approaches identified only 6 glycoforms on human milk IgG analysed between 1 and 16 weeks post-partum. These differences may in part result from differences in sample preparation, *i.e.*, we here analysed bovine milk IgG glycosylation in affinity-captured IgG, whereas Zhu *et al.* (2020) analysed glycopeptides enriched from a whole human milk proteolytic digest. Other studies, such as the one of Trbojević Akmačić *et al.* (2015), have shown that the human IgG glycosylation repertoire can also be quite diverse. However, conceptually, the number of bovine IgG glycoforms can be higher due to the presence of NeuGc and its combinations with NeuAc. All 6 human milk IgG glycoforms identified by Zhu *et al.* (2020) contained core fucosylation, whereas we determined here the bovine milk IgG to contain an average of 0.6 fucose residues *per* glycan at all investigated lactation time points (Figure 3.4B, second panel from the top). Zhu *et al.* (2020) found a single sialylated glycoform, *i.e.*, N4H5A1F1, making up 21% of the abundance of all glycoforms at 1 wk, and gradually decreasing to 12% at 16 wk. In terms of similarities shared with the bovine IgG glycosylation, all 6 human milk *N*-glycoforms are amongst the most abundant ones in the repertoire of 44 bovine glycoforms (Figure 3.4). The physiological reason for the differences between the inter-species diversity of glycoforms is still unknown. However, we speculate they may relate to the different functionality of the bovine milk IgG between the two species. With IgG only constituting a minor component of the human milk immunoglobulome (Ateyo and Alter, 2021), and with infant gut closure having occurred prior to birth (Weström *et al.*, 2020), it is unlikely that IgG is actively transported into human milk. Conversely, bovine milk IgG is actively transported into the bovine colostrum (Delouis, 1978), it needs to be resistant to the digestive proteases in the calf's gastrointestinal tract, it needs to be actively transported from the calf's intestine into its blood circulation (Godden *et al.*, 2019), and it needs to have a sufficiently long half-life in the calf's

circulation to provide passive immunity while the calf can develop its own system immunity. Therefore, it can be speculated that the high diversity and complexity of the bovine milk IgG glycoforms (Figure 3.4A), and the differences in biochemical features of the glycans in bovine colostrum from mature bovine milk (Figure 3.4B), are correlated with the binding affinity of the FcRn receptors that need to transport and protect the IgG from degradation, and also determine the proteolytic susceptibility of the IgG.

A very interesting feature of the bovine milk IgG glycosylation was the fact that NeuAc and NeuGc sialylation occurred simultaneously in the samples, at times even both on the same glycoprotein (Figure 3.4, Figure S3.12). We found here NeuAc sialylation to be a unique hallmark of bovine colostrum IgG, with the NeuAc/NeuGc ratio decreasing from 1.91 in the 0.5-d colostrum to 0.03 in the 28.0-d mature milk. This is interesting, particularly from a human nutrition perspective. While being a widespread and abundant sialic acid in the glycoconjugates of most mammalian species, NeuGc cannot be produced by the human body and is therefore a non-human sialic acid (Irie *et al.*, 1998). Consequently, NeuGc is immunogenic to humans, with anti-NeuGc antibodies detectable even in healthy individuals (Zhu and Hurst, 2002). The high NeuAc sialylation of the particularly abundant IgG in bovine colostrum renders its glycosylation more human-like, making bovine colostrum IgG furthermore interesting from the perspective of biological functionality in human nutrition. The function of high sialylation of bovine colostrum IgG is hypothesised to be the extension of the IgG half-life; non-sialylated proteins are removed from circulation by binding to asialoglycoprotein receptors (ASGRs) in the liver and are subsequently directed to degradation (Stockert, 1995). The ASGRs have a high affinity for glycan-terminal galactoses in the absence of sialic acids (D'souza and Devarajan, 2015). While ASGRs are primarily present in the liver, Stockert (1995) reported their presence to a lower extent also in the gastro-intestinal tract of rats. Species-specific differences may exist in the distribution and binding affinity of the ASGRs. However, by similarity it can be speculated that the higher sialylation of bovine colostrum IgG is meant to extend its half-life by protecting it from ASGR recognition in the gastro-intestinal tract and circulation of the calf.

Our research lays the foundation for future investigations into the biological functionality of the here described heterogeneity, diversity and dynamic lactational changes in bovine milk IgG *N*-glycosylation. These findings are not only important from the perspective of the healthy development of the calf's own immunity in the context of bovine animal health management, but also from the perspective of understanding the bioactivity of bovine milk components on human health, and

particularly the potential role of bovine milk IgG in *e.g.*, providing passive mucosal immunity to formula-fed infants.

Materials and Methods

Chemicals and reagents

Bovine serum IgG (I5506) and IgM (I8135), enriched from pooled bovine serum, dithiothreitol (DTT), Tris(2-carboxyethyl)phosphine (TCEP), Tris, chloroacetamide (CAA), sodium deoxycholate (SDC) and formic acid (FA) were purchased from Sigma-Aldrich (Darmstadt, Germany). The Pierce BCA Protein Assay Kit (23225) for the bicinchoninic acid (BCA) assay, Imperial Protein Stain (24615), Pierce Spin Columns (69705), Dulbecco's phosphate-buffered saline (DPBS), Pierce IgG elution buffer (1856202) and neutralization buffer (1856281) were sourced from Thermo Scientific, Rockford, Illinois, USA. Trifluoroacetic acid (TFA) was purchased from Fisher Scientific (Landsmeer, The Netherlands). PepMap Trap Cartridges (5 μ m C18 300 μ m X 5 mm) were sourced from Thermo Fisher Scientific (Germering, Germany). XT sample buffer, 18-well 12% Criterion™ XT Bis-Tris Precast Gels and XT MOPS running buffer and Precision Plus Protein™ Dual Color Standards used for sodium-dodecyl sulphate – polyacrylamide gel electrophoresis (SDS-PAGE) were all purchased from Bio-Rad (Veenendaal, The Netherlands). Ultrapure MilliQ water was prepared with the system sourced from Merck Millipore, Darmstadt, Germany. Sequencing-grade endoproteinase GluC and broad range protease inhibitor cocktail (cOmplete, EDTA-free) were sourced from Roche Diagnostics (Mannheim, Germany).

Sequencing-grade trypsin was sourced from Promega (Madison, Wisconsin, USA). Sequencing-grade lysyl endopeptidase (LysC) was sourced from Wako Chemicals (Richmond, Virginia, USA). Oasis PRiME HLB 96-well plates (10 mg sorbent *per* well) for solid phase extraction (SPE) were sourced from Waters (Etten-Leur, The Netherlands). C18 column material (Poroshell 120 EC-C18, 2.4 μ m) was sourced from Agilent Technologies (Amstelveen, The Netherlands). Protein G slurry (Protein G Sepharose 4 Fast Flow, GE17-0618-02) was sourced from GE Healthcare (Chicago, Illinois, USA).

Bovine milk samples

Bovine milk samples from four individual Dutch Friesian-Holstein cows, *i.e.*, referred to as Cow 1, Cow 2, Cow 3 and Cow 4, were obtained from a farm in The Netherlands. Sampling started within 12 h after calving and continued up until four weeks after calving, with samples taken at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28.0 d post-partum, as indicated in Figure 3.1A. Additional samples were taken from Cow 2 at 3.5 and 4.0 d, and Cow 3 at 14.5 d post-partum. Calving dates of all

four cows were in the interval of January 24th-28th, 2022. Milking was carried out twice daily at fixed times. The collected samples were stored at -20 °C until further analyses. The samples collected up until 3 d will be further referred to as bovine colostrum, samples collected at 28 d as mature bovine milk, and samples collected in between 3 and 28 d as transitional bovine milk.

Protein content and gel electrophoresis of bovine milk samples

The protein content of the bovine milk samples was determined by the BCA assay using the Pierce BCA Protein Assay Kit and 96-well plate method, according to manufacturer's instructions. SDS-PAGE under both reducing (with DTT) and non-reducing (without DTT) conditions was performed as previously described (Gazi *et al.*, 2022). The bovine milk samples, and the bovine serum IgG and IgM, were treated with XT sample buffer, loaded onto 18-well 12% Criterion™ XT Bis-Tris Precast Gels, and electrophoresis was run in XT MOPS running buffer. The bovine milk and immunoglobulin samples were loaded at 20 and 4 µg protein/well, respectively. Precision Plus Protein™ Dual Color Standards were run on the gels in parallel with the samples for protein size reference. The gels were stained with Imperial Protein Stain according to manufacturer's instructions, and destained in ultrapure MilliQ water.

Capturing IgG from the bovine milk samples

For capturing IgG from bovine milk, the milk samples were warmed for 5 min at 37 °C. Broad range protease inhibitor cocktail was added at a ratio of 1 tablet to 23 mL of bovine milk. Sodium azide (NaN₃) was added to a final concentration of 0.02% to the bovine milk samples. Skimming of the bovine milk samples was performed by centrifugation at 1,500 × g and 37 °C for 10 min. Volumes of 100 µL of diluted Protein G slurry were packed into Pierce Spin Columns. The columns were conditioned with DPBS. Volumes of 10, 20, 50, 150, 250, 350, 400 and 450 µL of skim bovine milk were loaded from lactation time points 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5 and 4.0 d, respectively. DPBS was added to a total skim bovine milk + DPBS volume of 700 µL. The bovine milk samples from 7-28 d after calving were loaded directly at the volume of 700 µL. The bovine milk/Protein G slurries were gently mixed on a rotary shaker for 1 h at room temperature. Following capturing of the IgG, the beads were washed two times with 150 µL of DPBS, and one time with ultrapure MilliQ water. The captured IgG was then eluted with Pierce IgG elution buffer and treated with neutralization buffer. The samples were stored at 4 °C until further analysis. The protein content of the captured IgG samples was determined spectrophotometrically using a NanoDrop 2000 instrument (Thermo Fisher Scientific, Waltham, Massachusetts, USA) at 280 nm with the pre-set method for

IgG quantification. The capturing efficiency was corrected based on the capturing from the bovine serum IgG reference, and the IgG concentrations were calculated back for every bovine milk sample. To verify the composition of the captured sample material, the samples were analysed by non-reducing SDS-PAGE as described above. Bovine serum IgG was used as reference, and all samples were loaded at 2 µg protein/well.

Analysis of bovine milk proteome and captured IgG CH2 N-glycosylation by mass spectrometry

Amounts of 10 µg of protein from each whole bovine milk sample, captured IgG from bovine milk, and bovine serum IgG were denatured, reduced and alkylated in a buffer containing 10 mM TCEP, 100 mM Tris, 40 mM CAA and 1% (m/v) SDC, as previously described by Gazi *et al.* (2022). The whole bovine milk protein samples were digested with a ratio of 1:50 (m/m) GluC:protein for 4 h at room temperature, followed by the addition of 1:50 (m/m) trypsin:protein and overnight digestion at 37 °C. Digestion of the IgG samples was performed overnight at 37 °C with a mixture of 1:70 (m/m) LysC:protein and 1:50 (m/m) trypsin:protein. The enzymatic reactions were stopped and SDC was precipitated by adjusting the sample pH to the range of 1.5-2.0 with 10% TFA. The peptides were extracted by SPE using Oasis PRiME HLB 96-well plates according to the instructions of the manufacturer. Following extraction, the samples were dried using a vacuum concentrator and stored at -20 °C until further analysis.

The dried samples were reconstituted in 2% FA, and amounts of 800 ng of digested whole bovine milk protein or 100 ng digested IgG were injected *per* run. The samples were analysed using an Ultimate 3000 UHPLC system (Thermo Fisher Scientific, Germering, Germany) coupled online to an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA). Reversed-phase separation was achieved using PepMap Trap Cartridges and C18 analytical columns (50 cm length, 50 µm inner diameter). Mobile-phase solvent A consisted of 0.1% FA in water, and mobile-phase solvent B consisted of 0.1% FA in 80% ACN. Total method runtime for the whole bovine milk protein digests was 120 min at a flow rate of 300 nL/min, where the peptides were loaded at 9% B, followed by 1 min ramp from 9 to 13% B, the elution of the peptides during a 100 min linear gradient of 13-44% B, 3 min ramp from 44 to 99% B, 4 min washing of the column with 99% B, 1 min decrease from 99 to 9% B and 10 min equilibration to 9% B. The IgG digests were analysed by a similar method, with the following modifications: 60 min total method run time and 40 min gradient elution time. MS scans were recorded at a resolution of 120,000, with an automatic gain control (AGC) target of 400,000, 50

ms maximum injection time (IT) and a scan range of m/z 350-2,000. Data-dependent MS2 scans were recorded at a resolution of 60,000, AGC target of 50,000, 50 ms maximum IT and a scan range of m/z 120-4,000.

Each sample was injected three times with identical chromatography and MS method, but with different MS/MS methods; this provided triplicate information at precursor level, and complementary results at fragmentation level. The different fragmentation methods consisted of a) higher-energy collisional dissociation (HCD) at a normalised collision energy of 29; b) oxonium ion-triggered stepped HCD at normalised collision energy steps of 10, 25 and 40%; c) oxonium ion-triggered electron-transfer/higher-energy collisional dissociation (ET_HCD). The oxonium ion-triggered hybrid fragmentation methods favoured the fragmentation of specifically glycopeptides. The list of oxonium ions used for triggering hybrid fragmentation is provided in supplementary Table S3.2.

The mass spectrometry raw data and complete Byonic search results of the glycoproteomics analyses on the captured IgG have been deposited to the ProteomeXchange Consortium (Deutsch *et al.*, 2020) via the PRIDE (Perez-Riverol *et al.*, 2022) partner repository with the dataset identifier PXD037755.

Protein database optimisation, database search and identification of bovine milk proteome

For optimal results, and overcoming the limitations of incomplete, poorly annotated and highly redundant bovine protein databases, a customised protein database was created for the searches. The one protein sequence *per* gene version of the *Bos taurus* (taxon ID 9913) reference proteome (ID UP000009136) of January 8th, 2022 was downloaded from UniProt (<https://www.uniprot.org/proteomes/UP000009136>) in FASTA format on March 3rd, 2022. This database was further processed by removing residual duplicate sequences (*i.e.*, protein sequences with the same gene name), adding complete sequence where only fragment sequences were present, and updating the FASTA headers to the current headers of the date of download, *i.e.*, March 3rd, 2022. Considering that the bovine immunoglobulins in UniProt were listed under the names “uncharacterised protein” or “Ig domain-containing protein”, functional bovine immunoglobulin heavy and light chain, constant and variable sequences were downloaded from the reference directory of the international ImMunoGeneTics information system (IMGT RefSeq, <https://www.imgt.org/vquest/refseqh.html>). The constant immunoglobulin heavy chain sequences were assembled from the respective fragment sequences of the constant heavy (CH) domains and hinge (H) regions. A single representative allele

for each immunoglobulin gene sequence was kept in the working protein database. The contaminants FASTA file from the installation folder of the MaxQuant software was processed to remove all bovine contaminants. Database searches were performed on MaxQuant v 1.5.3.30 against the three databases described above combined, using default settings unless otherwise specified. The built-in contaminants feature of MaxQuant was deactivated. Digestion mode was set to specific against Trypsin/P and GluC, allowing a maximum of 2 missed cleavages. Methionine oxidation, protein N-terminus acetylation, and serine and threonine phosphorylation were searched as variable modifications. Cysteine carbamidomethylation was set as a fixed modification. Minimum peptide length was allowed at 5 amino acids, and maximum peptide mass was limited at 10,000 Da. Protein quantification was performed on a minimum of 2 unique + razor peptides. Label-free quantification was carried out using the intensity based absolute quantification (iBAQ) values. FTMS recalibration was activated. Following the database search, the protein groups table was further post-processed to remove non-bovine contaminants, reverse identifications, variable immunoglobulin sequences, proteins identified by less than 2 unique peptides, and proteins identified with an Andromeda score below 20. Supplementary Figure S3.14 illustrates the shortcomings of commonly-used SwissProt and UniProt bovine protein databases, and the advantages of using our optimised bovine protein database.

Protein and glycan database optimisation, database search and identification of IgG CH2 domain *N*-glycopeptides

The database search was performed with Byonic v4.5.2 (Protein Metrics, Cupertino, California, USA), a search engine specialised in the identification of glycopeptides. Considering the high apparent purity of the captured IgG samples based on the SDS-PAGE results (supplementary Figure S3.13), the samples were searched against a simplified database containing only the immunoglobulin sequences from IMGT RefSeq as described above. Decoys were added to the protein database. The *N*-glycan database used was based on the Byonic built-in database of 309 mammalian *N*-glycans. The default database was expanded to include complex and hybrid glycan antennae of *N*, *N'*-diacetyllactosamine (LacdiNAc) and all combinations of LacdiNAc and *N*-acetyllactosamine (LacNAc). To confirm which types of sialic acid to include into the glycan database, we verified MS2 chromatographic traces and spectra of relevant 0.5-d colostrum and 28.0-d mature milk samples from raw files resulting from methods employing higher-energy collisional dissociation. Supplementary Figure S3.15 shows oxonium ion evidence for various sialic acids, including unmodified and acetylated *N*-acetylneuraminic acid (NeuAc), unmodified

and acetylated *N*-glycolylneuraminic acid (NeuGc), and unmodified deaminated neuraminic acid (Kdn). The occurrence of NeuAc and NeuGc was undeniably confirmed, and they were found to overlap with oxonium ions derived from HexNAc residues. Acetylated residues and Kdn, however, could not be detected above noise levels. Based on this information, the database was expanded to include sialylation with both *N*-acetylneuraminic acid (NeuAc) and the non-human *N*-glycolylneuraminic acid (NeuGc), as well as all combinations thereof. The resulting database containing a total of 2440 *N*-glycan compositions that were used for identifying the IgG *N*-glycopeptides, is shown in supplementary Table S3.3. Default search parameters were used, unless otherwise specified. Trypsin cleavage sites were defined C-terminal of arginine and lysine residues with 2 missed cleavages, but digestion specificity was set to non-specific. Fragmentation type was set either to HCD for the samples analysed with HCD or product ion-triggered stepping HCD, or to both HCD & EThcD for the samples analysed with product ion-triggered EThcD. Cysteine carbamidomethylation was set as a fixed modification. Methionine and tryptophan oxidation, and N-terminal cyclisation of glutamine and glutamic acid to pyroglutamic acid were all searched as rare variable modifications, whereas *N*-glycosylation was searched as a common variable modification. A maximum of one rare and one common variable modifications were allowed *per* peptide. Protein false discovery rate (FDR) was set to 1% or 20 reverse counts. Further post-processing included filtering of the data based on $|\text{Log Prob}| \geq 1$ and $\text{score} \geq 150$. Based on the identified glycan composition, proposed glycan structures were built using GlycoWorkBench 2.1 build 146, according to the symbol nomenclature for glycan representation of the Consortium for Functional Glycomics (Varki *et al.*, 2009).

Author contributions

I.G., T.H. and A.J.R.H. conceptualised the project; I.G. performed all experiments; I.G. performed data curation and analysis with support from K.R.R.; I.G. was responsible for visualisation of the results and writing of the manuscript; all authors contributed to the interpretation of the findings and provided feedback for editing; A.J.R.H. and A.G. secured the funding for the project. All authors have read and agreed to the final version of the manuscript.

Acknowledgements

We acknowledge support from the Dutch Research Council (NWO) in the framework of the Innovation Fund for Chemistry (NWO SATIN project 731.017.202). K.R.R. acknowledges support from NWO Veni project VI.Veni.192.058.

References

- Atyeo C, Alter G. 2021. The multifaceted roles of breast milk antibodies. *Cell*, 184:1486-1499.
- Baumrucker CR, Macrina AL, Bruckmaier RM. 2022. Colostrogenesis: Role and Mechanism of the Bovine Fc Receptor of the Neonate (FcRn). *J Mammary Gland Biol Neoplasia*:1-35.
- Bourlieu C, Ménard O, Bouzerzour K, Mandalari G, Macierzanka A, Mackie AR, Dupont D. 2014. Specificity of infant digestive conditions: some clues for developing relevant in vitro models. *Crit Rev Food Sci Nutr*, 54:1427-1457.
- Burton RE, Kim S, Patel R, Hartman DS, Tracey DE, Fox BS. 2020. Structural features of bovine colostrum immunoglobulin that confer proteolytic stability in a simulated intestinal fluid. *J Biol Chem*, 295:12317-12327.
- Chucru TM, Monteiro J, Lima A, Salvadori M, Junior JK, Miglino MA. 2010. A review of immune transfer by the placenta. *J Reprod Immunol*, 87:14-20.
- D'souza AA, Devarajan PV. 2015. Asialoglycoprotein receptor mediated hepatocyte targeting—Strategies and applications. *J Control Release*, 203:126-139.
- Delouis C. 1978. Physiology of colostrum production. *Ann Rech Vet*. p. 193-203.
- Deutsch EW, Bandeira N, Sharma V, Perez-Riverol Y, Carver JJ, Kundu DJ, García-Seisdedos D, Jarnuczak AF, Hewapathirana S, Pullman BS. 2020. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res*, 48:D1145-D1152.
- Dingess KA, Gazi I, van den Toorn HW, Mank M, Stahl B, Reiding KR, Heck AJ. 2021. Monitoring Human Milk β -Casein Phosphorylation and O-Glycosylation Over Lactation Reveals Distinct Differences between the Proteome and Endogenous Peptidome. *Int J Mol Sci*, 22:8140.
- Dingess KA, Hoek M, van Rijswijk DM, Tamara S, den Boer MA, Damen MJ, Barendregt A, Romijn M, Juncker HG, van Keulen BJ. 2022. Humans have distinct repertoires of IgA1. *bioRxiv*.
- Feeney S, Gerlach JQ, Slattery H, Kilcoyne M, Hickey RM, Joshi L. 2019. Lectin microarray profiling and monosaccharide analysis of bovine milk immunoglobulin G oligosaccharides during the first 10 days of lactation. *Food Sci Nutr*, 7:1564-1572.

Gazi I, Franc V, Tamara S, van Gool MP, Huppertz T, Heck AJ. 2022. Identifying glycation hot-spots in bovine milk proteins during production and storage of skim milk powder. *Int Dairy J*, 129:105340.

Godden SM, Lombard JE, Woolums AR. 2019. Colostrum management for dairy calves. *Vet Clin North Am Food Anim Pract*, 35:535-556.

Goonatilleke E, Huang J, Xu G, Wu L, Smilowitz JT, German JB, Lebrilla CB. 2019. Human milk proteins and their glycosylation exhibit quantitative dynamic variations during lactation. *J Nutr*, 149:1317-1325.

Groves M, Greenberg R. 1977. Bovine homologue of β 2-microglobulin isolated from milk. *Biochem Biophys Res Commun*, 77:320-327.

Groves ML, Basch JJ, Gordon WG. 1963. Isolation, characterization, and amino acid composition of a new crystalline protein, lactollin, from milk. *Biochemistry*, 2:814-817.

Hinneburg H, Stavenhagen K, Schweiger-Hufnagel U, Pengelley S, Jabs W, Seeberger PH, Silva DV, Wührer M, Kolarich D. 2016. The art of destruction: optimizing collision energies in quadrupole-time of flight (Q-TOF) instruments for glycopeptide-based glycoproteomics. *J Am Soc Mass Spectrom*, 27:507-519.

Hirabayashi J, Yamada M, Kuno A, Tateno H. 2013. Lectin microarrays: concept, principle and applications. *Chem Soc Rev*, 42:4443-4458.

Irie A, Koyama S, Kozutsumi Y, Kawasaki T, Suzuki A. 1998. The molecular basis for the absence of N-glycolylneuraminic acid in humans. *J Biol Chem*, 273:15866-15871.

Jansen BC, Bondt A, Reiding KR, Lonardi E, De Jong CJ, Falck D, Kammeijer GS, Dolhain RJ, Rombouts Y, Wührer M. 2016. Pregnancy-associated serum N-glycome changes studied by high-throughput MALDI-TOF-MS. *Sci Rep*, 6:1-10.

Jasion VS, Burnett BP. 2015. Survival and digestibility of orally-administered immunoglobulin preparations containing IgG through the gastrointestinal tract in humans. *Nutr J*, 14:1-8.

Johansen F, Braathen R, Brandtzaeg P. 2000. Role of J chain in secretory immunoglobulin formation. *Scand J Immunol*, 52:240-248.

Kim J, Bronson C, Wani MA, Oberyshyn TM, Mohanty S, Chaudhury C, Hayton WL, Robinson JM, Anderson CL. 2008. β 2-Microglobulin deficient mice catabolize IgG more rapidly than FcRn- α -chain deficient mice. *Exp Biol Med*, 233:603-609.

Lindh E. 1975. Increased resistance of immunoglobulin A dimers to proteolytic degradation after binding of secretory component. *J Immunol*, 114:284-286.

Marnila P, Korhonen H. 2011. Milk Proteins | Immunoglobulins. In: Fuquay JW editor. *Encyclopedia of Dairy Sciences*: Academic Press. p. 807-815.

McGrath BA, Fox PF, McSweeney PL, Kelly AL. 2016. Composition and properties of bovine colostrum: a review. *Dairy Sci Technol*, 96:133-158.

Mostov KE. 1994. Transepithelial transport of immunoglobulins. *Annu Rev Immunol*, 12:63-84.

Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M. 2022. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res*, 50:D543-D552.

Praetor A, Hunziker W. 2002. β 2-microglobulin is important for cell surface expression and pH-dependent IgG binding of human FcRn. *J Cell Sci*, 115:2389-2397.

Raju TS, Briggs JB, Borge SM, Jones AJ. 2000. Species-specific variation in glycosylation of IgG: evidence for the species-specific sialylation and branch-specific galactosylation and importance for engineering recombinant glycoprotein therapeutics. *Glycobiology*, 10:477-486.

Reiding KR, Bondt A, Franc V, Heck AJ. 2018. The benefits of hybrid fragmentation methods for glycoproteomics. *TrAC, Trends Anal Chem*, 108:260-268.

Reily C, Stewart TJ, Renfrow MB, Novak J. 2019. Glycosylation in health and disease. *Nat Rev Nephrol*, 15:346-366.

Renegar KB. 2005. Passive immunization: systemic and mucosal. *Mucosal Immunol*:841.

Rohrer J, Basumallick L, Hurum D. 2013. High-performance anion-exchange chromatography with pulsed amperometric detection for carbohydrate analysis of glycoproteins. *Biochemistry (Moscow)*, 78:697-709.

Rojas-Macias MA, Mariethoz J, Andersson P, Jin C, Venkatakrishnan V, Aoki NP, Shinmachi D, Ashwood C, Madunic K, Zhang T. 2019. Towards a standardized bioinformatics infrastructure for N-and O-glycomics. *Nature communications*, 10:1-10.

Ruhaak LR, Xu G, Li Q, Goonatileke E, Lebrilla CB. 2018. Mass spectrometry approaches to glycomic and glycoproteomic analyses. *Chem Rev*, 118:7886-7930.

Schjoldager KT, Narimatsu Y, Joshi HJ, Clausen H. 2020. Global view of human protein glycosylation pathways and functions. *Nat Rev Mol Cell Biol*, 21:729-749.

Stockert RJ. 1995. The asialoglycoprotein receptor: relationships between structure, function, and expression. *Physiol Rev*, 75:591-609.

Trbojević Akmačić I, Ventham NT, Theodoratou E, Vučković F, Kennedy NA, Krištić J, Nimmo ER, Kalla R, Drummond H, Štambuk J. 2015. Inflammatory bowel disease associates with proinflammatory potential of the immunoglobulin G glycome. *Inflamm Bowel Dis*, 21:1237-1247.

Ulfman LH, Leusen JH, Savelkoul HF, Warner JO, Van Neerven RJ. 2018. Effects of bovine immunoglobulins on immune function, allergy, and infection. *Front Nutr*, 5:52.

van Leeuwen SS, Schoemaker RJ, Timmer CJ, Kamerling JP, Dijkhuizen L. 2012. N- and O-glycosylation of a commercial bovine whey protein product. *J Agric Food Chem*, 60:12553-12564.

Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Marth JD, Bertozzi CR, Hart GW, Etzler ME. 2009. Symbol nomenclature for glycan representation. *Proteomics*, 9:5398-5399.

Weaver DM, Tyler JW, VanMetre DC, Hostetler DE, Barrington GM. 2000. Passive transfer of colostrum immunoglobulins in calves. *J Vet Intern Med*, 14:569-577.

Weström B, Arévalo Sureda E, Pierzynowska K, Pierzynowski SG, Pérez-Cano F-J. 2020. The immature gut barrier and its importance in establishing immunity in newborn mammals. *Front Immunol*, 11:1153.

Yu H, Shu J, Li Z. 2020. Lectin microarrays for glycoproteomics: an overview of their use and potential. *Expert Review of Proteomics*, 17:27-39.

Zhu A, Hurst R. 2002. Anti-N-glycolylneuraminic acid antibodies identified in healthy human serum. *Xenotransplantation*, 9:376-381.

Zhu J, Lin Y-H, Dingess KA, Mank M, Stahl B, Heck AJ. 2020. Quantitative longitudinal inventory of the N-glycoproteome of human milk from a single donor reveals the highly variable repertoire and dynamic site-specific changes. *J Proteome Res*, 19:1941-1952.

Overview of supplementary information for this chapter

Supplementary figures

Supplementary tables

Supplementary figures

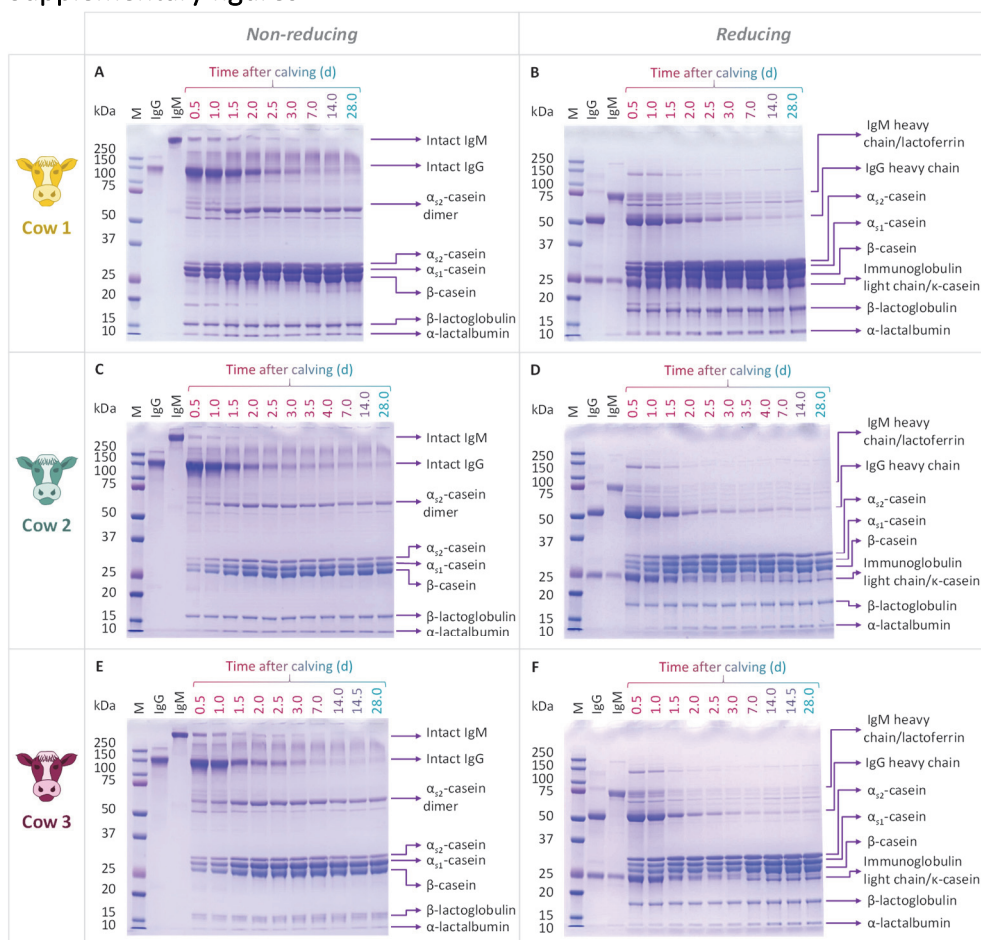


Figure S3.1 - SDS-PAGE gel images of the samples collected from Cow 1 (A and B), Cow 2 (C and D) and Cow 3 (E and F), analysed under non-reducing (A, C and E) and reducing (B, D and F) conditions. All bovine milk samples were loaded on the gel at 20 μ g protein/well. Bovine serum IgG and IgM were loaded on the gel at 4 μ g protein/well. M = molecular mass marker.

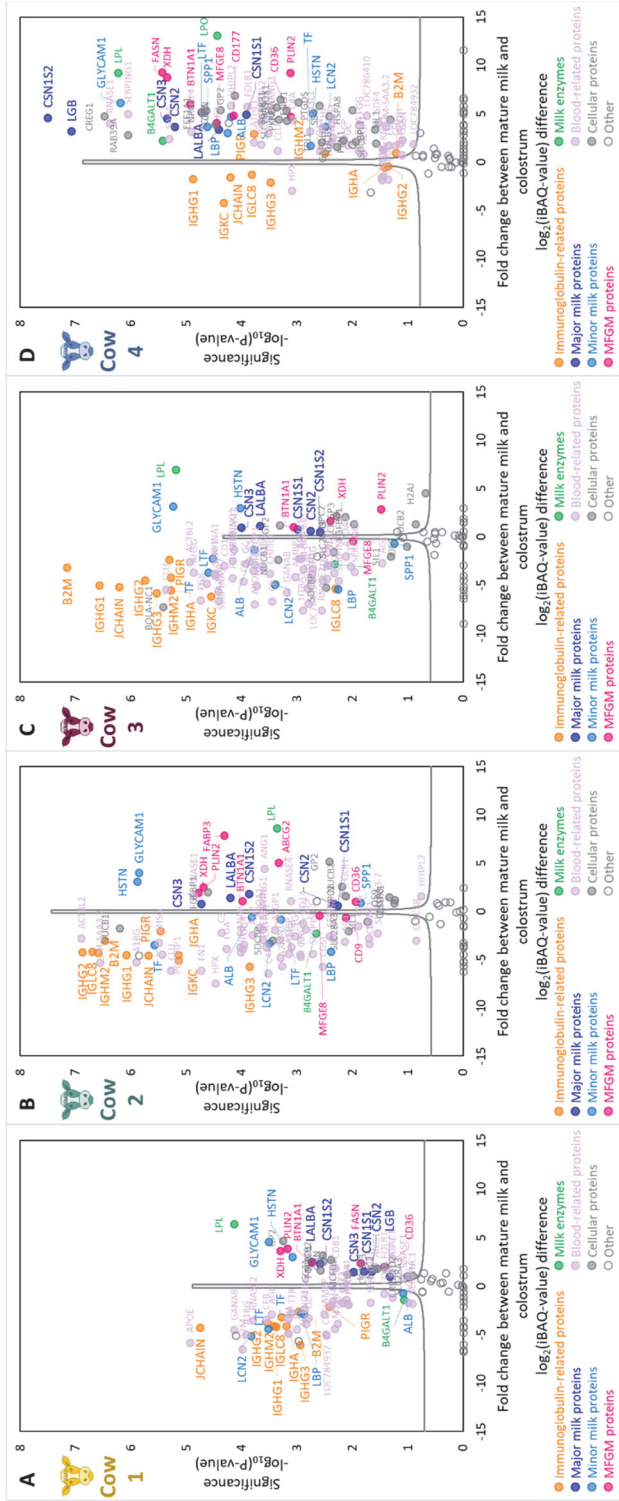


Figure S3.2 - Volcano plots depicting the proteins with significantly different abundances in the proteomes of the samples collected at 0.5 and 28 d after calving from the four individual cows: (A) Cow 1, (B) Cow 2, (C) Cow 3 and (D) Cow 4. While the most abundant colostrum and milk proteins show the same trend, Cow 4 appears to represent an outlier, with blood serum and cellular related proteins being significantly increased in the mature milk.

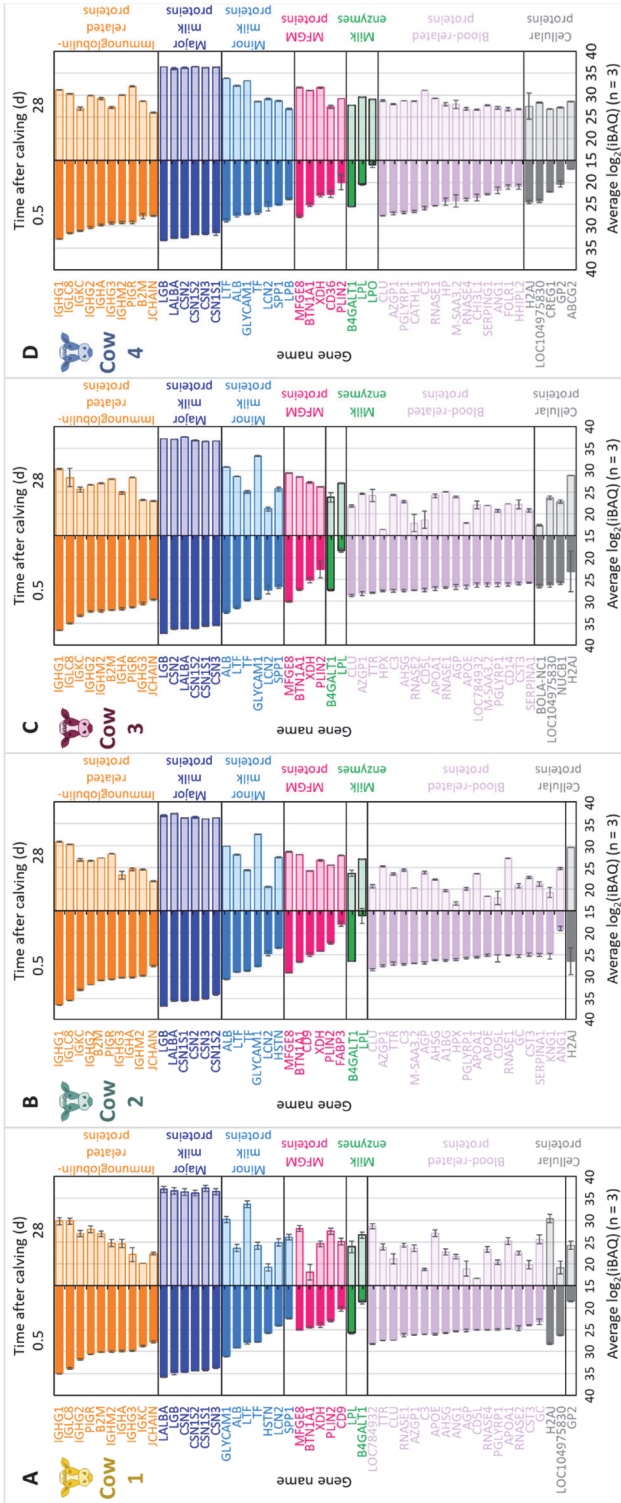


Figure S3.3 - Differences in intensity Based Absolute Quantification abundance ($\log_2(\text{ibaQ})$) between the 50 most abundant protein sequences present in the proteomes of samples collected at 0.5 (left) and 28 (right) d after calving from the four individual cows: (A) Cow 1, (B) Cow 2, (C) Cow 3 and (D) Cow 4. The 50 most abundant protein sequences were determined as explained in the caption of Figure 3.2. The error bars indicate the standard deviation of each protein sequence ibaQ -value from triplicate measurements.

NeuAc sialylation is a hallmark of bovine colostrum IgG glycosylation

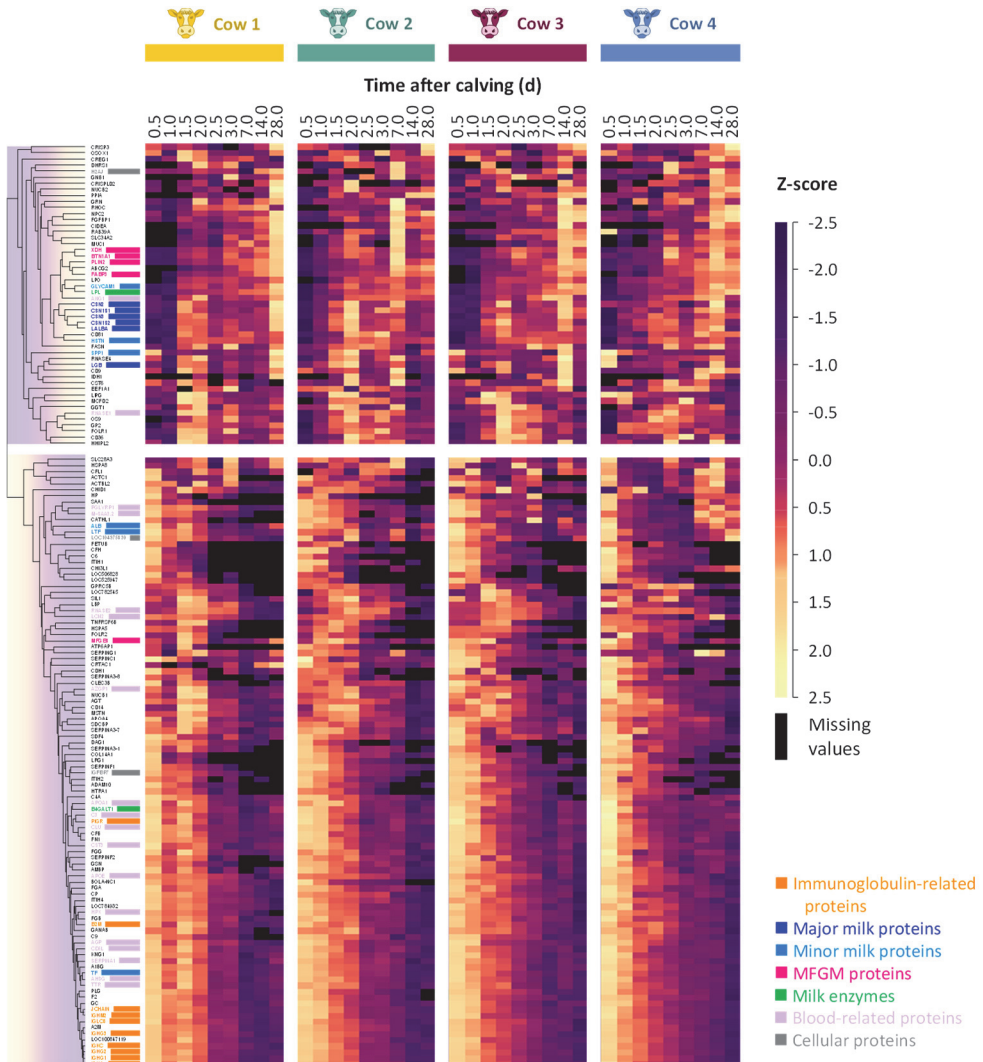


Figure S3.4 - Dynamic proteome of bovine milk during lactation in the four individual cows, i.e., Cow 1, Cow 2, Cow 3 and Cow 4, determined by bottom-up mass spectrometry. The figure depicts a heatmap visualizing hierarchical clustering of individual proteins based on the Z-scores of $\log_2(\text{iBAQ})$ values across time. Highlighted are the 50 most abundant proteins in bovine colostrum and mature milk combined; the colour codes depicting the different protein categories are consistent with Figure 3.2A and B.

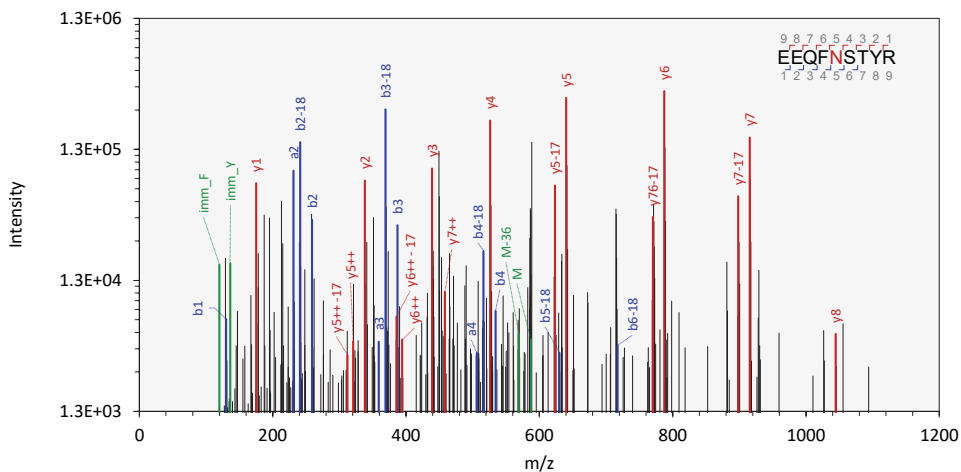


Figure S3.5 - Representative LC-MS/MS spectrum illustrating a non-glycosylated peptide covering the bovine IgG N-glycosylation motif, identified and annotated with Byonic v4.5.2. The illustrated spectrum is scan number 3877 in the run of IgG captured from the 1 d colostrum of Cow 2, analysed with oxonium ion-triggered stepped higher-energy collisional dissociation.

NeuAc sialylation is a hallmark of bovine colostrum IgG glycosylation

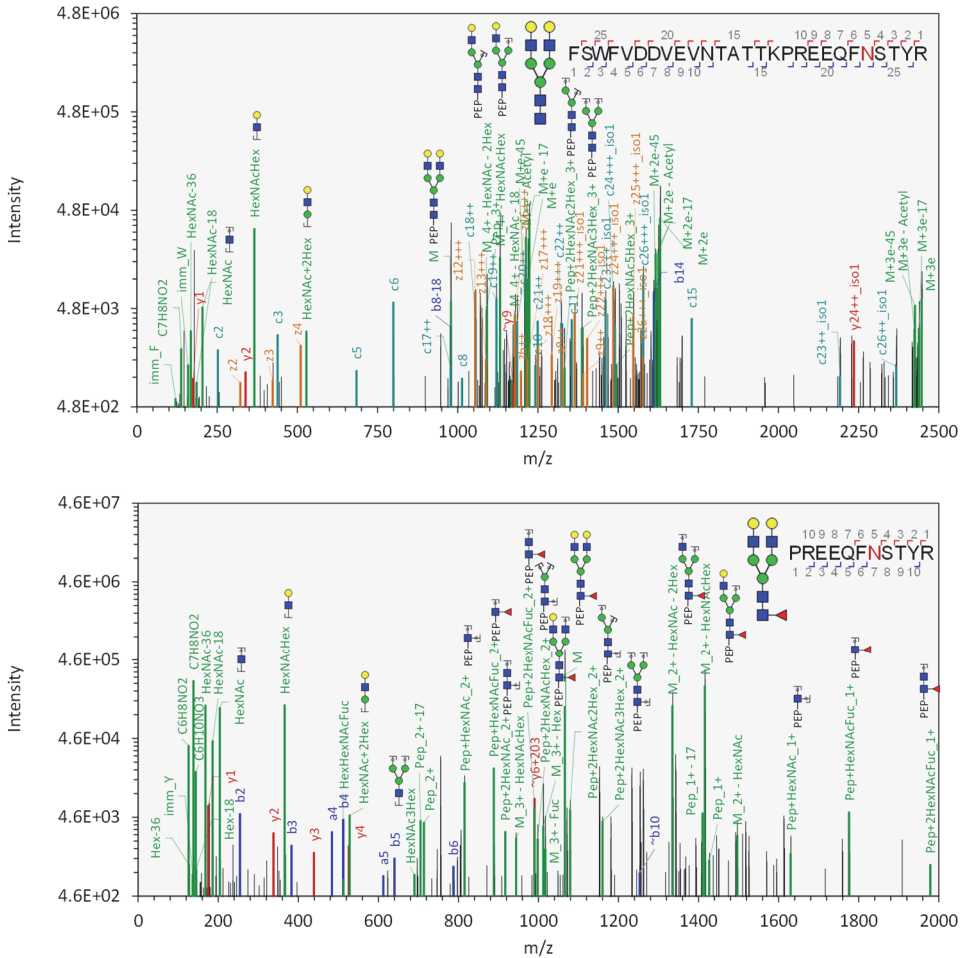


Figure S3.6 - Representative LC-MS/MS spectra illustrating bovine IgG glycopeptides occupied with neutral glycans, identified and annotated with Byonic v4.5.2. The spectrum on the top, illustrating a non-fucosylated glycopeptide, is scan number 10692 in the run of IgG captured from the 14 d transitional milk of Cow 1, analysed with oxonium ion-triggered electron-transfer/higher-energy collisional dissociation. The spectrum on the bottom, illustrating a fucosylated glycopeptide, is scan number 2990 in the run of IgG captured from the 0.5-d colostrum of Cow 2, analysed with oxonium ion-triggered stepped higher-energy collisional dissociation. Proposed intact glycan, oxonium ion and peptide-bound glycan fragment structures were built using GlycoWorkBench 2.1 build 146.

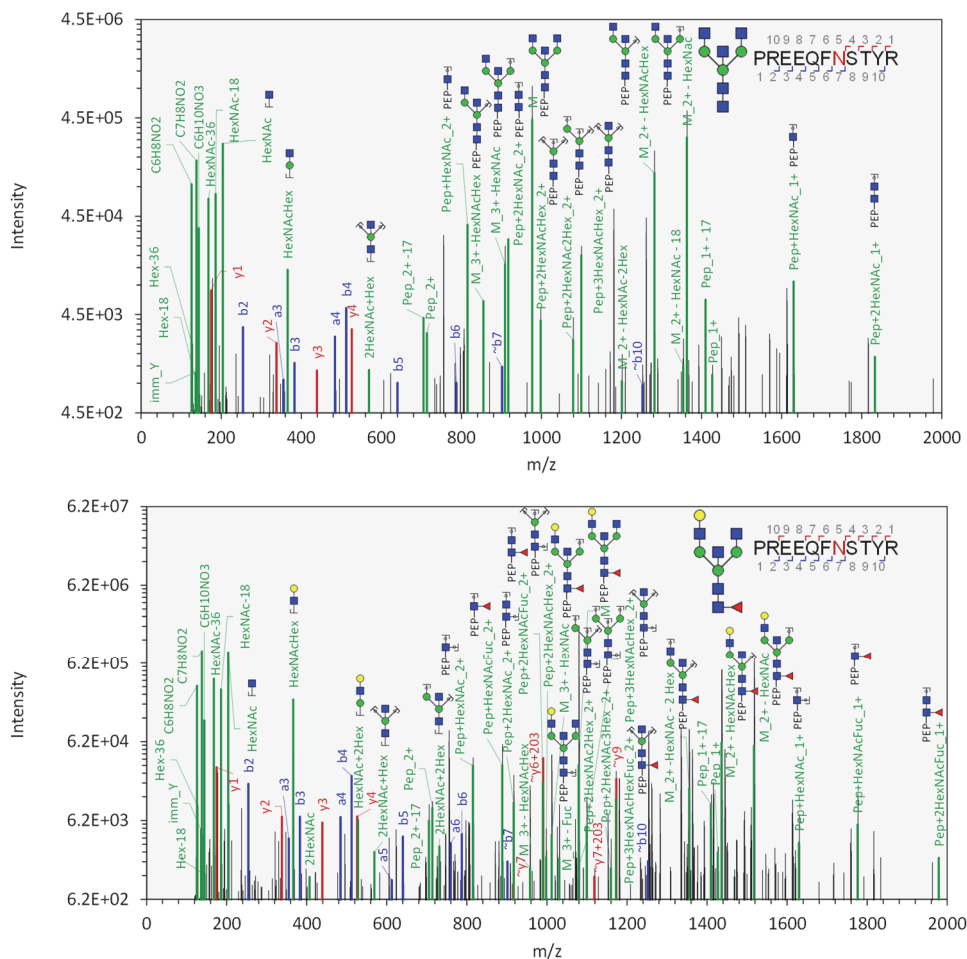


Figure S3.7 - Representative LC-MS/MS spectra illustrating bovine IgG glycopeptides occupied with neutral glycans containing bisecting GlcNAc, identified and annotated with Byonic v4.5.2. The spectrum on the top, illustrating a non-fucosylated glycopeptide, is scan number 2722 in the run of IgG captured from the 2.5 d colostrum of Cow 1, analysed with oxonium ion-triggered stepped higher-energy collisional dissociation (HCD). The spectrum on the bottom, illustrating a fucosylated glycopeptide, is scan number 2627 in the run of IgG captured from the 1 d colostrum of Cow 2, analysed with oxonium ion-triggered stepped HCD. The ions characteristic for the fragmentation of the bisecting GlcNAc-containing-glycopeptides, i.e., $\text{Pep}+3\text{HexNAcHex}_2+$ ($m/z = 1099.4851$) in the top spectrum and $\text{Pep}+3\text{HexNAcHex}_2+$ ($m/z = 1099.4877$) and $\text{Pep}+3\text{HexNAcHexFuc}_2+$ ($m/z = 1172.5178$) in the bottom spectrum were annotated manually. Proposed intact glycan, oxonium ion and peptide-bound glycan fragment structures were built using GlycoWorkBench 2.1 build 146.

NeuAc sialylation is a hallmark of bovine colostrum IgG glycosylation

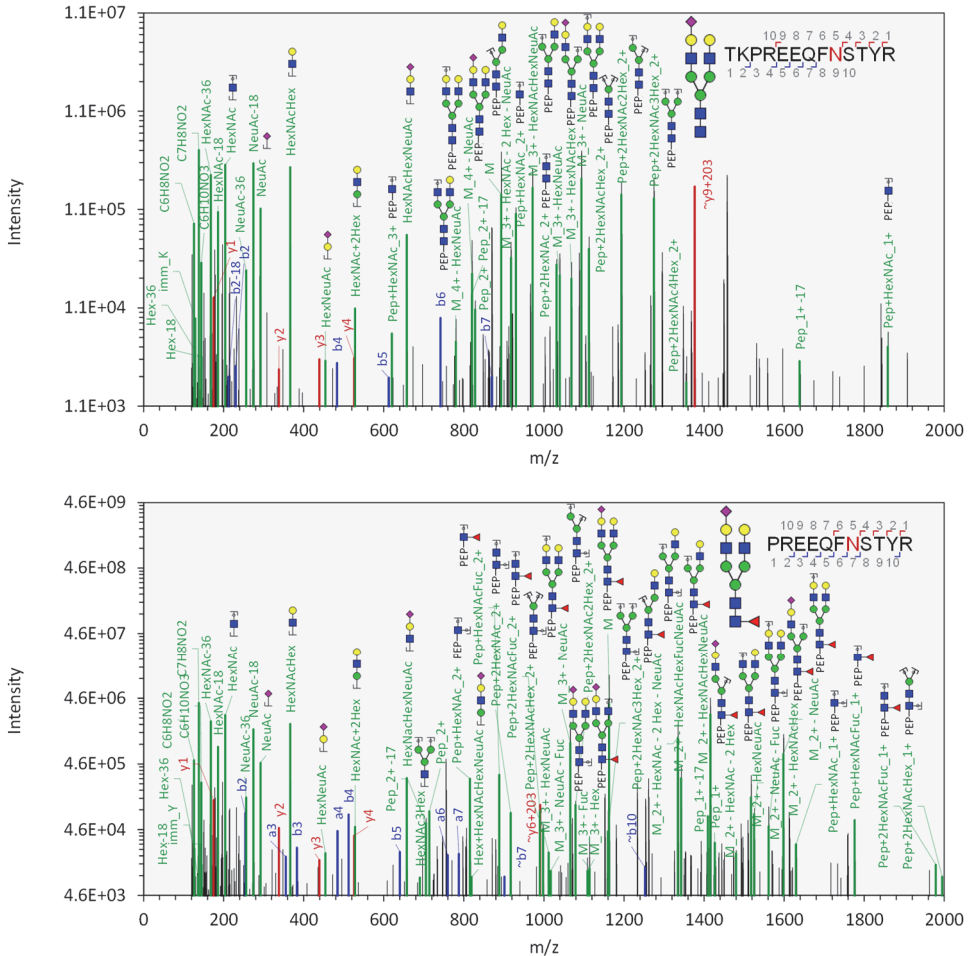


Figure S3.8 - Representative LC-MS/MS spectra illustrating bovine IgG glycopeptides occupied with glycans sialylated with NeuAc, identified and annotated with Byonic v4.5.2. The spectrum on the top, illustrating a non-fucosylated glycopeptide, is scan number 2494 in the run of IgG captured from the 0.5-d colostrum of Cow 4, analysed with oxonium ion-triggered higher-energy collisional dissociation. The spectrum on the bottom, illustrating a fucosylated glycopeptide, is scan number 3061 in the run of IgG captured from the 2.5 d colostrum of Cow 4, analysed with oxonium ion-triggered stepped HCD. Proposed intact glycan, oxonium ion and peptide-bound glycan fragment structures were built using GlycoWorkBench 2.1 build 146.

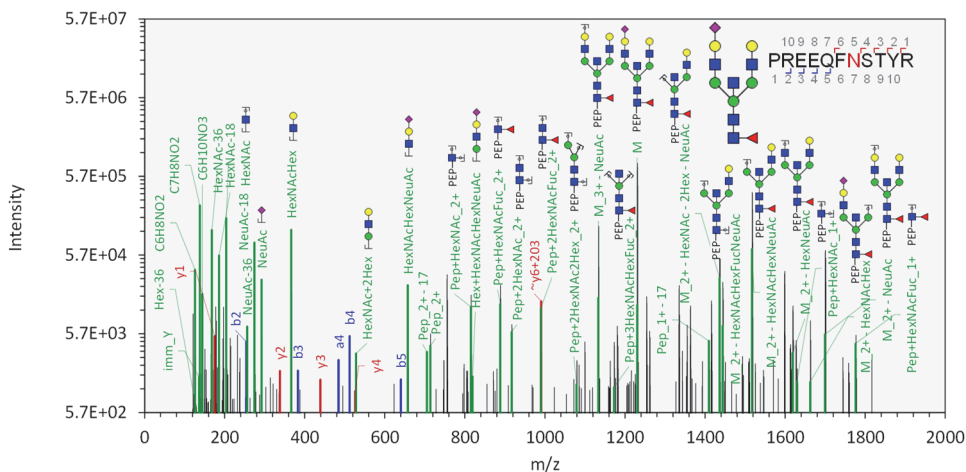


Figure S3.9 - Representative LC-MS/MS spectrum illustrating a bovine IgG glycopeptide occupied with a NeuAc sialylated glycan, containing fucosylation and bisecting GlcNAc, identified and annotated with Byonic v4.5.2. The illustrated spectrum is scan number 2989 in the run of IgG captured from the 2.5 d colostrum of Cow 3, analysed with oxonium ion-triggered stepped higher-energy collisional dissociation. The ion characteristic for the fragmentation of the bisecting GlcNAc-containing-glycopeptides, i.e., $\text{Pep}+3\text{HexNAcHexFuc}_2+$ ($m/z = 1172.5120$), was annotated manually. Proposed intact glycan, oxonium ion and peptide-bound glycan fragment structures were built using GlycoWorkBench 2.1 build 146.

NeuAc sialylation is a hallmark of bovine colostrum IgG glycosylation

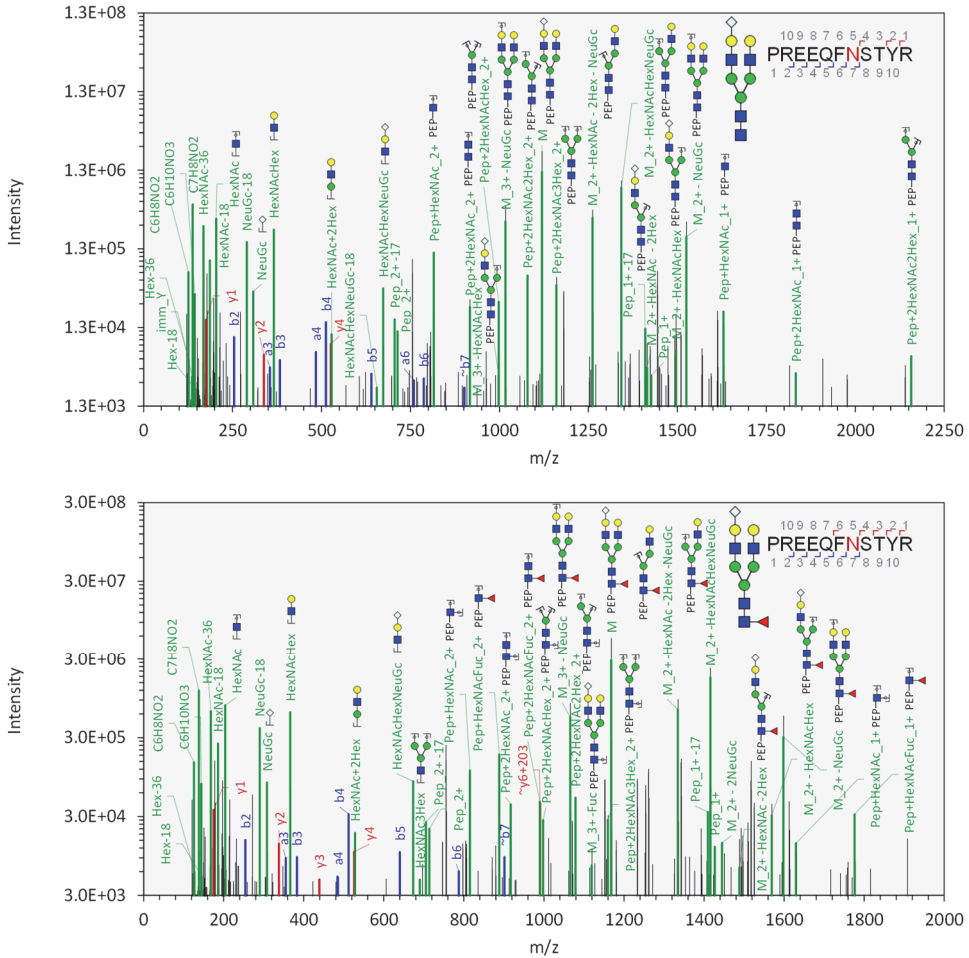


Figure S3.10 - Representative LC-MS/MS spectra illustrating bovine IgG glycopeptides occupied with glycans sialylated with NeuGc, identified and annotated with Byonic v4.5.2. The spectrum on the top, illustrating a non-fucosylated glycopeptide, is scan number 3037 in the run of IgG captured from the 1 d colostrum of Cow 3, analysed with oxonium ion-triggered higher-energy collisional dissociation. The spectrum on the bottom, illustrating a fucosylated glycopeptide, is scan number 2979 in the run of IgG captured from the 0.5-d colostrum of Cow 4, analysed with oxonium ion-triggered stepped HCD. Proposed intact glycan, oxonium ion and peptide-bound glycan fragment structures were built using GlycoWorkBench 2.1 build 146.

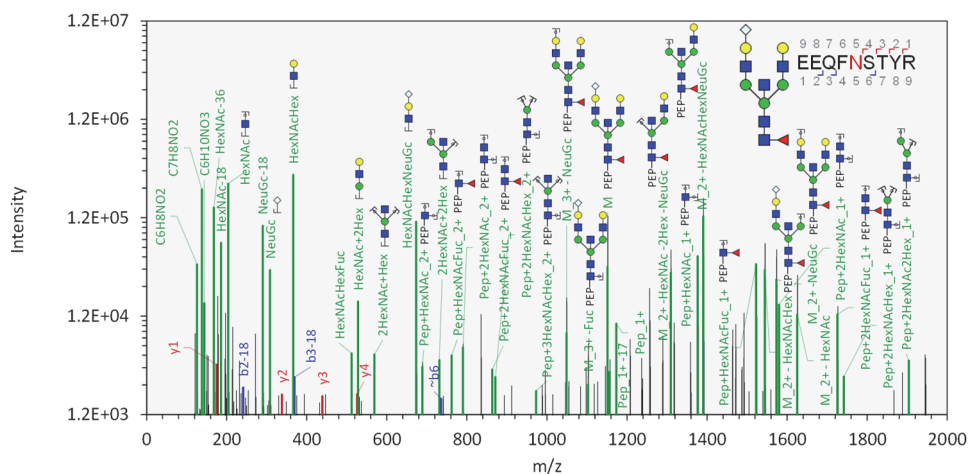
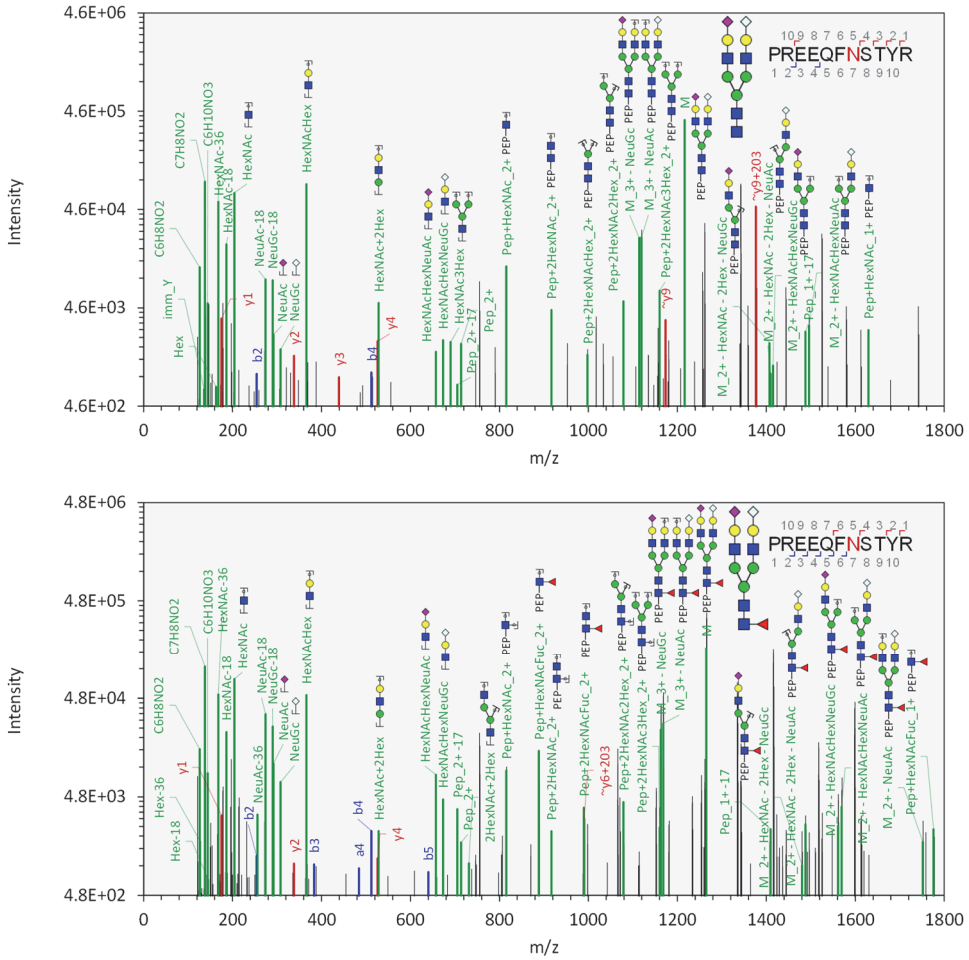


Figure S3.11 - Representative LC-MS/MS spectrum illustrating a bovine IgG glycopeptide occupied with a NeuGc sialylated glycan, containing fucosylation and bisecting GlcNAc, identified and annotated with Byonic v4.5.2. The illustrated spectrum is scan number 3462 in the run of IgG captured from the 1 d colostrum of Cow 2, analysed with oxonium ion-triggered stepped higher-energy collisional dissociation. The ion characteristic for the fragmentation of the bisecting GlcNAc-containing-glycopeptides, i.e., *Pep+3HexNAcHex_2+* ($m/z = 972.9130$), was annotated manually. Proposed intact glycan, oxonium ion and peptide-bound glycan fragment structures were built using GlycoWorkBench 2.1 build 146.

NeuAc sialylation is a hallmark of bovine colostrum IgG glycosylation



3

Figure S3.12 - Representative LC-MS/MS spectra illustrating bovine IgG glycopeptides occupied with glycans sialylated with both NeuAc and NeuGc, identified and annotated with Byonic v4.5.2. The spectrum on the top, illustrating a non-fucosylated glycopeptide, is scan number 3620 in the run of IgG captured from the 1.5 d colostrum of Cow 4, analysed with oxonium ion-triggered higher-energy collisional dissociation. The spectrum on the bottom, illustrating a fucosylated glycopeptide, is scan number 3492 in the run of IgG captured from the 1.5 d colostrum of Cow 1, analysed with oxonium ion-triggered stepped HCD. Proposed intact glycan, oxonium ion and peptide-bound glycan fragment structures were built using GlycoWorkBench 2.1 build 146.

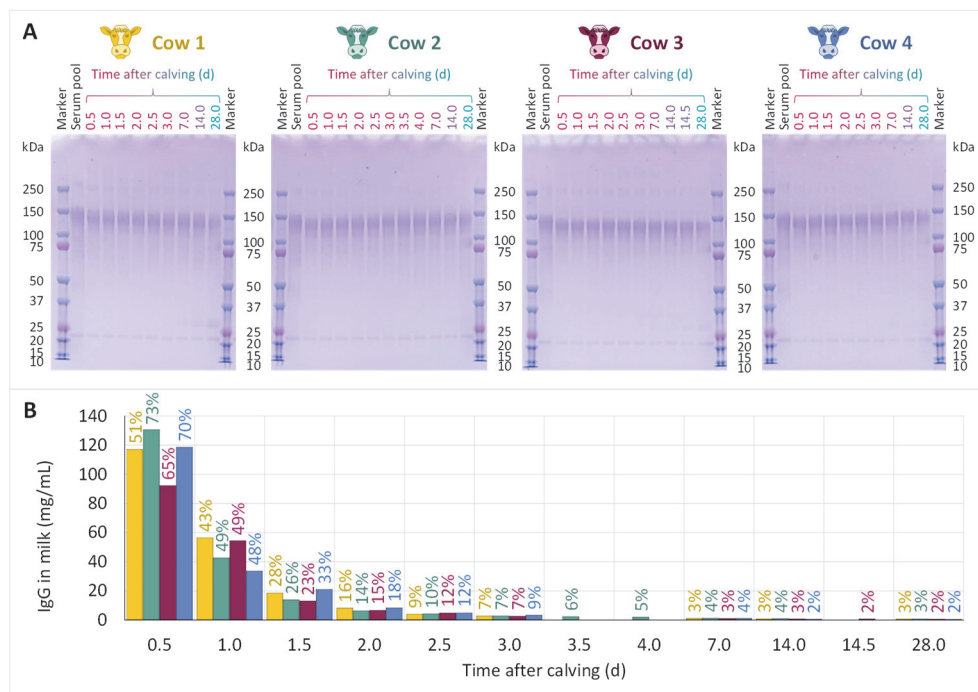


Figure S3.13 - IgG captured with Protein G from the bovine milk samples collected during lactation from the four cows, i.e., Cow 1 (■), Cow 2 (■), Cow 3 (■) and Cow 4 (■). (A) SDS-PAGE gel images of the captured IgG samples compared to the reference bovine serum pool IgG. All samples were prepared in non-reducing conditions and were loaded on the gels at 2 μ g protein/well. (B) Concentration of IgG in the bovine milk samples determined by measuring the amount of captured IgG by Nanodrop, and correcting it for the capturing efficiency assessed on the bovine serum pool IgG. The data labels indicate the percentage of total protein in the sample made up by IgG.

NeuAc sialylation is a hallmark of bovine colostrum IgG glycosylation

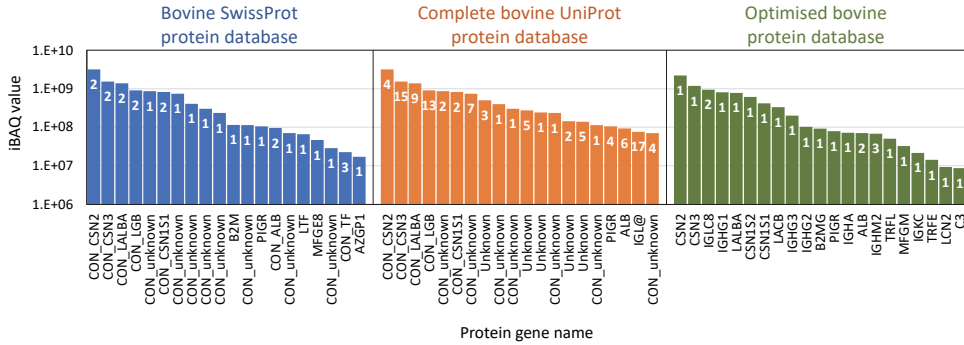


Figure S3.14 - Effect of protein database on search results of peptide-centric LC-MS/MS data. Illustrated are the 20 protein sequences with the highest iBAQ-values identified in the bovine colostrum of Cow 1 at 0.5 d post-partum using a search as described in materials and methods with the following databases: 1) the bovine SwissProt protein database (*Bos taurus*, taxon ID 9913, downloaded on June 3rd, 2022) + automated MaxQuant contaminants database, 2) the bovine UniProt protein database (*Bos taurus*, taxon ID 9913, downloaded on June 3rd, 2022) + automated MaxQuant contaminants database, and 3) the optimised bovine protein + contaminants database, respectively. The data labels indicate the numbers of proteins identified per protein group. Due to its incompleteness, i.e., ~6,000 out of total ~20,000 protein sequences encoded by unique genes, the database search against SwissProt failed to identify the major immunoglobulin components of bovine colostrum. The database search against the complete UniProt protein database did also not result in identification of the immunoglobulin components, due to incomplete annotation of protein and gene names, resulting in the identification of uncharacterized proteins. Furthermore, due to the high redundancy of the complete UniProt protein database (i.e., ~47,000 protein sequences for ~20,000 protein-encoding genes), high numbers of proteins were identified per protein group (the number in the bar), negatively impacting the quantification results. The presence of bovine protein sequences in the automated MaxQuant contaminants database further resulted in the identification of major bovine colostrum and milk proteins as sample contaminants. The optimised bovine protein database overcame the shortcomings of the previous two databases, considerably improving the identification and relative quantification of the key colostrum and milk protein components.

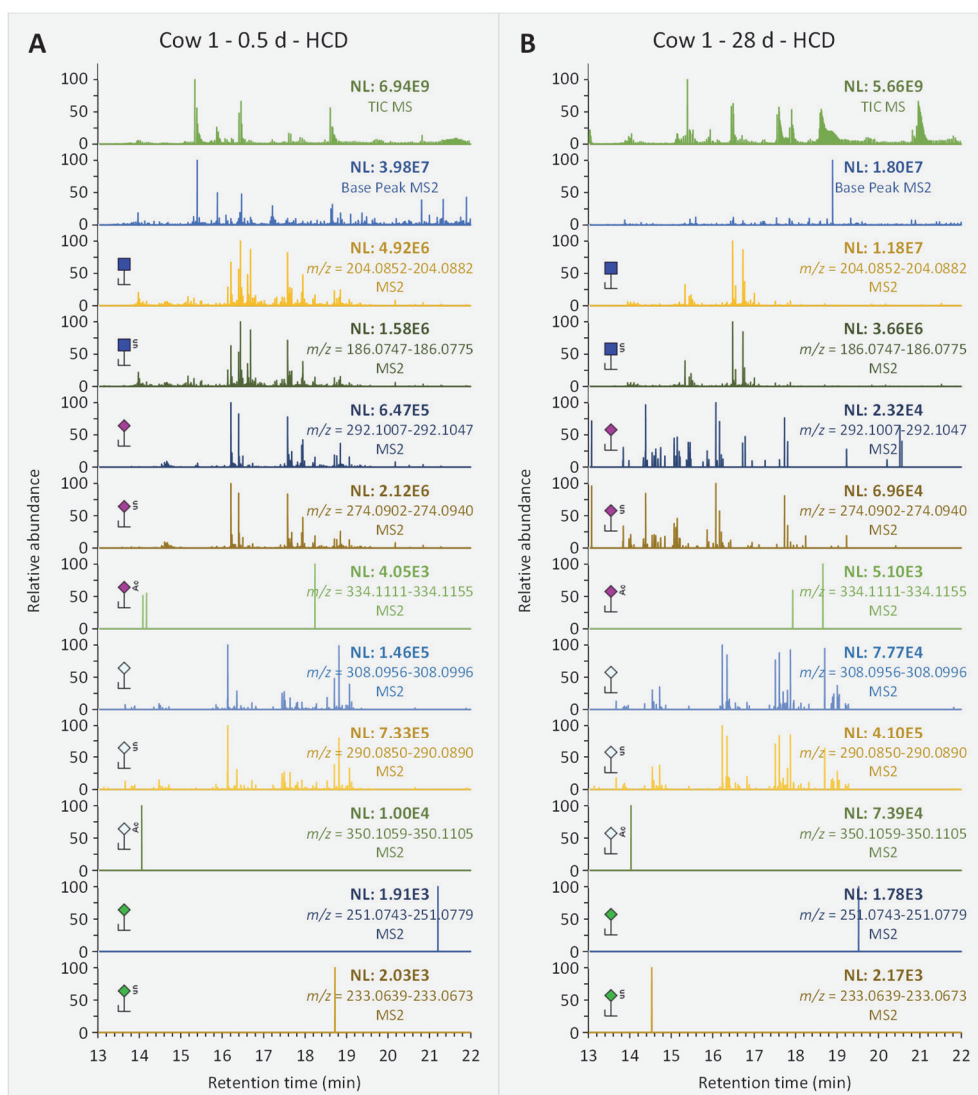


Figure S3.15 - Oxonium ion evidence for the occurrence of different types of sialic acid in the N-glycosylation of bovine IgG from (A) colostrum at 0.5 d post-partum and (B) mature milk from 28 d post-partum from Cow 1. Presented here is the retention time window of 13-22 min, wherein the Byonic search picked up all IgG glycopeptides. From top to bottom in both panels (A) and (B), the data represents: total ion count (TIC MS), base peak MS2, and MS2 traces of the following oxonium ions: N-acetylhexosamine (HexNAc), HexNAc-H₂O, N-acetylneuraminic acid (NeuAc), NeuAc-H₂O, acetylated NeuAc, N-glycolylneuraminic acid (NeuGc), NeuGc-H₂O, acetylated NeuGc, deaminated neuraminic acid (Kdn) and Kdn-H₂O. The illustration of the oxonium ions was done with GlycoWorkBench 2.1 build 146, according to the symbol nomenclature for glycan representation of the Consortium for Functional Glycomics (Varki, Cummings et al. 2009).

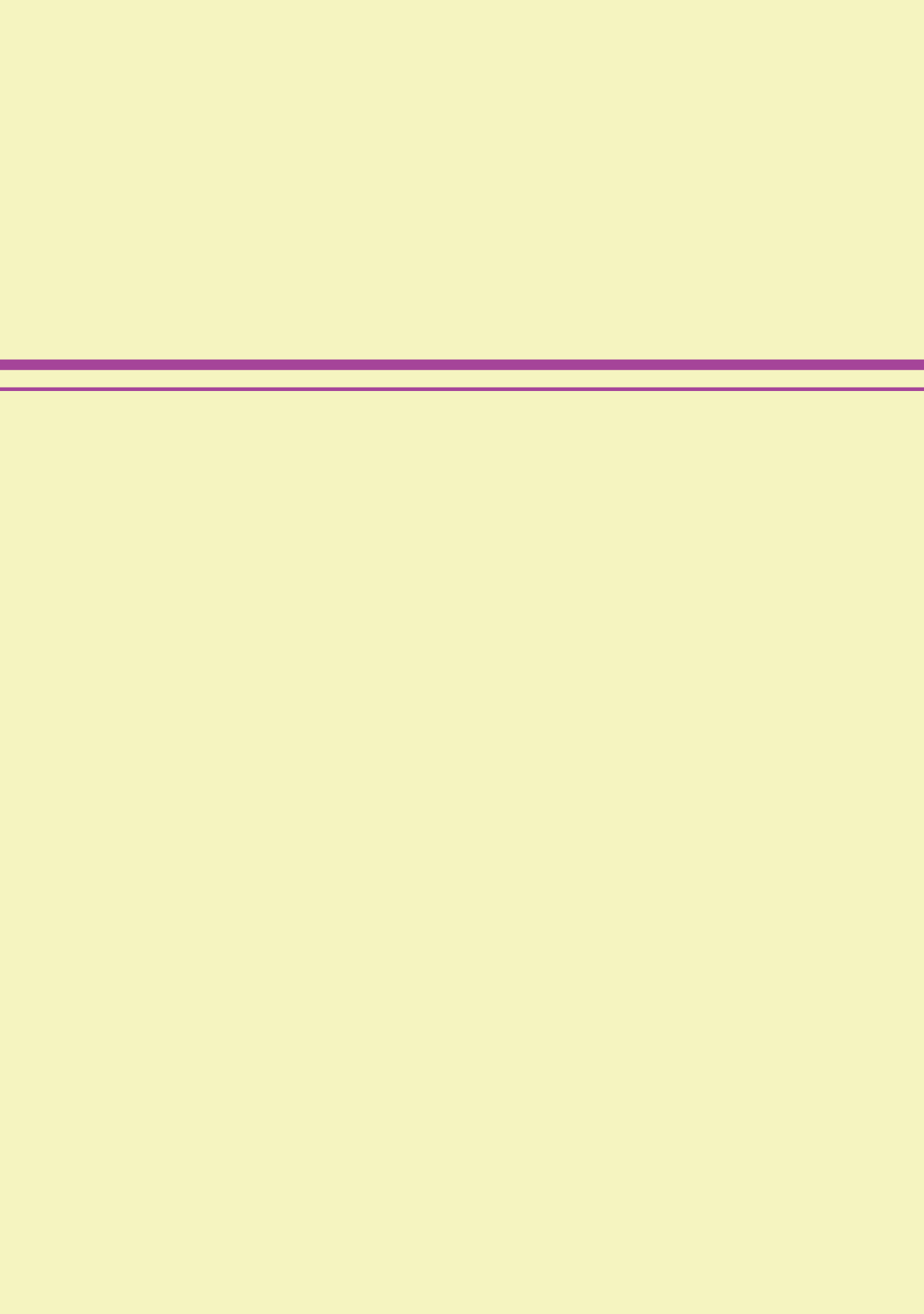
Supplementary tables

Table S3.2 - List of oxonium ions used for triggering hybrid fragmentation of the glycopeptides.

<i>m/z</i>	Chemical formula	Glycan composition
127.039	[C ₆ H ₇ O ₃] ⁺	Hex - 2(H ₂ O)
145.05	[C ₆ H ₉ O ₄] ⁺	Hex - H ₂ O
163.06	[C ₆ H ₁₁ O ₅] ⁺	Hex
243.026	[C ₆ H ₁₂ O ₈ P] ⁺	Hex-PO ₄
405.079	[C ₁₂ H ₂₂ O ₁₃ P] ⁺	Hex ₂ -PO ₄
138.055	[C ₇ H ₈ NO ₂] ⁺	HexNAC - CH ₆ O ₃
168.066	[C ₈ H ₁₀ NO ₃] ⁺	HexNAC - 2(H ₂ O)
186.076	[C ₈ H ₁₂ NO ₄] ⁺	HexNAC - H ₂ O
204.087	[C ₈ H ₁₄ NO ₅] ⁺	HexNAC
274.092	[C ₁₁ H ₁₆ NO ₇] ⁺	NeuAc - H ₂ O
292.103	[C ₁₁ H ₁₈ NO ₈] ⁺	NeuAc
366.14	[C ₁₄ H ₂₄ NO ₁₀] ⁺	HexNAC-Hex
407.166	[C ₁₆ H ₂₇ N ₂ O ₁₀] ⁺	HexNAC ₂
512.197	[C ₂₀ H ₃₄ NO ₁₄] ⁺	HexNAC-Hex-dHex
657.235	[C ₂₅ H ₄₁ N ₂ O ₁₈] ⁺	HexNAC-Hex-NeuAc
290.087	[C ₁₁ H ₁₆ NO ₈] ⁺	NeuGc - H ₂ O
308.098	[C ₁₁ H ₁₈ NO ₉] ⁺	NeuGc

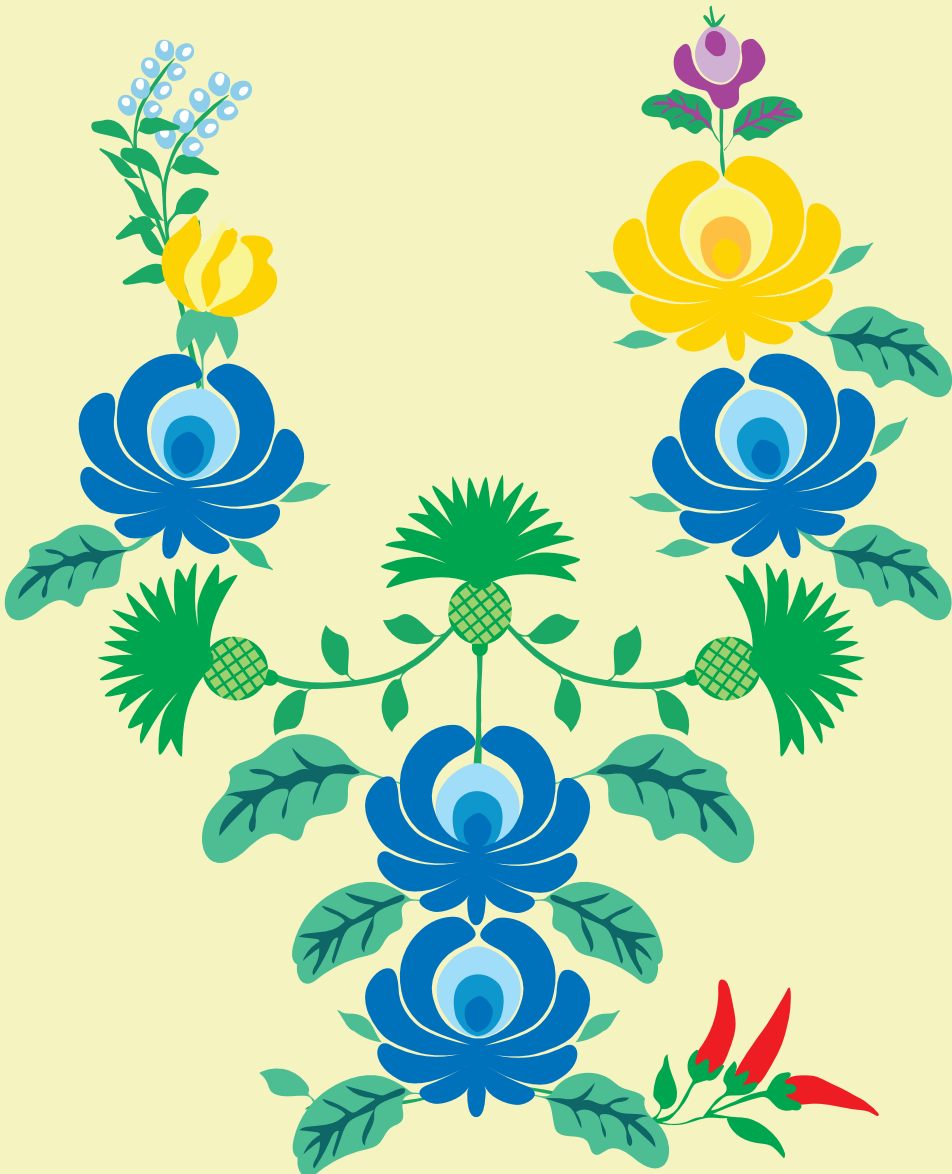
HexNAC = *N*-acetylhexosamine; Hex = hexose;
dHex = deoxyhexose; NeuAc = *N*-acetylneuraminic acid;
NeuGc = *N*-glycolylneuraminic acid.

Supplementary Tables S3.1 and S3.3 have not been printed in this book due to their large sizes. These tables can be accessed online at <https://doi.org/10.1093/glycob/cwad001>.



CHAPTER 4

The *N*-glycoproteome of bovine mammary secretions from colostrum to mature milk, with detailed insights into the *N*-glycosylation of α -lactalbumin



Chapter 4. The *N*-glycoproteome of bovine mammary secretions from colostrum to mature milk, with detailed insights into the *N*-glycosylation of α -lactalbumin

Inge Gazi, Karli R. Reiding, André Groeneveld, Jan Bastiaans,

Thom Huppertz and Albert J. R. Heck

Abstract

Protein glycosylation is a complex post-translational modification that directly impacts a plethora of biological processes and functions. Glycosylated proteins are highly prevalent in milk, contributing to the functional properties of the milk. Here we investigated the changes in the *N*-glycoproteome in colostrum and milk sampled from four individual cows, each at 9 time points in the interval of 0.5-28.0 d after calving. Our data indicate that changes in the *N*-glycoproteome from colostrum transitioning to mature milk occurred at two levels: 1) relative changes in (*N*-glyco)protein abundances, and 2) changes in protein-specific *N*-glycosylation during lactation. The bovine colostrum *N*-glycoproteome primarily consisted of blood serum-derived proteins, the signature biochemical feature of which was identified as *N*-glycolylneuraminic acid (NeuGc) sialylation. The relative abundance of colostrum *N*-glycoproteins swiftly decreased in the first 3 d after calving, with a consequent decrease of NeuGc sialylation. Unlike colostrum, the *N*-glycoproteome of transitional and mature bovine milk was characterised by *N*-acetylneuraminic acid (NeuAc) sialylation. These findings further indicated that colostrum and mature milk *N*-glycoproteins had largely different origins, having been produced by different cells. Of all identified *N*-glycoproteins, we presented a detailed insight on α -lactalbumin, the second most abundant whey protein in bovine milk, and of particular importance for the infant formula industry. We identified *N*-glycosylation in bovine α -lactalbumin, consisting of a repertoire of complex glycan structures at both Asn45 and Asn74. We found similar glycan profiles in all four cows, with partial site occupancies averaging at 35% and 4% for Asn45 and Asn74, respectively, with no substantial changes in occupancy over lactation. Fucosylation, sialylation primarily with NeuAc, but also with NeuGc, and a high ratio of LacdiNAc/LacNAc antennae were characteristic biochemical features of the identified α -lactalbumin glycans. The glycoproteoform profile showed lactational changes, with significant increases in neutral glycans and number of LacNAc antennae *per* glycan, and significant decreases in numbers of NeuAc residues and LacdiNAc antennae *per* glycan. Building on the complexity and dynamics of bovine colostrum and milk *N*-glycosylation, further studies are warranted to investigate the potential functionality of these changes, also in the context of human nutrition and health.

Introduction

Colostrum and milk are multifunctional biological fluids meant to nourish and support the immune, digestive and cognitive development of new-borns and developing young infants (McGrath *et al.*, 2016). Mammary secretions contain many classes of bioactive components, one of them being the carbohydrate fraction. Carbohydrates in milk are present free in the form of lactose and free

oligosaccharides, and in glycoconjugates such as glycolipids and glycoproteins with *N*- and *O*-linked glycans (Gopal and Gill, 2000). Protein glycosylation, in particular, is of interest due to its function to modulate protein interactions and activity, increase stability against proteolytic degradation, bind pathogens, and promote gut health and a balanced microbiome (Krištić and Lauc, 2017). Milk protein glycosylation shows glycoprotein- and site-specificity (O’Riordan *et al.*, 2014), and undergoes dynamic changes throughout lactation (Gazi *et al.*, 2023, Takimori *et al.*, 2011, Zhu *et al.*, 2020). With milk, of which the majority is represented by bovine milk, occupying a core position in a balanced diet owing to its functional and nutritional properties, the study of bovine milk protein glycosylation is necessary to understand its biological functionality in the context of human nutrition.

Many of the studies published to date investigating bovine milk glycosylation rely on the analysis of enzymatically released *N*-glycans (Takimori *et al.*, 2011, Valk-Weeber *et al.*, 2020) and/or the analysis of deglycosylated peptides (Cao *et al.*, 2019, Xiao *et al.*, 2022). These approaches convey a general picture of the milk *N*-glycome, and provide indirect information on the *N*-glycoproteome itself. However, in a previous study we instead presented a glycoproteomics approach for the study of bovine milk IgG constant heavy domain 2 (CH2) *N*-glycosylation, where we identified glycopeptide amino acid sequence, and *N*-glycan structure, composition and localization from the direct analysis of *N*-glycopeptides (Gazi *et al.*, 2023). There, we found bovine milk IgG CH2 *N*-glycosylation consisting of complex diantennary glycans containing on average 0.6 fucose and 0.2 *N*-glycolylneuraminic acid (NeuGc) residues *per* glycan, with a small proportion of the glycan structures also containing bisecting *N*-acetylglucosamine (GlcNAc). However, the IgG CH2 *N*-glycan biochemical feature that showed the strongest lactational dependency was NeuAc, which was identified as the hallmark of bovine colostrum IgG *N*-glycosylation, significantly decreasing in the first days after calving, and barely detectable on mature bovine milk IgG. Here, we applied a similar approach to characterise the whole detectable *N*-glycoproteome of bovine colostrum and milk in samples from four individual cows collected at 0.5-28.0 d after calving, without pre-enrichment of glycoproteins or glycopeptides.

As part of the glycoproteomics findings described in this chapter, we present a detailed analysis of the *N*-glycosylation of bovine α -lactalbumin. α -Lactalbumin is one of the most abundant proteins in milk, it is exclusively produced in the lactating mammary gland, and it is present in the milk of all mammals (Brew, 2013). In human milk, α -lactalbumin makes up approximately 22% of the total protein content, constituting a source of bioactive peptides and essential amino acids (Layman *et al.*, 2018). For this reason, bovine α -lactalbumin is a protein of interest in the

production of humanised, bovine milk-based infant formula. α -Lactalbumin is a member of the lysozyme superfamily, sharing structural similarity with lysozyme C, from which it presumably evolved by gene duplication and mutation (Mckenzie and White Jr, 1991). Nevertheless, while they share structural similarity, α -lactalbumin is functionally very different from lysozyme C (Qasba *et al.*, 1997). α -Lactalbumin is the regulatory subunit in lactose synthase, the enzyme responsible for the synthesis of lactose for milk (Brodbeck *et al.*, 1967). Lactose synthase consists of beta-1,4-galactosyltransferase 1 (B4GALT1) and α -lactalbumin. B4GALT1 catalyses the reaction between uridine diphosphate galactose (UDP-Gal) and GlcNAc, with the formation of *N*-acetyllactosamine (LacNAc). However, the presence of α -lactalbumin changes the affinity of B4GALT1 from GlcNAc to glucose, leading to the synthesis of lactose (Brew *et al.*, 1968). Bovine α -lactalbumin contains two *N*-glycosylation motifs at Asn45 (Asn-Asp-Ser) and Asn74 (Asn-Ile-Ser), and a third non-canonical motif at Asn71 (Asn-Ile-Cys). Partial occupancy of Asn45 has been shown in previous studies, primarily with complex diantennary *N*-glycans containing *N,N'*-diacetyllactosamine (LacdiNAc) antennae, with on average 10% of total α -lactalbumin occurring as a glycoprotein (Barman, 1970, Chandrika, 1999, Hindle and Wheelock, 1970, Hopper and McKenzie, 1973, Slangen and Visser, 1999, Valk-Weeber *et al.*, 2020). Here, we present a detailed characterisation of α -lactalbumin glycosylation, including information on glycan composition, structure and localization, which are prerequisite to understanding the biological functionality of the glycoprotein.

Results

Bovine colostrum and milk *N*-glycoproteome

To study the characteristics and lactational changes of the whole bovine milk *N*-glycoproteome, we worked with the sample set described in **Chapter 3**, performing substantial orthogonal data analyses. The work presented in **Chapter 3** was protein-specific and focused on IgG purified from colostrum, transitional and mature milk. In contrast, in the current chapter, we explored the full *N*-glycoproteome of these samples, without prior enrichment of the glycoproteins or glycopeptides. For this purpose, we analysed 36 individual bovine milk samples collected from 4 individual Dutch Friesian-Holstein cows referred to as Cow 1, Cow 2, Cow 3 and Cow 4 at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28.0 d after calving. To illustrate the changes in the *N*-glycoproteome, we depict in Figure 4.1 the general characteristics of the bovine milk *N*-glycome at the earliest and latest time points of our monitoring, *i.e.*, in the 0.5-d colostrum and 28.0-d mature milk, summed over the four cows. The majority of the identified glycan compositions corresponded to diantennary complex structures, neutral or sialylated, with or without core fucosylation. Of the

neutral glycans, the greater part was represented by high-mannose glycans, followed to a lesser extent by complex glycans. A minor part of all glycan compositions detected corresponded to hybrid *N*-glycans. Higher glycan diversity can be observed in the 0.5-d colostrum than in the 28.0-d mature milk. The main notable difference between the *N*-glycome of 0.5-d colostrum and 28.0-d mature milk concerns the glycan species sialylated with NeuGc. As depicted in Figure 4.1, while NeuGc and NeuAc sialylation were identified in 21 and 46%, respectively, of the *N*-glycopeptide spectrum matches (gPSMs) in the colostrum, in the mature milk only NeuAc sialylation was still detectable (44% of the gPSMs). The glycan compositions from colostrum also included a considerable proportion of glycans simultaneously sialylated with both NeuAc and NeuGc, which were no longer detectable in the mature milk.

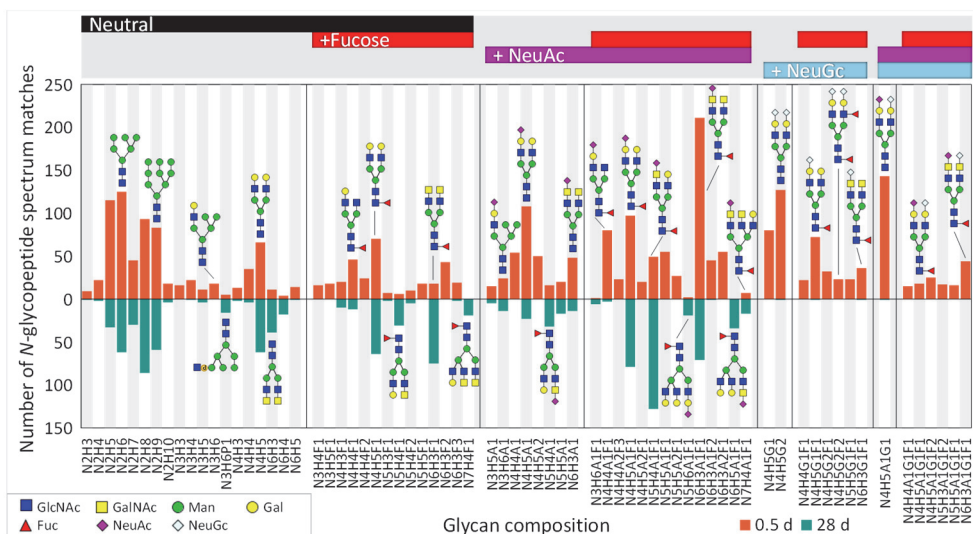


Figure 4.1 – Comparison of the overall characteristics of the bovine mammary secretion *N*-glycome between 0.5-d colostrum (orange, top) and 28.0-d mature milk (teal, bottom). The data represents absolute numbers (≥ 15) of *N*-glycopeptide spectrum matches summed across the four cows, i.e., Cow 1, Cow 2, Cow 3 and Cow 4. The glycan compositions are grouped based on their biochemical features, i.e., neutral (black), fucosylated (red), and sialylated with *N*-acetylneuraminic acid (NeuAc; magenta) and *N*-glycolylneuraminic acid (NeuGc; light blue), and combinations thereof. Proposed glycan structures were built using GlycoWorkBench 2.1 build 146, according to the symbol nomenclature for glycan representation of the Consortium for Functional Glycomics (Varki et al., 2009). The glycan structures were built for visualisation of the glycan composition; however, a glycan composition may be representative of several structural isoforms aside from the illustrated one. GlcNAc = *N*-acetylglucosamine; GalNAc = *N*-acetylgalactosamine; Man = mannose; Gal = galactose; Fuc = fucose.

Our experimental approach was focused on the analysis of the *N*-glycoproteome itself. This allowed us not only to extract information on the *N*-glycome of the different milk samples, but also to connect each identified *N*-glycan composition

and structure to the glycoprotein it occurred on, including localization of the glycan on the amino acid backbone of the protein. Details on the *N*-glycoproteome of the bovine mammary secretions analysed during lactation are shown in Figure 4.2. A total of 18 *N*-glycoproteins were identified, of which 12 and 4 were significantly more abundant in the proteomes of 0.5-d colostrum and 28.0-d mature milk, respectively, whereas the remaining 2 did not differ significantly in their abundances between colostrum and mature milk (Figure 3.2 from **Chapter 3**, Figure 4.2). The glycoproteins that were significantly more abundant in colostrum included the immunoglobulins and related proteins, *i.e.*, IgG, IgA and IgM heavy chain constant regions (IGHG1, IGHA and IGHM2, respectively), the immunoglobulin joining chain (JCHAIN) that facilitates the oligomerisation of IgA and IgM, the polymeric immunoglobulin receptor (PIGR) that is a precursor to the secretory component (SC) in the structure of secretory IgA and IgM (sIgA and sIgM), and lactoferrin (LTF). Next to this, (blood) serum-derived proteins included serotransferrin (TF), fetuin A (AHSG), zinc- α 2-glycoprotein (AZGP1), cellular proteins clusterin (CLU), lysosomal protein ribonuclease 2 (RNASE2), and soluble form of membrane protein cluster of differentiation 14 (CD14). Conversely, the glycoproteins with significantly higher abundance in the mature milk when compared to colostrum included α -lactalbumin (LALBA), lactophorin (GLYCAM1, also known as proteose peptone component 3, *i.e.*, PP3 in bovine milk), and milk fat globule membrane proteins butyrophilin subfamily 1 member A1 (BTN1A1) and pancreatic zymogen granule membrane glycoprotein 2 (GP2). The two proteins that did not show significant differences between their abundances in the proteomes of colostrum or mature milk were milk fat globule membrane proteins lactadherin (MFGE8) and platelet glycoprotein 4 (CD36, also known as the PAS-4 protein in bovine milk).

Relative abundances of the *N*-glycoproteins in the total proteome *per* lactation time point are shown in Figure 4.2A. The abundance of a given *N*-glycoprotein in the total proteome was calculated as the iBAQ value of the *N*-glycoprotein expressed as a percentage of total iBAQ values *per* lactation time point. Only the proteins on which *N*-glycosylation was detected in our study are illustrated in Figure 4.2; not depicted in Figure 4.2 are the proteins that made up the remaining proportion of the sample proteome, of which no *N*-glycopeptides were identified in this study. Insights on total proteome at the different lactation time points and *per* individual cow can be found in **Chapter 3**, Figure 3.2 and Supplementary Figures S3.2-S3.4. With respect to protein relative abundances in the total proteome of the sample, 0.5-d colostrum was dominated by IgG and α -lactalbumin (Figure 4.2A). With the gradual decrease in IgG abundance during lactation, α -lactalbumin became the

single most abundant *N*-glycoprotein in the 28.0-d mature milk, followed to a much lesser extent by lactophorin, with all other *N*-glycoproteins representing minor components of the proteome (Figure 4.2A). The greater part of *N*-glycopeptides in colostrum (0.5-3.0 d after calving) originated from the *N*-glycoproteins that were also significantly more abundant in the 0.5-d colostrum than in the 28.0-d mature milk (Figure 4.2B). In the transitional (7.0 and 14.0 d) and mature (28.0 d) milk samples, over half of the *N*-glycopeptides originated from *N*-glycoproteins significantly more abundant in the mature milk, primarily α -lactalbumin, and lactophorin and CD36. Furthermore, Figure 4.2B also illustrates a lactational decrease in the total *N*-glycopeptide spectral counts, relative to the same protein content, from colostrum to mature milk. Considering that spectral counts correlate with abundance and, therefore, concentration, it can be estimated that mature milk proteins contained comparatively less *N*-glycosylation than colostrum proteins. Figure 4.2C shows the relative contribution of each *N*-glycoprotein to the biochemical features of the identified glycan compositions. Corresponding to the decrease of sialylation, the proportion of neutral glycans increased during lactation. The neutral glycans of colostrum were primarily found on IgG and IgM glycopeptides, whereas a gradual increase in neutral glycans from α -lactalbumin and PIGR glycopeptides was observed during lactation. Fucosylated glycopeptides were found primarily from PIGR and IgG in colostrum, gradually being taken over by α -lactalbumin during lactation. The percentage of fucosylated glycopeptides from lactophorin increased during lactation and was higher in transitional and mature milk, than in colostrum. Overall, a 10% increase in the percentage of fucosylated glycans was observed during lactation. The total levels of NeuAc sialylation were comparable throughout lactation, although the origin of the NeuAc sialylated *N*-glycopeptides showed a lactational shift from primarily colostrum proteins to mature milk proteins. NeuGc sialylation, as also indicated in Figure 4.1, was found to decrease from a total of 21% in the 0.5-d colostrum to <1% in the 28.0-d mature milk. NeuGc sialylation was almost exclusively associated with proteins significantly more abundant in colostrum, particularly immunoglobulins IgG, IgA, IgM, including PIGR and JCHAIN, and blood serum proteins fetuin A and serotransferrin. A minor contribution of up to 1% of the total NeuGc sialylated *N*-glycopeptides was made by α -lactalbumin.

The average *N*-glycosylation site occupancies calculated for every identified *N*-glycoprotein are illustrated in Figure 4.3. While for the most glycoproteins full occupancy of the *N*-glycosylation sites was detected, α -lactalbumin and milk fat globule membrane proteins butyrophilin and lactadherin were found to consistently have partial occupancy in the range of 21-42%, with no clear lactational

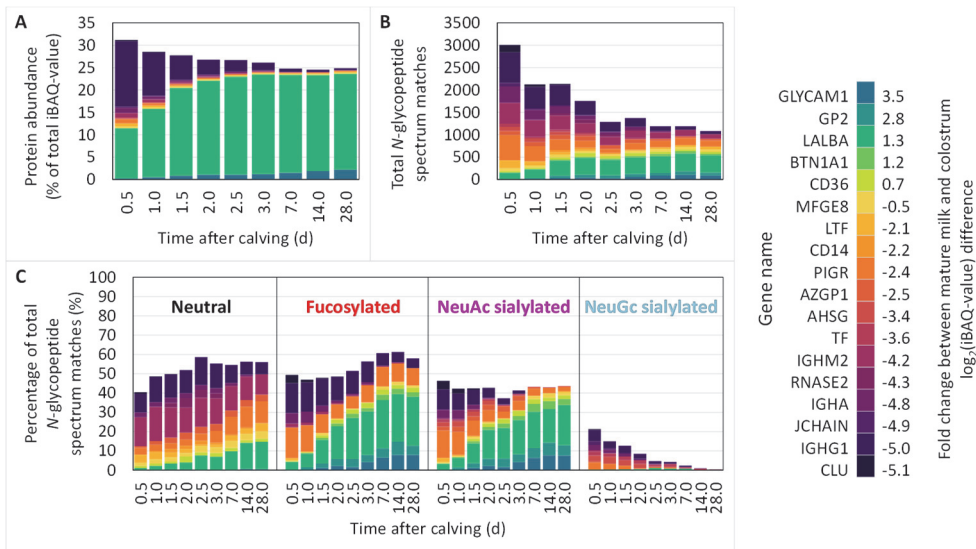


Figure 4.2 – Changes in the N-glycoproteome of bovine mammary secretions during lactation at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28.0 d after calving. All data is summed across the four individual cows and illustrates compositional changes in time normalised to total protein content. The identified N-glycoproteins are indicated by their gene names and ordered based on their fold change from the 0.5-d colostrum to the 28.0-d mature milk (differences in their intensity Based Absolute Quantification values, i.e., $\log_2(\text{iBAQ-value})$, between 28.0-d mature milk and 0.5-d colostrum). (A) Protein abundances calculated as the iBAQ-value of a given N-glycoprotein expressed as the percentage of total iBAQ-values per lactation time point. (B) Absolute numbers of N-glycopeptide spectrum matches illustrated per identified N-glycoprotein and lactation time point. (C) Biochemical features, i.e., neutral (black), fucosylated (red), and sialylated with N-acetylneuraminic acid (NeuAc; magenta) and N-glycolylneuraminic acid (NeuGc; light blue), of the N-glycans as a function of the glycoprotein they belong to, and during lactation. The results are calculated as numbers of N-glycopeptide spectrum matches of a given glycoprotein expressed as a percentage of total N-glycopeptide spectrum matches per time point.

dependency. CD14, ribonuclease 2 and IgA showed apparent lactational changes in N-glycosylation site occupancy. In the case of missing data, there was either no detection of the N-glycosylation site(s) of a given protein, or no detection of the protein altogether.

The glycoproteome, analysed directly from the milk samples without enrichment or purification of the proteins, was highly influenced by the changes in the relative abundance of the glycoproteins in the proteome. Supplementary Figure S4.1 illustrates the effect of IgG glycosylation analysed without (left panel) or with (right panel) first purifying the IgG from the colostrum and milk samples. During lactation, the IgG content gradually decreased from 51-73% of the total protein content in the 0.5-d colostrum to 2-3% of total protein content in the 28.0-d mature milk (Chapter 3, Supplementary Figure S3.13B). This was also reflected in the IgG

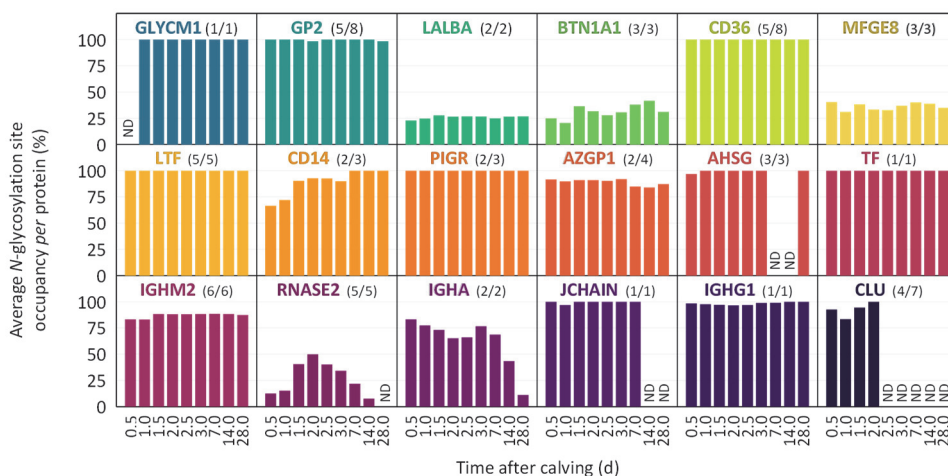


Figure 4.3 – Average N-glycosylation site occupancy per protein during lactation at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28.0 d after calving. These results are calculated based on spectral scans summed across the four individual cows. The average N-glycosylation site occupancy per protein is calculated by averaging the site occupancies of all detected N-glycosylation sites of a given protein. The N-glycoproteins are indicated by their gene names. The numbers in parentheses indicate (number of experimentally detected N-glycosylation motifs/total number of N-glycosylation motifs in the protein sequence). ND = not detected.

spectral counts of the putative N-glycopeptide detected without enrichment, shown in Supplementary Figure S4.1. The identification of IgG glycosylation was consequently influenced by the change in IgG abundance, introducing artefacts in the characterization of protein glycosylation during lactation. When IgG sialylation was analysed without purification of IgG from colostrum and milk (Supplementary Figure S4.1, left panel), the numbers of both NeuAc and NeuGc sialic acids *per* glycan apparently decreased. However, when IgG was first purified from the mammary secretion samples, followed by the analysis of its heavy chain constant region glycosylation at constant IgG content (Supplementary Figure S4.1, left panel), it was revealed that while the NeuAc sialylation decreased during lactation, the NeuGc sialylation was constant relative to the protein. In a similar way, Supplementary Figure S4.2 illustrates that the number of glycosylated peptide spectrum matches (gPSMs) is proportional to the total number of detections of ribonuclease 2, CD14 and IgA. The apparent lactational changes in average N-glycosylation site occupancy for these proteins observed in Figure 4.3 may therefore be caused by the changes in the abundance of the proteins in question, as shown in the case of IgG in Figure S4.1.

Lactational changes in the N-glycosylation of α -lactalbumin

As observed from Figure 4.2, α -lactalbumin was found to be one of the most

The N-glycoproteome from colostrum to mature milk: insights into α-lactalbumin

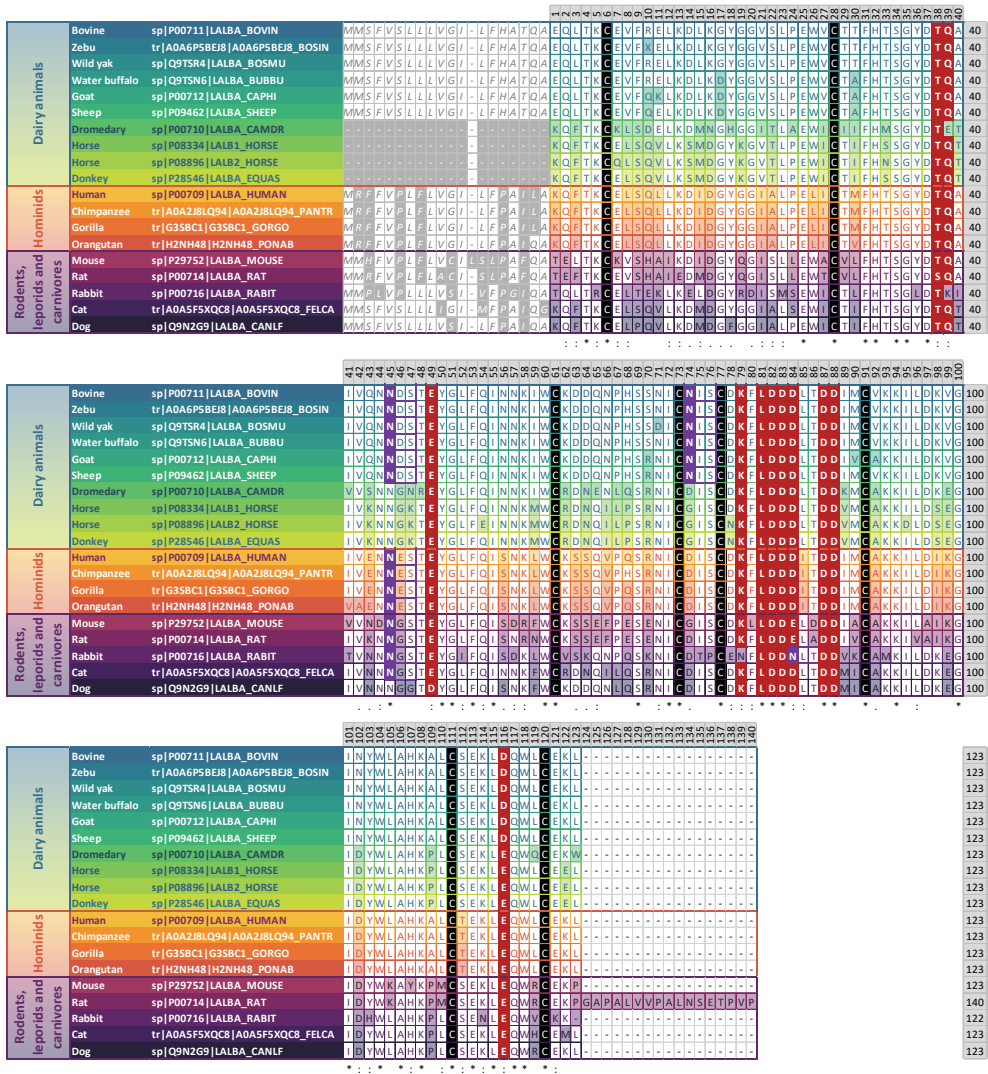


Figure 4.4 – Amino acid sequence alignment of α-lactalbumin across different species. The different protein sequences are identified by their UniProt header. The sequence alignment was performed on UniProt.org with CLUSTAL O(1.2.4). Numbering starts at the N-termini of the mature proteins, with the signal peptide sequences depicted in grey. Highlighted are the amino acid residues that differ from the bovine α-lactalbumin in the orthologous proteins. All cysteine residues are conserved between the different species, and are highlighted in black. Highlighted in red are the amino acid residues responsible for binding a single Ca²⁺ ion per protein; the Ca²⁺ ion-binding amino acid residues in the orthologous proteins were annotated based on sequence similarity to those of the human protein, as described in UniProt. The N-glycosylation motifs observed in the primary sequence are surrounded by thick purple borders, with the putatively N-glycosylated asparagine residue highlighted in purple.

abundant N-glycoproteins at every lactation time point, as well as a main component of the proteome during lactation (Figure 3.2B, Figure 3.4 and

supplementary Table S3.1 from **Chapter 3**). Furthermore, bovine milk α -lactalbumin is a protein of interest in humanised bovine milk-based infant formula, due to human α -lactalbumin being a major component of human milk, in which it accounts for 41% (*m/m*) of whey protein and 28% (*m/m*) of total protein (Heine *et al.*, 1991). Taking these into consideration, we further focused on the detailed lactational characterisation of α -lactalbumin *N*-glycosylation in the samples from the four individual cows.

The amino acid sequence of bovine α -lactalbumin contains two putative *N*-glycosylation motifs, at Asn45 and Asn74, as highlighted in Figure 4.4. Asn45 is conserved in most of the species depicted in Figure 4.4, including in humans. Asn74 is only present in cattle (bovine, zebu, wild yak and water buffalo) and in small ruminants (goat and sheep), but does not occur in any of the other species. Next to the conserved Asn45, rabbit α -lactalbumin contains an additional *N*-glycosylation motif at Asn84, which is not conserved in any of the other species. It can be further observed from Figure 4.4 that Asn45 occurs in different motifs between the different categories of species: Asn-Asp-Ser in dairy animals, Asn-Glu-Ser in hominids, and Asn-Gly-Ser in rodents, leporids and carnivores. Here, we found glycosylation occurring on both Asn45 and Asn74 sites in bovine α -lactalbumin, with the occupancy of each site during lactation in the individual cows shown in Figure 4.5. Both *N*-glycosylation sites exhibited partial occupancy, averaging at 35% for Asn45 and 4% for Asn74.

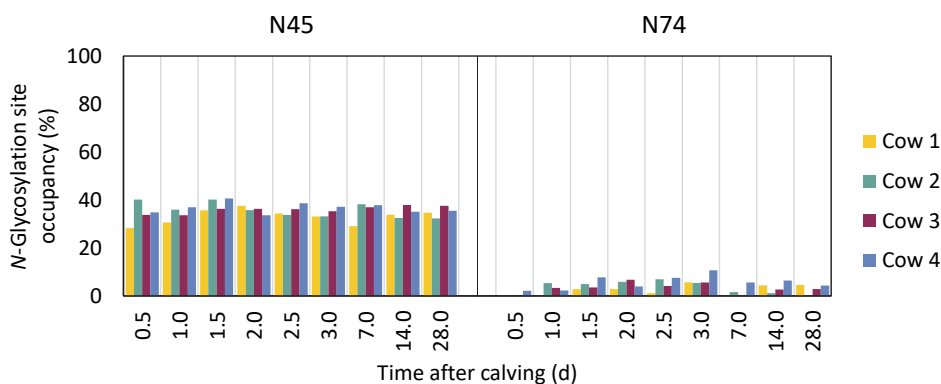


Figure 4.5 – *N*-Glycosylation site occupancy of Asn45 and Asn74 in the mature chain of bovine milk α -lactalbumin in samples collected at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28 d after calving from the four individual cows, i.e., Cow 1, Cow 2, Cow 3 and Cow 4. These results were determined based on scan numbers.

No lactational trend or differences between the individual cows were observed

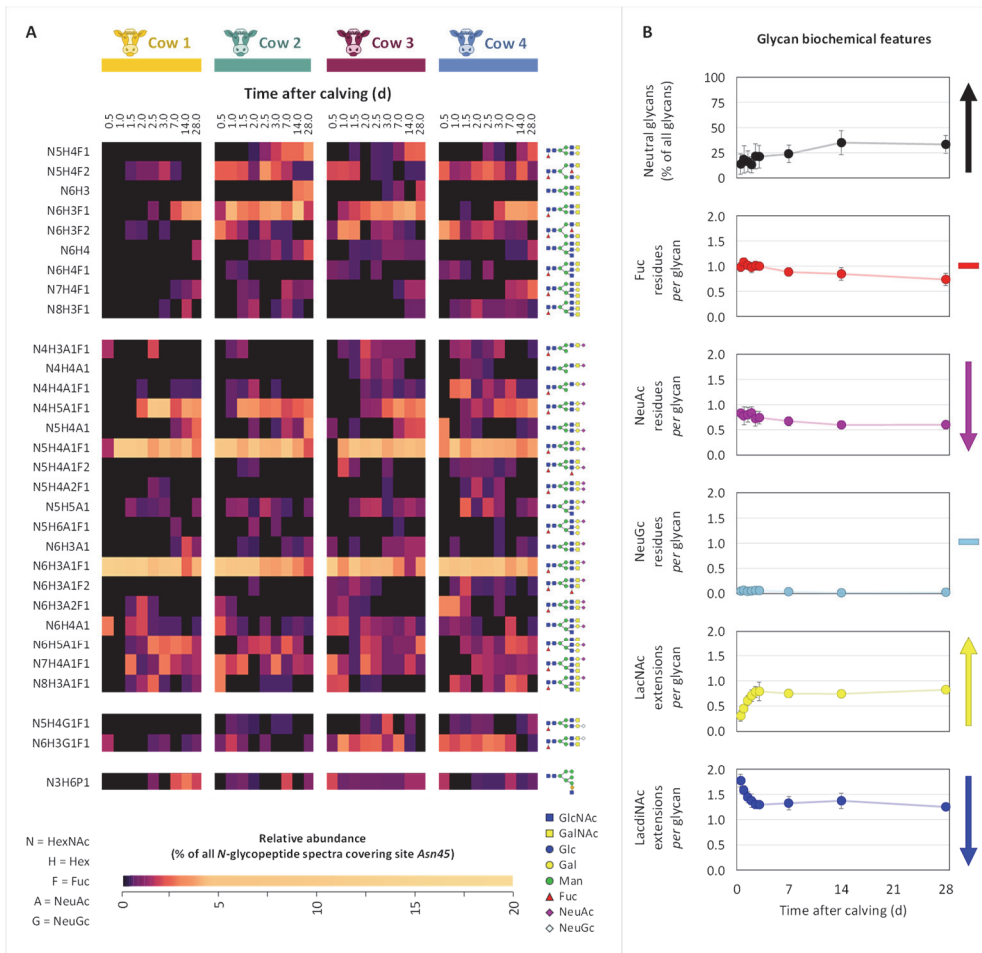


Figure 4.6 – Changes in the N-glycosylation profiles of bovine α -lactalbumin site Asn45 in the interval 0.5–28.0 d after calving in the bovine milk of the four cows, i.e., Cow 1, Cow 2, Cow 3 and Cow 4. (A) Heatmap depicting the macro- and micro-heterogeneity of the N-glycosylation determined based on spectral counts. Normalisation was performed relative to all spectra covering the glycosylation site, whereby the sum of all glycoforms and the non-glycosylated site amounts to 100% per time point. Clustering was on N-glycan biochemical features, from top to bottom: neutral, sialylated with NeuAc, sialylated with NeuGc, and phosphomannose. The glycan composition is indicated to the left of the heatmap and to the right are proposed corresponding glycan structures. (B) Lactational dynamics of the N-glycan biochemical features averaged across the four cows. From top to bottom, the panels describe changes in neutral glycans, fucosylation, sialylation with NeuAc or NeuGc, and quantification of LacNAc and LacdiNAc antennae, respectively. The error bars represent the standard deviation of values from the four cows. The increasing, decreasing and no change trends were determined based on Student's *t*-test probability determined between the values at 0.5 and 28.0 d at a significance level of 0.05. Abbreviations: Gal = galactose; Glc = glucose; Man = mannose; Hex = hexose; GalNAc = N-acetylgalactosamine; GlcNAc = N-acetylglucosamine; HexNAc = N-acetylhexosamine; Fuc = fucose; NeuAc = N-acetylneuraminic acid; NeuGc = N-glycolylneuraminic acid; LacNAc = N-acetyllactosamine; LacdiNAc = N,N'-diacetyllactosamine.

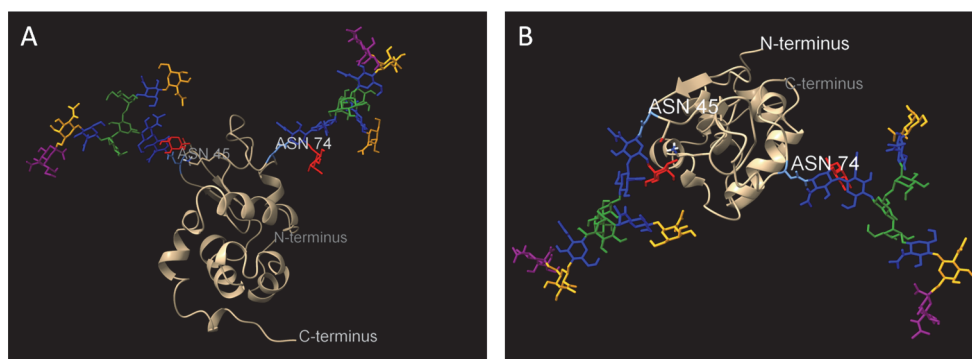


Figure 4.7 – Model of doubly-N-glycosylated bovine α -lactalbumin occupied with the N6H3A1F1 glycan at Asn45 and Asn74. The glycan structures were added to the 1F6S PDB structure of bovine α -lactalbumin using the Glycoprotein Builder tool of GLYCAM-Web (glycam.org). (A) Side and (B) top view of the glycosylated protein. Asn45 and Asn74 are highlighted in light blue. The monosaccharides in the glycan structures are coloured according to Varki et al. (2009): N-acetylglucosamine – blue, N-acetylgalactosamine – yellow, mannose – green, N-acetylneuraminic acid – purple, and fucose (F) – red.

with respect to site occupancy (Figure 4.5). A detailed overview of the *N*-glycosylation of the (higher occupancy) Asn45 site is shown in Figure 4.6A, *per* individual cow and lactation time point. A total of 30 different glycan compositions were identified, mostly corresponding to complex mono-, di- and triantennary glycan structures. The biochemical properties of these glycans are summarised in Figure 4.6B. These results indicate that the common features of α -lactalbumin glycans at site Asn45 are fucosylation, NeuAc sialylation and LacdiNAc antennae, all of which show lactational changes. These biochemical and structural glycan features were verified by manual curation of the *N*-glycopeptide fragmentation spectra, as exemplified in panel B of Supplementary Figure S4.3. Fucosylation and NeuAc sialylation decreased from 1.0 ± 0.1 (mean \pm standard deviation) and 0.8 ± 0.1 residues/glycan in the colostrum (0.5-3.0 d) samples, to 0.7 ± 0.1 and 0.8 ± 0.1 residues/glycan in the 28.0-d mature milk, respectively. NeuGc sialylation was detected, albeit at very low levels of up to 0.1 residues/glycan across all analysed samples. The number of LacdiNAc antennae decreased on average from 1.8 ± 0.1 antennae/glycan in the 0.5-d colostrum to 1.3 ± 0.0 antennae/glycan in the 28.0-d mature milk, simultaneously with the increase in the number of LacNAc antennae from 0.3 ± 0.1 to 0.8 ± 0.1 antennae/glycan. These structural features of the glycans were manually verified at glycopeptide fragmentation spectrum level. An example of the fragmentation of a LacdiNAc-containing α -lactalbumin glycan is illustrated in Figure S4.3 comparatively to a bisecting GlcNAc-containing IgG glycan of same monosaccharide composition. Furthermore, due to the majority of the α -lactalbumin glycans bearing both fucosylation and NeuAc sialylation, and the small

mass difference between 1 NeuAc residue (291.0954 Da) and 2 fucose residues (292.1158 Da), the correctness of NeuAc and Fuc residue identification and assignment by Byonic v4.5.2 was also manually verified in the annotated spectra. The retention times of *N*-glycopeptides in reversed-phase chromatography were found to cluster based on the charge of the glycans, with increasing retention times in the order: neutral glycans < glycans sialylated with a single sialic acid residue (irrespective of whether NeuAc or NeuGc) < phosphorylated glycans < glycans sialylated with two sialic acid residues (Figure S4.4). Non-glycosylated peptides with the same amino acid sequence were found to have higher retention times than all *N*-glycosylated peptides (Figure S4.4). These features were also useful in verifying and correcting the identified glycan compositions.

One of the most abundantly-detected glycans of α -lactalbumin, *i.e.*, N6H3A1F1 was modelled into the structure of a doubly-*N*-glycosylated α -lactalbumin occupied at both Asn45 and Asn74. The structure of the glycoprotein is depicted in Figure 4.7.

Discussion

In this study, we monitored the lactational changes in bovine milk *N*-glycoproteome analysed directly from the mammary gland secretions in the range of colostrum (0.5 d after calving) to mature milk (28 d). Nevertheless, both *N*- and *O*-linked glycosylation are present in bovine mammary secretions. Unlike *N*-glycosylation, *O*-glycosylation does not follow a consensus amino acid sequence motif, and can theoretically occur on any Ser or Thr residue, which are also constituent amino acid residues of the *N*-glycosylation motif. For these reasons, the study of *O*-glycosylation requires different database search strategies and poses additional challenges for identifying and localizing *O*-glycans on glycopeptides. We therefore chose to focus our study only on the *N*-glycoproteome. While we observed protein-specific lactational changes in *N*-glycosylation, the major changes in the *N*-glycoproteome were driven by the lactational shift in protein composition from colostrum to mature milk. In the second part of this study, we performed a more detailed analysis of the *N*-glycosylation of α -lactalbumin, a major milk protein, a major contributor to the *N*-glycoproteome, and a protein of interest in the infant formula industry.

Bovine colostrum and milk *N*-glycoproteome

Takimori *et al.* (2011) investigated changes in bovine milk glycoproteins in the similar time frame of 1-28 d postpartum, albeit at the glycan level. In their study, the *N*-glycome of colostrum was found to substantially differ from that of transitional and mature milk. The ratio of NeuGc/NeuAc was highest in the 1d colostrum sample, and gradually decreased in time. Colostrum *N*-glycans were also

more highly sialylated than those in transitional and mature milk. The results of our study further substantiate the glycomics findings of Takimori *et al.* (2011), identifying NeuGc as a biochemical feature of primarily blood serum-derived proteins present in the colostrum (Figure 4.1, Figure 4.2C). The decline of blood serum-derived proteins during the transition from colostrum to mature milk resulted in the consequent decrease of NeuGc, and thereby also total sialylation in the milk *N*-glycoprofile (Figure 4.2). Takimori *et al.* (2011) attributed the lactational changes in *N*-glycome primarily to the qualitative and quantitative changes in IgG from colostrum to mature milk. While we have shown that IgG is the main protein of colostrum (Gazi *et al.*, 2023), and a main contributor to the colostrum *N*-glycoprofile (Figure 4.2), the measurable lactational changes in *N*-glycosylation were driven by a series of several glycoproteins whose relative abundances changed during lactation. (Valk-Weeber *et al.*, 2020) estimated the relative contribution of individual glycoproteins, recreating the *N*-glycoprofile of (mature bovine milk) acid whey based on lactophorin, α -lactalbumin, lactoferrin, IgG and lactoperoxidase. Except for lactoperoxidase, of which we detected no *N*-glycopeptides in our study, we also identified the other four proteins as main contributors to the bovine milk *N*-glycoproteome Figure 4.2.

The *N*-glycosylation of a given protein is directly influenced by the available glycosylation machinery of the cell producing the protein. This can be observed, for instance, in differences between the *N*-glycosylation of wild type and recombinant proteins expressed in different host cells (Croset *et al.*, 2012, Lin *et al.*, 2018, Yu *et al.*, 2010). With respect to our results, the differences in *N*-glycosylation and protein composition between colostrum and mature milk (Figure 4.1, Figure 4.2) suggests that colostrum and mature milk proteins have different origins and are produced by different cells.

Overall, our study found that protein *N*-glycosylation was much more diverse in colostrum than in mature milk, and that the *N*-glycoprofile of the samples was to a greater extent influenced by the lactational changes in protein composition, and to a lesser extent by the protein-specific changes in *N*-glycosylation. While NeuAc was previously found to be a specific feature of bovine colostrum IgG *N*-glycosylation (Gazi *et al.*, 2023), in this chapter we identified NeuGc as the signature feature of bovine colostrum *N*-glycosylation in general.

Lactational changes in the *N*-glycosylation of α -lactalbumin

α -Lactalbumin was further selected for an in-depth study due to it being abundantly present throughout lactation, as well as being a protein of interest for the production of humanised infant formula. While the putative *N*-glycosylation site at

Asn45 is conserved in the α -lactalbumin orthologues between many different species, considerable inter-species differences in site occupancy have been reported in literature, ranging from no occupancy *via* partial occupancy to full occupancy. Bovine α -lactalbumin has often been described as being partially glycosylated, with up to 10% (*m/m*) of the protein harbouring *N*-glycosylation (Barman, 1970, Chandrika, 1999, Hopper and McKenzie, 1973, Slangen and Visser, 1999, Valk-Weeber *et al.*, 2020), whereas here we found on average 35% site occupancy of Asn45 based on spectral counts Figure 4.3. The α -lactalbumin of other dairy animals, such as goats and water buffalos, has also been described to be partially glycosylated to a similar extent (Chianese *et al.*, 2004, D'Ambrosio *et al.*, 2008, Kim and Jimenez-Flores, 1994, Macgillivray *et al.*, 1979). While <1% of the human α -lactalbumin has been reported to harbour *N*-glycosylation at Asn71 found in the less common Asn-Xxx-Cys motif (Giuffrida *et al.*, 1997), no *N*-glycosylation of the human α -lactalbumin at Asn45 has been reported to date. Asn71 is conserved in the α -lactalbumin of most species, and has also been reported to be *N*-glycosylated at 1% occupancy in horse α -lactalbumin (Girardet *et al.*, 2004), which lacks the common Asn-Xxx-Ser/Thr *N*-glycosylation motifs in its amino acid sequence (Figure 4.4). In our study, the second bovine α -lactalbumin *N*-glycosylation site, with on average 4% site occupancy based on spectral counts (Figure 4.3), is found on peptides that harbour both Asn71 and Asn74. While the *N*-glycosylation site is believed to be Asn74 due to this residue being present in the consensus *N*-glycosylation motif (Asn-Xxx-Ser), localization at Asn71 cannot fully be excluded. Conversely, rat and rabbit α -lactalbumin have been reported to primarily occur as glycoproteins (Hopp and Woods, 1979, Prasad *et al.*, 1979, Prasad *et al.*, 1982). One of the possible explanations for these inter-species differences in Asn45 site occupancy has been attributed to the amino acid residue at position 46 (Hopp and Woods, 1979). The occurrence of an Asp (dairy animals, Figure 4.4) or Glu residue at position 46 (humans and other hominids, Figure 4.4) has been hypothesised to be less favourable for the glycosylation of Asn45, than the occurrence of Gly46 (rodents, leporids and carnivores, Figure 4.4). Previous studies investigating the effect of the Xxx amino acid residue in the Asn-Xxx-Ser/Thr motif on the *N*-glycosylation efficiency have found that the occurrence of Glu, and to a lesser extent Asp, correlated with decreased glycosylation efficiency when the hydroxy amino acid was Ser, with no detrimental effect in the case of Thr (Bause and Legler, 1981, Breuer *et al.*, 2001, Kasturi *et al.*, 1997, Malaby and Kobertz, 2014, Mononen and Karjalainen, 1984, Rao and Bernd, 2010, Shakin-Eshleman *et al.*, 1996). The same studies found that when Xxx was a small amino acid such as Gly, it correlated with high glycosylation efficiency. This is in line with our findings on the *N*-glycosylation of bovine α -lactalbumin Asn45, and the previously-mentioned

studies, where Asn45 followed by Glu was not glycosylated, followed by Asp was partially glycosylated, and followed by Gly was fully glycosylated. Nevertheless, other studies have found conflicting information, with feline and murine α -lactalbumin being described as consisting of 50% (Halliday *et al.*, 1990) and 10% glycoprotein (Nagamatsu and Oka, 1980), respectively, with the remaining percentage being made up by non-glycosylated α -lactalbumin, while Asn45 in both of these species is found in the Asn-Gly-Ser motif (Figure 4.4).

The mechanism behind the partial site occupancy of Asn45 in bovine α -lactalbumin is therefore not fully understood, and may be the result of decreased glycosyltransferase affinity towards the Asn-Asp-Ser motif. The low site occupancy of Asn74, or potentially Asn71, may be explained by the proximity of both of these sites to cysteine residues which are typically engaged in disulphide bridges. Asn74 is located between Cys73 and Cys77, while Asn71 is in a noncanonical motif containing Cys73, whereas (Figure 4.4). Non-glycosylated α -lactalbumin contains four disulphide bridges arranged in a structurally identical way to those in its homologue lysozyme: Cys6-Cys120, Cys28-Cys111, Cys61-Cys77 and Cys73-Cys91 (Vanaman *et al.*, 1970). *N*-Glycosylation of Asn74 or Asn71 may therefore interfere or compete with the formation of two or one of these disulphide bridges, respectively, explaining the low *N*-glycosylation occupancy observed in this region of the protein (Figure 4.3). Furthermore, Asn71 and Asn74 are localized in close proximity of the Lys79-Asp88 amino acid sequence region, which is engaged in the binding of one Ca^{2+} ion. In its holo form, α -lactalbumin has a more compact structure (Chrysin *et al.*, 2000), whereby Asn71 and Asn74 may be less surface available for glycosyltransferases.

Slangen and Visser (1999) analysed intact mass of glycosylated α -lactalbumin and of its enzymatically-released *N*-glycans by LC-MS. Based on the experimental masses, they proposed 14 monosaccharide compositions corresponding to complex di- and triantennary glycan structures with LacNAc and LacdiNAc antennae, with or without monofucosylation and/or monosialylation with NeuAc. Based on the deconvoluted mass spectrum of the intact glycosylated α -lactalbumin, Slangen and Visser (1999) found that the three most abundant glycoproteoforms harboured glycan compositions N5H4F1, N6H3F1 and N6H3A1F1, respectively. In a similar study, Chandrika (1999) proposed a total of 15 *N*-glycan compositions and structures for bovine α -lactalbumin based on the masses of released intact glycans and intact glycoproteins. They reported N6H3F1 and N6H3A1F1 as the main glycan compositions on bovine α -lactalbumin without and with sialic acid, respectively. In a more recent study, Valk-Weeber *et al.* (2020) analysed the *N*-glycosylation of whey proteins, combining the intact mass analysis of released *N*-glycans with

sequential exoglycosidase digestion of the glycans for composition and structure analysis. They identified 14 bovine α -lactalbumin *N*-glycan compositions and structures consisting of complex, primarily diantennary, and to a lesser extent, triantennary glycans, primarily having LacdiNAc, but also LacNAc antennae, with or without core fucosylation and/or monosialylation with NeuAc. In line with Slangen and Visser (1999) and Chandrika (1999), Valk-Weeber *et al.* (2020) also identified the N6H3F1 and N6H3A1F1 as two of the most abundant ones in glycosylated bovine α -lactalbumin. Our study expands the existing knowledge of bovine α -lactalbumin *N*-glycosylation with the localization-specific identification of a broader repertoire of *N*-glycans, including difucosylated and disialylated species, as well as glycans sialylated with NeuGc (Figure 4.6), with all identifications manually curated at fragmentation spectrum level. Furthermore, while *N*-glycosylation site occupancy did not change during lactation (Figure 4.5), our study shows the changes in *N*-glycosylation microheterogeneity of α -lactalbumin during the transition from colostrum to mature milk in four individual cows (Figure 4.6), in a similar way to the previously-described changes in *N*-glycosylation of bovine milk IgG (Gazi *et al.*, 2023). Nevertheless, the lactational changes in the glycan biochemical properties were not as pronounced for α -lactalbumin as in the case of IgG glycosylation.

The function of α -lactalbumin *N*-glycosylation is still not fully understood. Previous studies investigating the differences in lactose synthase activity between non-glycosylated and *N*-glycosylated α -lactalbumin found that both fractions exhibited equal activities (Barman, 1970, Proctor *et al.*, 1974).

In conclusion, owing to advances in mass spectrometry and glycoproteomics technology, we were able to consolidate and expand the existing knowledge of the *N*-glycome and *N*-glycoproteome changes from bovine colostrum to mature milk. Future studies need to further elucidate the biological functionality of the here described glycoproteome and glycoproteins.

Materials and Methods

In this chapter, we re-analysed the peptide-centric mass spectrometry data acquired for the whole bovine mammary gland secretions described in **Chapter 3**. Chemicals and reagents, bovine milk samples, determination of protein content, and analysis of full proteome by peptide-centric mass spectrometry are described in full detail in the Materials and Methods section of **Chapter 3**. Briefly, a set of 36 samples from four individual cows, referred to as Cow 1, Cow 2, Cow 3 and Cow 4, and sampled at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28.0 d after calving, were analysed by LC-MS/MS in triplicate, following proteolytic digestion. Hybrid oxonium

ion-triggered fragmentation methods were employed, dedicated for the analysis of glycopeptides (Reiding *et al.*, 2018).

In addition to the work described in **Chapter 3**, here we performed database search and identification of *N*-glycopeptides, and analysis of the full glycoproteome detectable within this experimental approach.

Database search and identification of bovine milk protein *N*-glycosylation

Based on the results of the bovine proteome identification (Figure 3.2, supplementary Figure S3.4, and supplementary Table S3.1 from **Chapter 3**), a small protein database containing only the identified sequences was created for the *N*-glycoprotein search. The database search for *N*-glycoproteins was performed with Byonic v4.5.2 (Protein Metrics, Cupertino, California, USA), a search engine specialised in the identification of glycopeptides. A decoy database was created by reversing the protein sequences of the target database. The *N*-glycan database used was the one containing 2440 compositions, including high mannose, and hybrid and complex glycan antennae of *N*-acetylglucosamine (GlcNAc), *N*, *N'*-diacetylglucosamine (GlcNAc₂) and all combinations thereof, as well as sialylation with *N*-acetylneuraminic acid (NeuAc) and the non-human *N*-glycolylneuraminic acid (NeuGc), as well as all combinations thereof, as described in **Chapter 3**. Proteolytic cleavage sites were defined C-terminal of arginine, lysine, and aspartic and glutamic acid residues with 3 missed cleavages, but digestion specificity was set to semi-specific. Fragmentation type was set either to HCD for the samples analysed with HCD or product ion-triggered stepping HCD, or to both HCD & ETHcD for the samples analysed with product ion-triggered ETHcD. Cysteine carbamidomethylation was set as a fixed modification. Methionine and tryptophan oxidation, N-terminal cyclisation of glutamine and glutamic acid to pyroglutamic acid, and serine, threonine and tyrosine phosphorylation were all searched as rare variable modifications, whereas *N*-glycosylation was searched as a common variable modification. A maximum of one rare and one common variable modification were allowed *per* peptide. Further post-processing included filtering of the data based on $|\text{Log Prob}| \geq 1$ and $\text{score} \geq 150$. Based on the identified glycan composition, proposed glycan structures were built using GlycoWorkBench 2.1 build 146, according to the symbol nomenclature for glycan representation of the Consortium for Functional Glycomics (Varki *et al.*, 2009).

References

Barman T. 1970. Purification and properties of bovine milk glyco- α -lactalbumin. *Biochim Biophys Acta*, 214:242-244.

Bause E, Legler G. 1981. The role of the hydroxy amino acid in the triplet sequence Asn-Xaa-Thr (Ser) for the N-glycosylation step during glycoprotein biosynthesis. *Biochemical Journal*, 195:639-644.

Breuer W, Klein RA, Hardt B, Bartoschek A, Bause E. 2001. Oligosaccharyltransferase is highly specific for the hydroxy amino acid in Asn-Xaa-Thr/Ser. *FEBS Lett*, 501:106-110.

Brew K. 2013. α -Lactalbumin. In: McSweeney PLH, Fox PF editors. *Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition*. Boston, MA: Springer US. p. 261-273.

Brew K, Vanaman TC, Hill RL. 1968. The role of alpha-lactalbumin and the A protein in lactose synthetase: a unique mechanism for the control of a biological reaction. *Proceedings of the National Academy of Sciences*, 59:491-497.

Brodbeck U, Denton W, Tanahashi N, Ebner KE. 1967. The isolation and identification of the B protein of lactose synthetase as α -lactalbumin. *J Biol Chem*, 242:1391-1397.

Cao X, Yang M, Yang N, Liang X, Tao D, Liu B, Wu J, Yue X. 2019. Characterization and comparison of whey N-glycoproteomes from human and bovine colostrum and mature milk. *Food Chem*, 276:266-273.

Chandrika UG. 1999. Glycosylation of bovine α -lactalbumin: a thesis presented in partial fulfilment of the requirements for the degree of Master of Philosophy in Biochemistry, at Massey University, New Zealand. Massey University.

Chianese L, Caira S, Lilla S, Pizzolongo F, Ferranti P, Pugliano G, Addeo F. 2004. Primary structure of water buffalo α -lactalbumin variants A and B. *J Dairy Res*, 71:14-19.

Chrysina ED, Brew K, Acharya KR. 2000. Crystal Structures of Apo- and Holo-bovine α -Lactalbumin at 2.2-Å Resolution Reveal an Effect of Calcium on Inter-lobe Interactions*. *J Biol Chem*, 275:37021-37029.

Croset A, Delafosse L, Gaudry J-P, Arod C, Glez L, Losberger C, Begue D, Krstanovic A, Robert F, Vilbois F, *et al.* 2012. Differences in the glycosylation of recombinant proteins expressed in HEK and CHO cells. *J Biotechnol*, 161:336-348.

D'Ambrosio C, Arena S, Salzano AM, Renzone G, Ledda L, Scaloni A. 2008. A proteomic characterization of water buffalo milk fractions describing PTM of

major species and the identification of minor components involved in nutrient delivery and defense against pathogens. *Proteomics*, 8:3657-3666.

Gazi I, Reiding KR, Groeneveld A, Bastiaans J, Huppertz T, Heck AJ. 2023. Key changes in bovine milk immunoglobulin G during lactation: NeuAc sialylation is a hallmark of colostrum immunoglobulin G N-glycosylation. *Glycobiology*, 33:115-125.

Girardet J-M, N'negue M-A, Egito A, Campagna S, Lagrange A, Gaillard J-L. 2004. Multiple forms of equine α -lactalbumin: evidence for N-glycosylated and deamidated forms. *Int Dairy J*, 14:207-217.

Giuffrida MG, Cavaletto M, Giunta C, Neuteboom B, Cantisani A, Napolitano L, Calderone V, Godovac-Zimmermann J, Conti A. 1997. The Unusual Amino Acid Triplet Asn-Ile-Cys Is a Glycosylation Consensus Site in Human α -Lactalbumin. *J Protein Chem*, 16:747-753.

Gopal PK, Gill HS. 2000. Oligosaccharides and glycoconjugates in bovine milk and colostrum. *Br J Nutr*, 84:69-74.

Halliday JA, Bell K, McKenzie HA, Shaw DC. 1990. Feline whey proteins: Identification, isolation and initial characterization of α -lactalbumin, β -lactoglobulin and lysozyme. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 95:773-779.

Heine WE, Klein PD, Reeds PJ. 1991. The importance of α -lactalbumin in infant nutrition. *J Nutr*, 121:277-283.

Hindle E, Wheelock J. 1970. Carbohydrates of bovine alpha-lactalbumin preparations. *Biochemical Journal*, 119:14P.

Hopp TP, Woods KR. 1979. Primary structure of rabbit. alpha.-lactalbumin. *Biochemistry*, 18:5182-5191.

Hopper K, McKenzie H. 1973. Minor components of bovine α -lactalbumin A and B. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 295:352-363.

Kasturi L, Chen H, Shakin-Eshleman SH. 1997. Regulation of N-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors. *Biochemical Journal*, 323:415-419.

Kim HHY, Jimenez-Flores R. 1994. Comparison of Milk Proteins Using Preparative Isoelectric Focusing Followed by Polyacrylamide Gel Electrophoresis. *J Dairy Sci*, 77:2177-2190.

Krištić J, Lauc G. 2017. Ubiquitous Importance of Protein Glycosylation. In: Lauc G, Wuhler M editors. *High-Throughput Glycomics and Glycoproteomics: Methods and Protocols*. New York, NY: Springer New York. p. 1-12.

Layman DK, Lönnerdal B, Fernstrom JD. 2018. Applications for α -lactalbumin in human nutrition. *Nutr Rev*, 76:444-460.

Lin Y-H, Franc V, Heck AJR. 2018. Similar Albeit Not the Same: In-Depth Analysis of Proteoforms of Human Serum, Bovine Serum, and Recombinant Human Fetuin. *J Proteome Res*, 17:2861-2869.

Macgillivray RT, Brew K, Barnes K. 1979. The amino acid sequence of goat α -lactalbumin. *Archives of Biochemistry and Biophysics*, 197:404-414.

Malaby HL, Kobertz WR. 2014. The middle X residue influences cotranslational N-glycosylation consensus site skipping. *Biochemistry*, 53:4884-4893.

McGrath BA, Fox PF, McSweeney PL, Kelly AL. 2016. Composition and properties of bovine colostrum: a review. *Dairy Sci Technol*, 96:133-158.

Mckenzie HA, White Jr FH. 1991. Lysozyme and α -lactalbumin: structure, function, and interrelationships. *Adv Protein Chem*, 41:173-315.

Mononen I, Karjalainen E. 1984. Structural comparison of protein sequences around potential N-glycosylation sites. *Biochim Biophys Acta*, 788:364-367.

Nagamatsu Y, Oka T. 1980. Purification and characterization of mouse α -lactalbumin and preparation of its antibody. *Biochemical Journal*, 185:227-237.

O'Riordan N, Kane M, Joshi L, Hickey RM. 2014. Structural and functional characteristics of bovine milk protein glycosylation. *Glycobiology*, 24:220-236.

Prasad R, Hudson BG, Butkowski R, Hamilton JW, Ebner KE. 1979. Resolution of the charge forms and amino acid sequence and location of a tryptic glycopeptide in rat α -lactalbumin. *J Biol Chem*, 254:10607-10614.

Prasad RV, Butkowski RJ, Hamilton JW, Ebner KE. 1982. Amino acid sequence of rat α -lactalbumin: a unique α -lactalbumin. *Biochemistry*, 21:1479-1482.

- Proctor SD, Wheelock JV, Davies DT. 1974. Heterogeneity of bovine α -lactalbumin. *Biochem Soc Trans*, 2:621-622.
- Qasba PK, Kumar S, Brew K. 1997. Molecular divergence of lysozymes and α -lactalbumin. *Critical reviews in biochemistry and molecular biology*, 32:255-306.
- Rao RSP, Bernd W. 2010. Do N-glycoproteins have preference for specific sequons? *Bioinformation*, 5:208.
- Reiding KR, Bondt A, Franc V, Heck AJ. 2018. The benefits of hybrid fragmentation methods for glycoproteomics. *TrAC, Trends Anal Chem*, 108:260-268.
- Shakin-Eshleman SH, Spitalnik SL, Kasturi L. 1996. The Amino Acid at the X Position of an Asn-X-Ser Sequon Is an Important Determinant of N-Linked Core-glycosylation Efficiency (*). *J Biol Chem*, 271:6363-6366.
- Slangen CJ, Visser S. 1999. Use of mass spectrometry to rapidly characterize the heterogeneity of bovine α -lactalbumin. *J Agric Food Chem*, 47:4549-4556.
- Takimori S, Shimaoka H, Furukawa JI, Yamashita T, Amano M, Fujitani N, Takegawa Y, Hammarström L, Kacs Kovics I, Shinohara Y. 2011. Alteration of the N-glycome of bovine milk glycoproteins during early lactation. *The FEBS journal*, 278:3769-3781.
- Valk-Weeber RL, Deelman-Driessen C, Dijkhuizen L, Eshuis-de Ruiter T, van Leeuwen SS. 2020. In depth analysis of the contribution of specific glycoproteins to the overall bovine whey N-linked glycoprofile. *J Agric Food Chem*, 68:6544-6553.
- Vanaman TC, Brew K, Hill RL. 1970. The disulfide bonds of bovine α -lactalbumin. *J Biol Chem*, 245:4583-4590.
- Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Marth JD, Bertozzi CR, Hart GW, Etzler ME. 2009. Symbol nomenclature for glycan representation. *Proteomics*, 9:5398-5399.
- Xiao J, Wang J, Gan R, Wu D, Xu Y, Peng L, Geng F. 2022. Quantitative N-glycoproteome analysis of bovine milk and yogurt. *Current Research in Food Science*, 5:182-190.
- Yu T, Guo C, Wang J, Hao P, Sui S, Chen X, Zhang R, Wang P, Yu G, Zhang L, et al. 2010. Comprehensive characterization of the site-specific N-glycosylation of wild-

type and recombinant human lactoferrin expressed in the milk of transgenic cloned cattle. *Glycobiology*, 21:206-224.

Zhu J, Lin Y-H, Dingess KA, Mank M, Stahl B, Heck AJ. 2020. Quantitative longitudinal inventory of the N-glycoproteome of human milk from a single donor reveals the highly variable repertoire and dynamic site-specific changes. *J Proteome Res*, 19:1941-1952.

Overview of supplementary information for this chapter

Supplementary figures

Supplementary figures

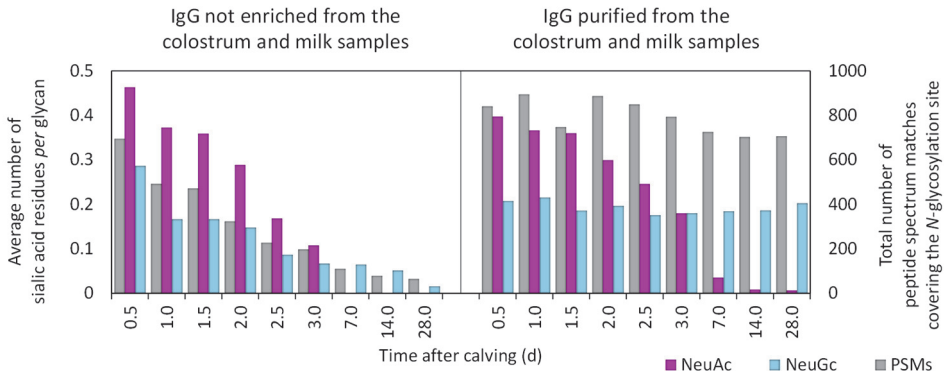


Figure S4.1 – Quantification of N-acetylneuraminic (NeuAc; magenta) and N-glycolylneuraminic (NeuGc; light blue) sialic acids, in the N-glycans identified in bovine milk IgG during lactation at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28.0 d after calving. These results are based on spectral counts that were summed across the four individual cows. Indicated on the secondary y-axis and plotted in grey are the total peptide spectrum matches (PSMs). The panel on the left shows the results obtained on bovine milk IgG analysed without enrichment from the colostrum and milk samples. The panel on the right shows the results from the analyses of IgG purified from the samples mentioned above, as described in **Chapter 3**.

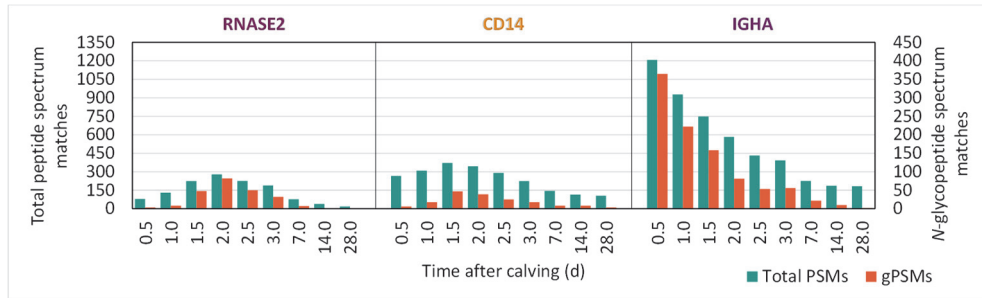


Figure S4.2 – Lactational changes in total peptide spectrum matches (PSMs; primary y-axis; teal) and N-glycopeptide spectrum matches (gPSMs; secondary y-axis; orange) of ribonuclease 2 (RNASE2), cluster of differentiation 14 (CD14) and IgA heavy chain constant (IGHA) at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28.0 d after calving. These results are summed across the four individual cows.

4

The N-glycoproteome from colostrum to mature milk: insights into α -lactalbumin

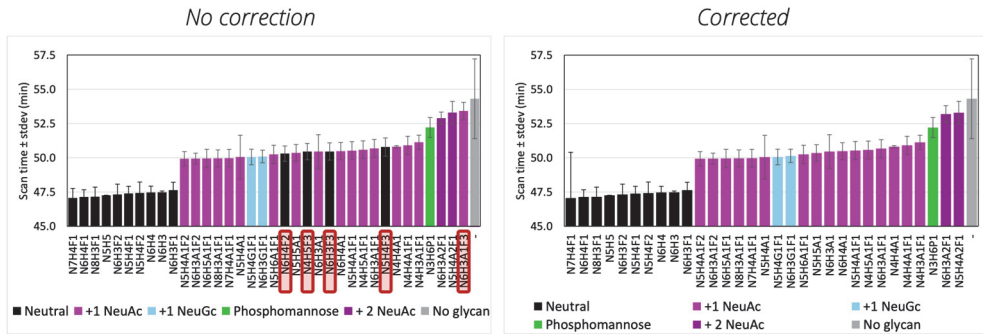
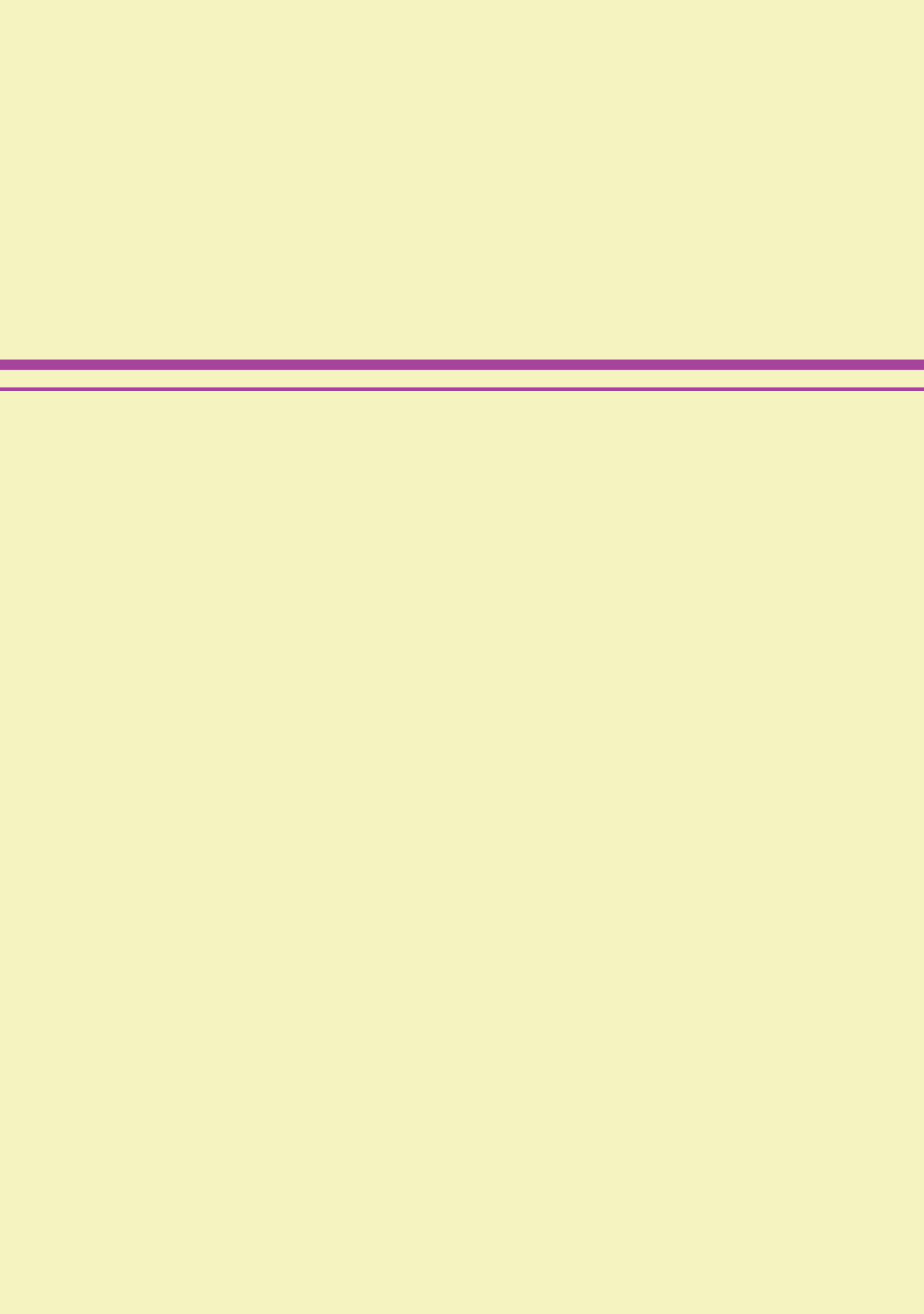


Figure S4.4 – Scan times of fragmentation spectra of all α -lactalbumin peptides with the amino acid sequence WVCTTFHTSGYDTQAIQNNDDSTE, covering N-glycosylation site Asn45. This data encompasses a total of 5526 peptide spectrum matches from the analysis in triplicate of each sample collected during lactation at 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 7.0, 14.0 and 28 d after calving from the four individual cows, i.e., Cow 1, Cow 2, Cow 3 and Cow 4. The glycan compositions are indicated based on their biochemical properties: neutral (black), sialylated with one NeuAc residue (magenta), sialylated with one NeuGc residue (light blue), sialylated with two NeuAc residues (dark magenta) and phosphomannose (green). Indicated in grey are the retention times of the non-glycosylated peptide. The plot on the left indicates the glycan compositions and their respective scan times as identified by Byonic v4.5.2. Highlighted in red are the glycan compositions that we found to have been misidentified by Byonic v4.5.2, by the misassignment of one NeuAc residue as two fucose residues. The plot on the right indicates the glycan compositions and their respective scan times following manual correction based on careful fragmentation spectra verification, illustrating the clustering of N-glycopeptides in the retention time dimension in reversed-phase chromatography based on the charge of their respective glycans.



CHAPTER 5

Summary and Outlook



Chapter 5. Summary and Outlook

Summary

When I started my PhD studies, I already had expertise in milk protein research, but no experience with mass spectrometry. I learned about the different aspects of mass spectrometry starting from basic principles, to advanced understanding of instrumentation, software, different tiers of analysis, and state-of-the-art technology. This opened a new world of possibilities for elevating milk protein research, which I strived to achieve during my time as PhD student in the NWO funded SATIN consortium (project 731.017.202) at Utrecht University. This thesis represents a selection of the research projects I have performed during my PhD studies, which focused on the characterisation of bovine milk proteins using high-resolution mass spectrometry.

Chapter 1 is divided into two parts, which together introduce the context in which this thesis is set. The first part of this chapter introduces the proteins of bovine milk, providing a brief historic overview of milk and dairy products in the human diet, the golden age of studying proteins following the sequencing of the genome, and the discovery of milk proteins, followed by detailed characterisation of the most abundant proteins of bovine milk. The second part of this chapter provides an introduction and chronicle of mass spectrometry, discussing commonly-used techniques in the study of milk proteins and introducing the modern mass spectrometric approaches applied for the work described in this PhD thesis (LC-MS/MS at intact protein level, oxonium ion-triggered hybrid fragmentation methods at glycopeptide level), highlighting their advantages over those of the classical approaches (LC-MS at intact protein level, protein glycosylation analysis at released glycan and/or deglycosylated peptide level). My insights into and analytical experience with mass spectrometry techniques, gained in over four years of my PhD studies, are summarised here, providing a detailed description of peptide-centric, protein-centric and glycoproteomics approaches. Particular emphasis is placed on the fact that no one technique can answer all research questions. A well-defined research question and careful experimental design are paramount to successfully obtaining the sought-after answers.

Heat treatment and drying are amongst the core technologies for ensuring microbial safety and extending the shelf life of dairy products, thereby also reducing food waste. However, thermal processing can result in unintended chemical modifications, with in particular glycation of primary amine residues on the proteins with reducing sugars, of which lactose is most prevalent in milk. To study the development of glycation, we focused on the six most abundant proteins of bovine milk (α_{s1} -, α_{s2} -, β - and κ -casein, α -lactalbumin and β -lactoglobulin), and

monitored the changes in their heterogeneity during production and storage of skim milk powder. For this purpose, we used a hybrid approach, combining protein-centric LC-MS(/MS) with peptide-centric multi-protease LC-MS/MS to identify and monitor proteoform development. This work forms the basis of **Chapter 2**. At intact protein level, we verified protein identification by database search of the LC-MS/MS data, after which we assigned proteoforms based on their intact masses obtained from LC-MS. This allowed us to identify the native proteoforms of the six most abundant bovine milk proteins, consisting of a combination of different genetic variants and/or different extents of post-translational modification, with in particular phosphorylation of Ser/Thr residues. We further monitored changes in proteoform profiles as a result of increases in heterogeneity caused by glycation, identifying the extent of glycation *per* proteoform, and distributions of the glycated proteoforms at different stages of the production process and during storage. The protein-centric approach, however, did not provide site localization of the glycated residues. To obtain this information, we analysed the samples using a peptide-centric approach. The usual choice for tryptic digest of the proteins presented disadvantages with respect to sequence coverage in protein regions with either dense or sparse presence of tryptic cleavage sites, *i.e.*, Lys and Arg residues, particularly in caseins. Furthermore, glycated proteins exhibited suboptimal digestibility by trypsin, due to Lys and Arg residues also representing the prime glycation targets. We overcame this limitation by performing multi-protease digestions with a selection of four proteases (trypsin, chymotrypsin, and endoproteinases GluC and AspN), and combinations thereof, which yielded high protein sequence coverage, and identification of the glycation hot spots. We hereby expanded on the current understanding of milk protein glycation, having identified new glycation hot spots and discussing their localization in relation to protein structures.

In **Chapter 3**, we took a step back from dairy product processing, and investigated changes occurring in bovine mammary secretions during the transition from colostrum to mature milk, with emphasis on bovine milk immunoglobulin G (IgG). IgG is the critical component of bovine colostrum, essential for providing passive immunity to the calf, and supporting its immune development, as well as a biofunctional component in the context of human diet. We focused this study on the bovine IgG heavy chain constant region *N*-glycosylation, a post-translational modification that directly impacts the potency and interactions of IgG, amongst others. For this purpose, we analysed samples from four individual cows at 9 time points in the range of 0.5-28.0 d after calving. We analysed the samples by peptide-centric LC-MS/MS to map total sample proteome, and the lactational changes of

IgG relative to the rest of the proteome. In this chapter we also illustrate the impact of protein database on the identification and label-free quantification of bovine mammary secretion proteins, and describe a means by which we optimised the bovine protein database for improved identification and relative quantitation. In parallel, we used a peptide-centric glycoproteomics workflow on tryptic digests of IgG captured from the samples of the four individual cows across lactation from the 0.5-d colostrum to the 28.0-d mature milk. The principle of our glycoproteomics workflow is the selective hybrid fragmentation (by stepped higher-energy collisional dissociation [HCD] or by electron-transfer/higher-energy collisional dissociation [ET_hCD]) of glycopeptides, triggered on the detection of oxonium ions during regular HCD fragmentation, followed by the database search of the results with dedicated glycoproteomics software and manual curation of the identifications. While this approach of analysing glycopeptides is still in its infancy, it is in the course of being more widely adopted, and it presents clear advantages over the analysis at released glycans and at deglycosylated peptide levels. The hereby investigated lactational dynamics of bovine IgG *N*-glycosylation proved to be diverse and heterogeneous. Here we identified *N*-acetylneuraminic acid (NeuAc) sialylation as the key abundant characteristic of bovine colostrum *N*-glycans, significantly decreasing in the first days of lactation, and being barely detectable in mature bovine milk IgG. Studying the detailed molecular characteristics of bovine IgG and their dynamic changes during the transition of colostrum to mature milk is important both from the perspective of animal health, as well as in the context of biofunctional foods in human nutrition.

Next to IgG, many other glycoproteins are prevalent in milk. With protein glycosylation playing a role in many biological processes and functions, the biofunctional properties of milk are in part attributed to the milk glycoproteins. **Chapter 4** is therefore an extension of the work done in Chapter 3, to expand the investigation of *N*-glycosylation to the whole detectable *N*-glycoproteome, in the samples from the four individual cows at the 9 time points between 0.5 and 28.0 d after calving, without prior enrichment of glycoproteins or glycopeptides. The results of this chapter highlight the layered complexity of the *N*-glycoproteome of bovine mammary secretions during the transition from colostrum to mature milk. A first level of changes occurred as a result of the relative lactational changes in (*N*-glyco)protein abundances. Bovine colostrum *N*-glycoproteome was found to be dominated by IgG, secretory immunoglobulins sIgA and sIgM, and blood serum-derived proteins, including serotransferrin, fetuin A and zinc- α 2-glycoprotein. The *N*-glycans of colostrum *N*-glycoproteins featured sialylation with NeuAc, with *N*-glycolylneuraminic acid (NeuGc), or with both NeuAc and NeuGc. During the

transition from colostrum to mature milk, the proportion of immunoglobulins and blood serum-derived *N*-glycoproteins decreased considerably in the first 3 days after calving, in favour of the milk *N*-glycoproteins, including α -lactalbumin, lactophorin and the proteins of the milk fat globule membrane. The glycans of mature milk *N*-glycoproteins were identified to be primarily sialylated with only NeuAc. As a result, the defining feature of colostrum *N*-glycoproteome was identified as NeuGc sialylation. These findings raise interesting questions regarding the balance between the beneficial and the potentially detrimental health effects of bovine colostrum in human nutrition, due to NeuGc sialylation being immunogenic in humans. The second level of lactational changes in the *N*-glycoproteome were the protein-specific changes in *N*-glycosylation. This was also exemplified in Chapter 3 for IgG purified from the colostrum and milk samples, where, contrary to the trend of the general *N*-glycoproteome, we identified that NeuAc sialylation was the signature biochemical feature of colostrum IgG *N*-glycans. Here, we further characterised the *N*-glycosylation of bovine α -lactalbumin in depth, a protein that was abundantly present in all investigated samples and that is of particular interest, amongst others, in the production of humanised cow milk-based infant formula. We identified *N*-glycosylation at Asn45 and Asn74, both of which exhibited partial occupancy, which was not influenced by lactation time point or individual animal. The repertoire of identified *N*-glycans consisted of primarily diantennary complex structures, showing core fucosylation, NeuAc sialylation and a high ratio of LacdiNAc/LacNAc antennae as the main biochemical features. We found lactational decreases in the NeuAc sialylation of α -lactalbumin, however, not as pronounced as in the case of the previously-studied IgG. This work illustrates the complexity and dynamics of the colostrum and milk *N*-glycoproteome, both as a whole, as well as at protein-specific level. Understanding not only the overall *N*-glycoproteome of a mammary secretion sample, but also the protein-specific *N*-glycosylation, and changes thereof, is particularly important when, *e.g.*, producing and using milk protein ingredients in which one or more proteins are enriched or isolated.

Nederlandse samenvatting

Toen ik aan mijn promotieonderzoek begon, had ik al een brede kennis van melkeiwitonderzoek, maar geen ervaring met massaspectrometrie. Ik heb veel geleerd over de verschillende aspecten van massaspectrometrie, van de basisprincipes, tot geavanceerd begrip van instrumentatie, software, verschillende analyse-niveaus en grensverleggende ontwikkelingen in technologie. Dit opende een nieuwe wereld aan mogelijkheden om onderzoek naar melkeiwitten naar een hoger niveau te verheffen, iets waar ik naar streefde tijdens mijn tijd als

promovenda binnen het door NWO gesponsorde SATIN consortium (project 731.017.202) aan de Universiteit Utrecht. Dit proefschrift beschrijft een selectie van de onderzoeksprojecten die ik heb uitgevoerd tijdens mijn promotieonderzoek, gericht op de karakterisering van koemelkeiwitten met behulp van hoge-resolutie massaspectrometrie.

Hoofdstuk 1 is verdeeld in twee delen, die samen de context schetsen waarin dit proefschrift zich bevindt. Het eerste deel van dit hoofdstuk introduceert de koemelkeiwitten, met een kort historische overzicht van melk en zuivelproducten in het menselijke dieet, de huidige gouden eeuw van het bestuderen van eiwitten na het bepalen van de sequentie van het genoom en de ontdekking van melkeiwitten, gevolgd door gedetailleerde karakterisering van de meest voorkomende eiwitten in koemelk. Het tweede deel van dit hoofdstuk geeft een inleiding en een overzicht van massaspectrometrie, bespreekt veelgebruikte technieken in de studie van melkeiwitten en introduceert moderne massaspectrometrische methoden die in dit proefschrift worden toegepast, waarbij de voordelen daarvan worden toegelicht ten opzichte van de klassieke aanpakken. Mijn inzichten in, en analytische ervaring met, massaspectrometrische technieken, opgebouwd de in meer dan vier jaar van mijn promotieonderzoek, worden hier samengevat en vormen een gedetailleerde beschrijving van peptidecentrische, eiwitcentrische en glycoproteomics aanpakken. Bijzondere nadruk wordt gelegd op het feit dat geen enkele techniek alle onderzoeksvragen volledig kan beantwoorden en een zorgvuldig experimenteel ontwerp is van groot belang om met succes de gezochte antwoorden te verkrijgen.

Hittebehandeling en drogen behoren tot de kerntechnologieën voor het waarborgen van microbiële veiligheid en het verlengen van de houdbaarheid van zuivelproducten, waardoor ook voedselverspilling wordt verminderd. Thermische verwerking kan echter resulteren in onbedoelde chemische reacties, met name glycatie van primaire amineresiduen op eiwitten met reducerende suikers, vooral lactose wat het meest voorkomt in melk. Om de ontwikkeling van glycatie te bestuderen, hebben we ons gericht op de zes meest voorkomende eiwitten in koemelk (α_{s1} -, α_{s2} -, β - en κ -caseïne, α -lactalbumine en β -lactoglobuline), en hebben we de veranderingen in hun heterogeniteit tijdens het productieproces en opslag van magere melkpoeder gevolgd. Voor dit doel hebben we een hybride aanpak gebruikt, waarbij we eiwit-centrische LC-MS(/MS) hebben gecombineerd met peptide-centrische multi-protease LC-MS/MS om de ontwikkeling van eiwit variaties te identificeren en te volgen. De totale variatie van eiwitten die allemaal uit een gen voorkomen wordt ook wel het proteovormprofiel genoemd. Dit werk vormt de basis van **Hoofdstuk 2**. Op intact eiwitniveau hebben we de

eiwitidentificatie geverifieerd door het zoeken van de LC-MS/MS resultaten tegen een database, waarna we proteovormen hebben toegewezen op basis van hun intacte massa verkregen uit LC-MS. Hierdoor konden we de natieve proteovormen van de zes meest voorkomende koemelkeiwitten identificeren, bestaande uit een combinatie van verschillende genetische varianten en/of verschillende mate van posttranslationale modificatie, met name fosforylering van Ser en Thr residuen. We volgden verder de veranderingen in proteovormprofielen als gevolg van toename in heterogeniteit veroorzaakt door glycatie, waarbij we de mate van glycatie per proteovorm en distributies van de geglyceerde proteovormen in verschillende stadia van het productieproces en tijdens opslag hebben geïdentificeerd. De eiwitcentrische aanpak zorgde echter niet voor plaatsbepaling van de geglyceerde residuen. Om deze informatie te verkrijgen, hebben we de monsters geanalyseerd met behulp van een peptide-centrische aanpak. De gebruikelijke keuze voor tryptische digestie van de eiwitten vertoonde nadelen met betrekking tot sequentiedekking in eiwitgebieden met dichte of schaarse aanwezigheid van tryptische splitsingsplaatsen, d.w.z. Lys- en Arg-residuen, vooral in de caseïnes. Bovendien vertoonden geglyceerde eiwitten een suboptimale verteerbaarheid door trypsine, omdat Lys- en Arg-residuen ook glycatieplaatsen kunnen zijn. We hebben deze beperking overwonnen door multiproteasedigesties uit te voeren met een selectie van vier proteasen (trypsine, chymotrypsine en endoproteasen GluC en AspN) en combinaties daarvan, wat een hoge eiwitsequentiedekking en identificatie van de glycatiehotspots opleverde. Door de identificatie van nieuwe glycatiehotspots en het bespreken van hun lokalisatie in relatie tot de eiwitstructuur hebben we het huidige begrip van melkeiwitglycatie uitgebreid.

In **Hoofdstuk 3** hebben we de veranderingen die optreden in koemelkkliersecreties tijdens de overgang van biest naar rijpe melk onderzocht, met de nadruk op immunoglobuline G (IgG). IgG is de kritieke component van koeienbiest, essentieel voor het bieden van passieve immuniteit aan het kalf en het ondersteunen van de ontwikkeling van zijn immuunsysteem, evenals een biofunctionele component in de context van menselijke voeding. We hebben deze studie gericht op de *N*-glycosylering van de constante deel van de zware keten van koeien-IgG, een posttranslationale modificatie die rechtstreeks van invloed is op, onder andere, de activiteit en interacties van IgG. Hiervoor hebben we monsters van vier individuele koeien op 9 tijdstippen in het bereik van 0,5-28,0 dagen na het afkalven geanalyseerd. We hebben de monsters geanalyseerd met peptide-centrische LC-MS/MS om het totale proteoom en de lactatieveranderingen van IgG ten opzichte van de rest van het proteoom in kaart te brengen. In dit hoofdstuk illustreren we ook de invloed van de eiwitdatabase op de identificatie en labelvrije kwantificering

van de eiwitten van koemelkkliersecreties, en beschrijven we een manier waarop we de koeieneiwit database hebben geoptimaliseerd voor verbeterde identificatie en relatieve kwantificering. Tegelijkertijd hebben we een peptidecentrische glycoproteomics workflow gebruikt op tryptische digesties van IgG geïsoleerd uit de monsters van de vier individuele koeien tijdens de lactatie van de 0,5-dagen biest tot de 28,0-dagen rijpe melk. Het principe van onze glycoproteomics workflow is de selectieve hybride fragmentatie (door getrapte botsingsdissociatie met hogere energie [HCD] of door elektronenoverdracht/botsingsdissociatie met hogere energie [ETHCD]) van glycopeptiden, geactiveerd bij de detectie van oxonium ionen tijdens reguliere HCD fragmentatie, gevolgd door het doorzoeken van de resultaten in de database met speciale glycoproteomicssoftware en handmatige curatie van de identificaties. Hoewel deze aanpak van het analyseren van glycopeptiden nog nieuw is, wordt het steeds breder toegepast en biedt het duidelijke voordelen ten opzichte van de analyse op het niveau van vrijgemaakte glycanen en gedeglycosyleerde peptiden. De hierbij onderzochte lactatiedynamiek van koeien-IgG *N*-glycosylering bleek divers en heterogeen te zijn. We identificeerden *N*-acetylneuraminezuur (NeuAc)-sialylering als het belangrijkste overvloedige kenmerk van *N*-glycanen van koeienbiest, aanzienlijk afnemend in de eerste dagen van lactatie, en nauwelijks detecteerbaar in IgG van rijpe koemelk. Het bestuderen van de gedetailleerde moleculaire kenmerken van koeien IgG en hun dynamische veranderingen tijdens de overgang van biest naar rijpe melk, is belangrijk, zowel vanuit het perspectief van diergezondheid als in de context van biofunctioneel voedsel in menselijke voeding.

Naast IgG komen veel andere glycoproteïnen veel voor in melk. Omdat eiwitglycosylering een rol speelt in veel biologische processen en functies, worden de biofunctionele eigenschappen van melk gedeeltelijk toegeschreven aan de melkglycoproteïnen. **Hoofdstuk 4** is daarom een uitbreiding van het werk dat in Hoofdstuk 3 is gedaan, om het onderzoek naar *N*-glycosylering uit te breiden naar het gehele detecteerbare *N*-glycoproteoom, in de monsters van de vier individuele koeien op de 9 tijdstippen tussen 0,5 en 28,0 dagen na het afkalven, zonder voorafgaande verrijking van glycoproteïnen of glycopeptiden. De resultaten van dit hoofdstuk benadrukken de gelaagde complexiteit van het *N*-glycoproteoom van koemelkkliersecreties tijdens de overgang van biest naar rijpe melk. Een eerste niveau van veranderingen vond plaats als gevolg van de relatieve lactatieveranderingen in (*N*-glyco)proteïne-abundanties. Koeienbiest-*N*-glycoproteoom bleek te worden gedomineerd door IgG, secretoire immunoglobulinen sIgA en sIgM, en van bloedserum afgeleide eiwitten, waaronder serotransferrine, fetuïne A en zink- α 2-glycoproteïne. De *N*-glycanen van biest-*N*-

glycoproteïnen vertoonden sialylering met NeuAc, met *N*-glycolylneuraminezuur (NeuGc), of met zowel NeuAc als NeuGc. Tijdens de overgang van biest naar rijpe melk daalde het aandeel immunoglobulinen en van bloedserum afgeleide *N*-glycoproteïnen aanzienlijk in de eerste 3 dagen na het kalven, ten gunste van de melk-*N*-glycoproteïnen, waaronder α -lactalbumine, lactophorine en de eiwitten van het melkvetbolletjesmembraan. Van de glycanen van *N*-glycoproteïnen van rijpe melk werd vastgesteld dat ze primair gesialyleerd waren met alleen NeuAc. Hierdoor werd NeuGc-sialylering geïdentificeerd als het bepalende kenmerk van biest-*N*-glycoproteoom. Deze bevindingen roepen interessante vragen op met betrekking tot de balans tussen de gunstige en de potentieel schadelijke gezondheidseffecten van koeienbiest in menselijke voeding, omdat NeuGc-sialylering immunogeen is bij mensen. Het tweede niveau van lactatieveranderingen in het *N*-glycoproteoom waren de eiwit-specifieke veranderingen in *N*-glycosylering. Dit werd ook geïllustreerd in Hoofdstuk 3 voor IgG geïsoleerd uit de biest- en melkmonsters, waar we, in tegenstelling tot de trend van het algemene *N*-glycoproteoom, identificeerden dat NeuAc-sialylering de kenmerkende biochemische eigenschap was van biest IgG *N*-glycanen. Hier hebben we de *N*-glycosylering van koeien- α -lactalbumine verder gekarakteriseerd, een eiwit dat overvloedig aanwezig was in alle onderzochte monsters, en dat onder andere van bijzonder belang is bij de productie van gehumaniseerde zuigelingenvoeding op basis van koemelk. We hebben *N*-glycosylering geïdentificeerd op Asn45 en Asn74, die beide een gedeeltelijke bezetting vertoonden die niet werd beïnvloed door het tijdstip van lactatie of individueel dier. Het repertoire van geïdentificeerde *N*-glycanen bestond voornamelijk uit tweezijdige complexe structuren, met kernfucosylering, NeuAc-sialylering en een hoge verhouding van LacdiNAc/LacNAc-antennes als de belangrijkste biochemische eigenschappen. We vonden dalingen in de NeuAc-sialylering van α -lactalbumine gedurende de lactatie, echter niet zo uitgesproken als in het geval van het eerder bestudeerde IgG. Dit werk illustreert de complexiteit en dynamiek van het biest- en melk-*N*-glycoproteoom, zowel in het geheel als op eiwit-specifiek niveau. Het begrijpen van niet alleen het algehele *N*-glycoproteoom van een melkkliersecretiemonster maar ook van de eiwit-specifieke *N*-glycosylering en veranderingen daarvan, is vooral belangrijk bij bijvoorbeeld het produceren en gebruiken van melkeiwitingrediënten waarin een of meer eiwitten zijn verrijkt of geïsoleerd.

Outlook

I dare say we are standing on the shoulders of giants, looking back at a legacy of over two centuries of milk protein research (Berzelius, 1814, Braconnot, 1830), and

over one century of developments in mass spectrometry (Thomson, 1897). This naturally raises the question: *so is there still anything left to be learned about milk proteins?* And the shortest and most direct answer is: *Absolutely, yes!*

While the milk of many mammals, especially bovines and humans, has been studied extensively, there are nevertheless many knowledge gaps yet to be addressed. Oftentimes, the key to even realising there are knowledge gaps, and then to developing the missing knowledge and understanding, lies in breaking out of the comfortable circle of a traditional discipline, and venturing into different fields of study, to ultimately make use of an interdisciplinary approach. Put in very simple words, it is useful to take a step outside the box, and simply look at the matter at hand from a different perspective. For me, the box was dairy science, and I did not know that I was in a box, until I came to a group of non-dairy scientists, biologists, chemists and physicists. Here, I learned to look back at my projects from their various perspectives, to adopt methodologies from different fields to better answer my questions, and also to reciprocally use my dairy science expertise to help answer the questions of my peers in other disciplines.

A rule of thumb with mass spectrometry is that we are not likely to identify what we are not searching for. In other words, protein sequences or modifications of the proteins that may exist in the analysed samples are not likely to be identified if they are not included in the database search. Our previously-published study by Dingess *et al.* (2021) is a prime example of how a combined interdisciplinary approach yielded results that challenge the textbook knowledge of milk proteins. The learnings from modern glycoproteomics approaches prompted Dingess *et al.* (2021) to check oxonium ion traces in the MS/MS spectra of endogenous human milk peptidome samples analysed by LC-MS/MS with conventional HCD fragmentation. The samples appeared rich in oxonium ion traces, indicating the abundant presence of glycopeptides. Furthermore, since the most abundant endogenous peptides in those samples were derived from human β -casein, this brought about a full investigation into human β -casein glycosylation, although textbook knowledge of caseins from any mammal indicated that of all the caseins, only κ -casein is a glycoprotein with a part of the molecules harbouring *O*-glycosylation (Huppertz, 2013). In the study of Dingess *et al.* (2021) we identified *O*-glycosylation of human β -casein in endogenous peptides and also in the intact protein. We characterised the unmodified, phosphorylated and/or *O*-glycosylated proteoforms of human β -casein in two donors across lactation, as well as confirming site localization of the novel *O*-glycosylation using peptide-centric glycoproteomics. These findings further raised the question whether bovine milk caseins also harboured similar and elusive post-translational modifications.

Additional analyses for *O*-glycopeptide identification on the data acquired in Chapter 3, revealed the existence of novel *O*-glycosylation in all bovine caseins, as exemplified in Figure 5.1. While Figure 5.1 only shows a few representative casein *O*-glycopeptide spectra, *O*-glycosylation was identified at multiple sites in each of the bovine caseins. These results demonstrated for the first time that in fact all caseins, and not solely κ -casein, exist also as *O*-glycoproteins in bovine milk. Furthermore, κ -casein *O*-glycosylation has so far been known to occur only on the caseinomacropeptide (CMP) region of the protein (residues Met106-Val169 of the mature protein), also termed the glycomacropptide (GMP)(Huppertz, 2013). Our findings showed a novel *O*-glycosylation also in the para- κ -casein region of the protein (residues Pyr1-Phe105 of the mature protein), which has not been reported to be *O*-glycosylated until now. The expansion of knowledge of milk proteins brought on by the research described in this thesis was made possible thanks to modern mass spectrometry technologies. Mass spectrometry, proteomics and the emerging sub-field of glycoproteomics are undergoing a continuing revolution that constantly challenges and crosses the boundaries of what has been possible so far. The adoption and application of modern mass spectrometric technologies to address dairy research questions deeply and comprehensively is still somewhat limited, despite them being very powerful analytical tools prevalently used in biopharma and clinical fields.

Milk is an essential biofluid for the neonate and developing infants, and a core constituent of the diet throughout life. Milk is not just simply another food category that boils down to a macronutrient composition and energy content. Next to containing highly-bioavailable nutrients, milk constitutes a source of biofunctional components that contribute to gut health, the development of a healthy microflora and modulate the immune system, amongst others. These biological functionalities of milk constituents result from their interactions with each other, the changes they undergo during digestion, their interactions with gut microbiota, whether commensals, probiotics or (opportunistic) pathogens, their interactions with the intestinal mucosa, their interaction with specific receptors on the human cells, and so forth. At a molecular level, it is the structure and properties of the milk proteins, their surface composition, the occurrence, and nature of post-translational modifications, that dictate whether and how they change in the gastro-intestinal tract and how they interact further. Insights into all of these aspects can be gained by using and combining different tiers of mass spectrometry. Peptide-centric proteomics and glycoproteomics approaches provide information on amino acid sequence coverage and identification, characterization, site localization and micro-heterogeneity of post-translational modifications, as shown in Chapters 2-4.

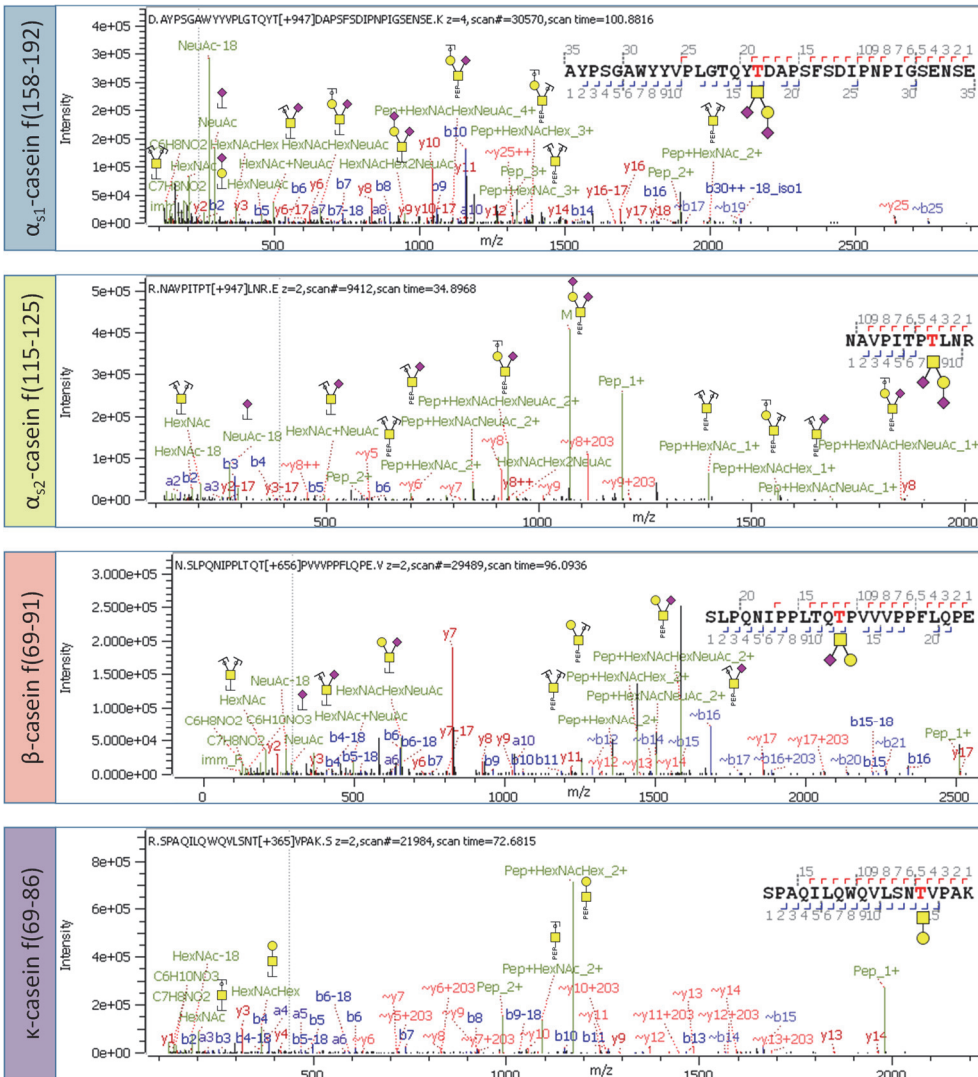


Figure 5.1 – Glycopeptide fragmentation spectra illustrating novel O-glycosylation discovered on each of the bovine milk caseins. The depicted b, y and B and Y fragment ions confirm sequencing of both the peptide backbone, and the glycan structure, confirming the existence of the glycopeptide. From top to bottom: α_{s1}-casein, α_{s2}-casein, β-casein and κ-casein. These results were obtained from additional analyses of the data presented in Chapter 3, namely database search for O-glycosylation of the bovine caseins using Byonic v4.5.2 (Protein Metrics, Cupertino, California, USA). Glycan residue structures were built using GlycoWorkBench 2.1 build 146, according to the symbol nomenclature for glycan representation of the Consortium for Functional Glycomics (Varki et al., 2009): ■ N-acetylgalactosamine (GalNAc), ● galactose (Gal), and ◆ N-acetylneuraminic acid (NeuAc). The uniqueness of the peptide sequences to the assigned proteins was verified with the Peptide search tool of UniProt.org.

Protein-centric approaches at single protein level provide information on the

proteoform composition and distribution, and the meta-heterogeneity of post-translational modifications within the protein, as shown in Chapter 2, complemented with the result of the peptide-centric approach. The protein-centric approach can be taken further, and native mass spectrometry and charge-detection mass spectrometry (CDMS) techniques can be used to investigate the interactions between different proteins in macromolecular protein assemblies. Structural features, dynamic properties and protein interactions can be further investigated by techniques such as Hydrogen/Deuterium eXchange Mass Spectrometry (HDX-MS) and crosslinking mass spectrometry (XL-MS). With this, I wish to highlight that many aspects about understanding milk proteins and their biological

and physical-chemical functionality are still understudied. However, thanks to recent technological advancements, modern analytical techniques that can be used to fill the knowledge gaps exist, providing ample opportunity for future milk protein research.

Thanks to the continuing development of the different mass spectrometric technologies, and the introduction of new-generation instruments that are more sensitive, more precise, faster and outperform previous instrumentation, it is now possible to generate new information, more information and faster than ever. However, does this also reveal relevant information? To somewhat rephrase this: *“Nice thesis, I see you found some new things. But what do they all mean?”* Just because we have the means to obtain a lot of information, this does not mean we always need to or that we should. Research studies should have a set aim based on which the study is designed to provide the optimal, necessary information, rather than becoming a data dump of information gathered for the sake of gathering. The beauty of science is that we set an aim, we work towards it, and we uncover so many more questions along the way, not all of which can be answered in a single project. This leads to new research ideas, new aims, and the design of new projects, ultimately paving the path to the important discoveries. Taking the work presented in Chapter 3 as an example, we aimed to and successfully characterised the *N*-glycosylation of IgG in the mammary secretions transitioning from colostrum to mature milk in the first month after calving. However, our findings raised many questions: what is the function of the NeuAc sialylation in colostrum IgG? Does it cause differential recognition of the IgG by the neonatal Fc receptor (FcRn), triggering either recirculation or cross-membrane transport of the IgG? Does it confer higher stability to the IgG to the digestive tract conditions and against the digestive enzymes of the calf? How do specific IgG *N*-glycoforms modulate its binding affinity and neutralising capacity? How do specific bovine IgG *N*-glycoforms interact with human Fc-gamma receptors (FcγR)? Do specific IgG *N*-glycoforms

exhibit better resistance to degradation in the human digestive tract? Are they friends or foes in human nutrition? These questions illustrate that despite having learned so many new insights, there is ample opportunity for future research into investigating which of these learnings are important, whether they are beneficial or detrimental, and what the underlying mechanism is that explains their function.

In a similar line of thought, if we take a step back at the above-mentioned novel casein *O*-glycosylation, this discovery also raises many questions. My personal question is: why is it there at all? I believe a living organism is optimised from an evolutionary standpoint, and the modifications we see occur there for a reason. If we look at κ -casein, *O*-glycoforms of it always exist in milk, but site occupancies are always partial. Why is it that we never see a milk where all κ -casein is non-glycosylated or a milk where all κ -casein is glycosylated? Is the case the same for the other caseins? So far, we have consistently detected the novel bovine casein *O*-glycosylation in all analysed milk samples, irrespective of lactation time, both in the milk from individual cows, and in pooled farm milk. The well-documented GMP *O*-glycosylation is linked to both biological functionality of the GMP, as well as physical-chemical functionality of the casein micelles, whereby casein micelles with a higher proportion of *O*-glycosylated κ -casein correlate with smaller micelle sizes and improved rennet-coagulation properties during cheesemaking. The occurrence of other casein proteoforms, such as specific genetic variants or composite genotypes, along with specific levels of phosphorylation also correlate with the physical-chemical functionality of the milk. This further raises the question whether and how the novel casein *O*-glycosylation impacts the biological and physical-chemical functionality of the milk and its proteins.

In summary, to circle back to initial paragraph: we are standing on the shoulders of giants, we have gained so much knowledge and understanding, and yet at times it feels like we still know so little. There is a bright future for milk protein research, and the future is now.

References

Berzelius J. 1814. Über Thierische Chemie. *Schweiggers Journal für Chemie Physik*, 11:261-280.

Braconnot H. 1830. Ueber den Käsestoff und die Milch, und deren neue Nutzenwendungen. *Annalen der Physik*, 95:34-47.

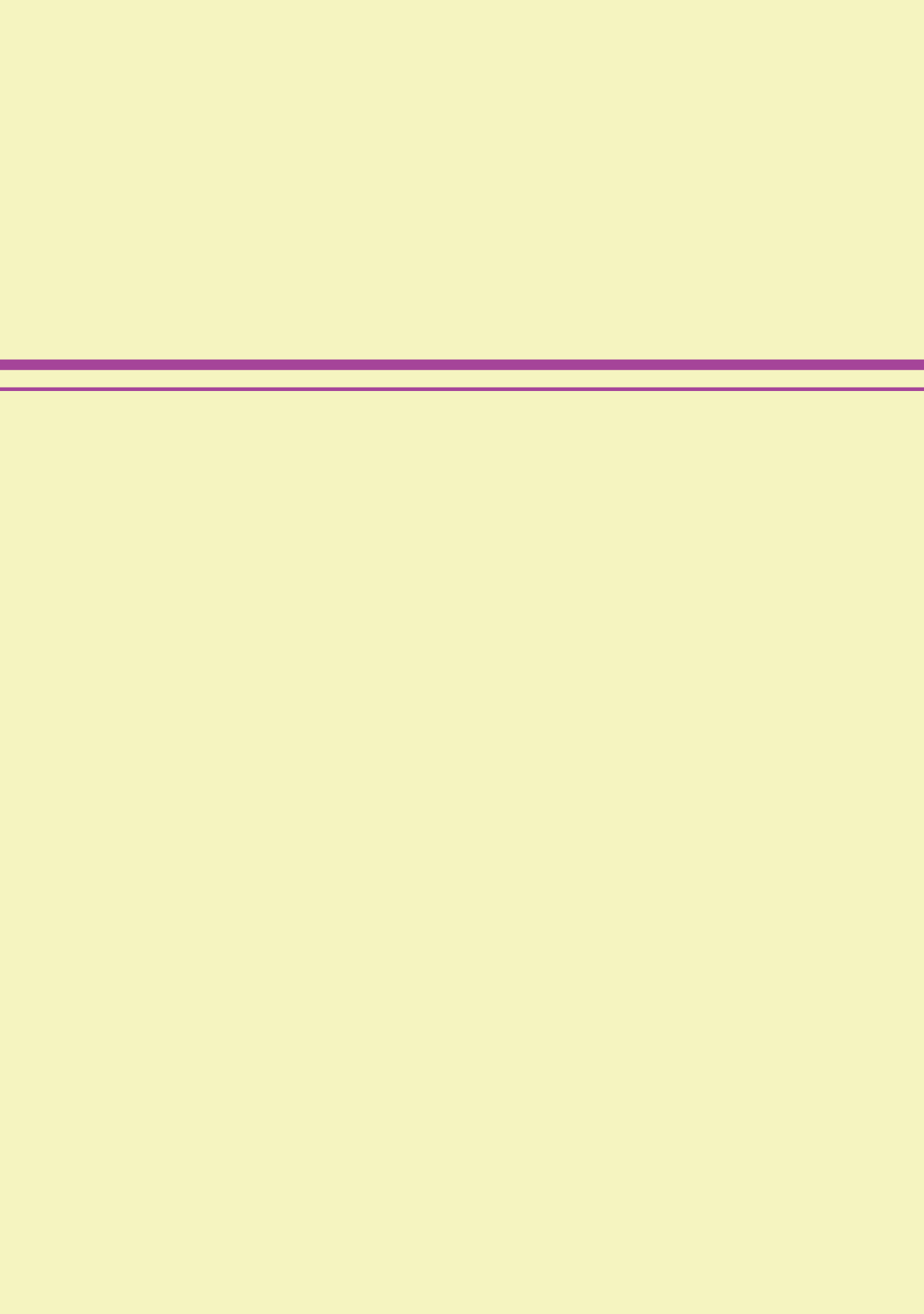
Dingess KA, Gazi I, van den Toorn HW, Mank M, Stahl B, Reiding KR, Heck AJ. 2021. Monitoring human milk β -casein phosphorylation and *O*-glycosylation over

lactation reveals distinct differences between the proteome and endogenous peptidome. *Int J Mol Sci*, 22:8140.

Huppertz T. 2013. Chemistry of the caseins. *Advanced dairy chemistry*: Springer. p. 135-160.

Thomson JJ. 1897. XL. Cathode rays. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 44:293-316.

Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Marth JD, Bertozzi CR, Hart GW, Etzler ME. 2009. Symbol nomenclature for glycan representation. *Proteomics*, 9:5398-5399.



ABOUT THE AUTHOR



About the author

Curriculum Vitae

I was born on Friday the 13th of February, 1987, in Iași, Romania, where my parents were both completing their Bachelor degrees in Engineering. I come from a multicultural background, born and raised in Romania, between the *Székely* ethnicity on my mother's side, and the *Qırımatar* one on my father's side.



In 2010, I obtained my BSc degree in Food Product Engineering from “Dunărea de Jos” University of Galați, Romania. After this, I relocated to The Netherlands and completed my MSc Food Technology degree in 2013 in the Dairy Science and Technology specialisation at Wageningen University and Research. During this time, I did my MSc internship at NIZO Food Research focusing on the physical-chemical functionality of milk proteins from a set of milk protein concentrates ranging from skim milk powder to milk protein isolate. Due to the success of my project, my internship was extended to a grand total of 15 months, and this work later materialised into multiple peer-reviewed scientific publications. Following my MSc graduation, I worked at NIZO Food Research for 5 years as Project Manager and Dairy Scientist. During this time, I focused on many areas of dairy science, with milk proteins consistently as a central theme.

It has long been my ambition to complete my foundation as scientist with a PhD degree. In 2018, I therefore left NIZO Food Research and embarked on a new journey in the Biomolecular Mass Spectrometry and Proteomics group at Utrecht University, pursuing a PhD project under the supervision of Professor Albert J. R. Heck, within the NWO SATIN consortium (project 731.017.202), a partnership between Utrecht University, Leiden University Medical Centre, Royal FrieslandCampina, Royal DSM and Roche. This thesis titled ***Unravelling elusive mysteries of bovine milk proteins by mass spectrometry*** summarises a great part of my doctoral studies.

Over the past 10 years, I have developed myself into a respected scientist in the field of dairy science, to which my portfolio of scientific publications is a testimony. I have been a valued reviewer for many years and as of 2021, an Editorial Board member of the International Dairy Journal, Elsevier. Starting January 2023, I have

joined the International Dairy Journal in a new position as Associate Editor. Milk proteins have been a core focus of my work across the years. My current interests with regards to milk proteins are related to understanding their biology, their intended biological functionality, the modulating role of their post-translational modifications therein, and their role in human health and nutrition. I am looking forward to continuing my work into this direction as a post-doctoral researcher starting in 2023 at Biomolecular Mass Spectrometry and Proteomics under the continued guidance of Prof. Dr. Albert J. R. Heck and Dr. Karli R. Reiding, and in collaboration with Prof. Dr. Thom Huppertz and André Groeneveld from Royal FrieslandCampina.

List of publications

1. **Gazi, I.**, Reiding, K. R., Groeneveld, A., Bastiaans, J., Huppertz, T., & Heck, A. J. R. (2023). Key changes in bovine milk immunoglobulin G during lactation: NeuAc sialylation is a hallmark of colostrum immunoglobulin G N-glycosylation. *Glycobiology*. <https://doi.org/10.1093/glycob/cwad001>
2. Daniloski, D., McCarthy, N. A., **Gazi, I.**, & Vasiljevic, T. (2022). Rheological and structural properties of acid-induced milk gels as a function of β -casein phenotype. *Food Hydrocolloids*, *131*, 107846. <https://doi.org/10.1016/j.foodhyd.2022.107846>
3. Huppertz, T., & **Gazi, I.** (2022). Caseins and casein micelles. In *Understanding and improving the functional and nutritional properties of milk* (pp. 155-185). Burleigh Dodds Science Publishing Limited. <https://doi.org/10.19103/AS.2022.0099.04>
4. **Gazi, I.**, Franc, V., Tamara, S., van Gool, M. P., Huppertz, T., & Heck, A. J. (2022). Identifying glycation hot-spots in bovine milk proteins during production and storage of skim milk powder. *International Dairy Journal*, *129*, 105340. <https://doi.org/10.1016/j.idairyj.2022.105340>
5. **Gazi, I.**, Johansen, L.B., Huppertz, T., 2022. Heterogeneity, fractionation, and isolation. In: McSweeney, P.L.H., McNamara, J.P. (Eds.), *Encyclopedia of Dairy Sciences*, 3rd Ed., Academic Press, Oxford, pp. 881–893. <https://doi.org/10.1016/B978-0-12-818766-1.00278-6>
6. Dingess, K. A., **Gazi, I.**, van den Toorn, H. W., Mank, M., Stahl, B., Reiding, K. R., & Heck, A. J. (2021). Monitoring human milk β -casein phosphorylation and O-glycosylation over lactation reveals distinct differences between the proteome and endogenous peptidome. *International Journal of Molecular Sciences*, *22*(15), 8140. <https://doi.org/10.3390/ijms22158140>
7. **Gazi, I.**, Kim, H., Paterson, A. H., & Huppertz, T. (2020). Measuring viscosity of supersaturated lactose solutions using dynamic light scattering. *International Dairy Journal*, *102*, 104596. <https://doi.org/10.1016/j.idairyj.2019.104596>
8. Huppertz, T., & **Gazi, I.** (2017). Ingredients from milk for use in food and non-food products: from commodity to value-added ingredients. In van Belzen, N. (Ed.), *Achieving sustainable production of milk, Volume 1: Milk*

composition, genetics and breeding (Chapter 3, pp. 121-144). Cambridge, UK: Burleigh Dodds Science Publishing Limited.
<https://doi.org/10.4324/9781351114165>

9. Huppertz, T., **Gazi, I.**, Luyten, H., Nieuwenhuijse, H., Alting, A., & Schokker, E. (2017). Hydration of casein micelles and caseinates: Implications for casein micelle structure. *International Dairy Journal*, 74, 1-11.
<https://doi.org/10.1016/j.idairyj.2017.03.006>
10. Settachaimongkon, S., van Valenberg, H. J., **Gazi, I.**, Nout, M. R., van Hooijdonk, T. C., Zwietering, M. H., & Smid, E. J. (2016). Influence of *Lactobacillus plantarum* WCFS1 on post-acidification, metabolite formation and survival of starter bacteria in set-yoghurt. *Food Microbiology*, 59, 14-22. <https://doi.org/10.1016/j.fm.2016.04.008>
11. Huppertz, T., **Gazi, I.** (2016). Lactose in dairy ingredients: Effect on processing and storage stability. *Journal of Dairy Science*, 99(8), 6842-6851.
<https://doi.org/10.3168/jds.2015-10033>
12. Huppertz, T., & **Gazi, I.** (2015). Milk protein concentrate functionality through optimised product-process interactions. *New Food magazine*, 18(1), 12-17.
13. **Gazi, I.**, & Huppertz, T. (2015). Influence of protein content and storage conditions on the solubility of caseins and whey proteins in milk protein concentrates. *International Dairy Journal*, 46, 22-30.
<https://doi.org/10.1016/j.idairyj.2014.09.009>
14. **Gazi, I.**, & Huppertz, T. (2015). Casein-whey protein interactions for optimizing milk protein functionality. *Agro FOOD Industry Hi Tech, Food Proteins/Pre-Probiotics*, 26(2), 11-14.
15. Crowley, S. V., Desautel, B., **Gazi, I.**, Kelly, A. L., Huppertz, T., & O'Mahony, J. A. (2015). Rehydration characteristics of milk protein concentrate powders. *Journal of Food Engineering*, 149, 105-113.
<https://doi.org/10.1016/j.jfoodeng.2014.09.033>
16. Crowley, S. V., Boudin, M., Chen, B., **Gazi, I.**, Huppertz, T., Kelly, A. L., & O'Mahony, J. A. (2015). Stability of milk protein concentrate suspensions to in-container sterilisation heating conditions. *International Dairy Journal*, 50, 45-49. <https://doi.org/10.1016/j.idairyj.2015.05.009>

17. **Gazi, I.**, Vilalva, I. C., & Huppertz, T. (2014). Plasmin activity and proteolysis in milk protein ingredients. *International Dairy Journal*, 38(2), 208-212. <https://doi.org/10.1016/j.idairyj.2013.11.012>
18. Crowley, S. V., Megemont, M., **Gazi, I.**, Kelly, A. L., Huppertz, T., & O'Mahony, J. A. (2014). Heat stability of reconstituted milk protein concentrate powders. *International Dairy Journal*, 37(2), 104-110. <https://doi.org/10.1016/j.idairyj.2014.03.005>
19. Crowley, S. V., **Gazi, I.**, Kelly, A. L., Huppertz, T., & O'Mahony, J. A. (2014). Influence of protein concentration on the physical characteristics and flow properties of milk protein concentrate powders. *Journal of Food Engineering*, 135, 31-38. <https://doi.org/10.1016/j.jfoodeng.2014.03.005>

Acknowledgements

Acknowledgements

I embarked into this journey towards a PhD degree as an enthusiastic, starry-eyed student, excited for a new adventure, naïvely thinking I know exactly what I signed up for. It turns out I had no idea... Working in a world-leading group, surrounded by exceptionally knowledgeable scientific minds specialised into a vast array of fields, it has at times left me feeling as a small fish in a big pond. At the same time, it has been inspiring and motivating to work with such enthusiastic and talented people, and I have learned in these past four years and several months so much more than I thought I would. Overall, it has been a very rewarding experience that contributed to making me the person and professional I am today, and for that I am grateful.

Albert, I am thankful for you offering me the opportunity to complete this PhD project under your supervision. Hecklab is a unique place where constantly-developing state-of-the-art technology meets leading-expertise, pushing the limits of anyone who is fortunate enough to join the group. Had I joined any other group, I don't believe I could have gained the knowledge, insights, experience and interests that I have now. You always told me to make the PhD project my own, and so I did. Working with you has helped me to become more independent, build my confidence, and open my mind to new perspectives. I am grateful for the opportunity to continue my work under your supervision in the coming couple of years after my PhD.

Thom, I still remember that day when I first came to NIZO for my MSc internship interview. I was dressed to impress, I arrived 30 min late after getting on the wrong bus and being lost in Ede, and I was entirely certain I will not get that position anymore. Despite it all, I still managed to make a good impression that day, and more than 10 years later, here we are still collaborating. Throughout these years, you've given me the room, the opportunities, and the nudges I needed to develop myself. You still believed in me at times when I didn't believe in myself, and you've put a great deal of trust in me with the responsibilities you've extended me. No matter how busy, you always had a moment for me when I needed your opinion. Thank you for your continued support.

I am thankful for the NWO SATIN consortium (project 731.017.202) through which my PhD project was funded. To all the consortium members, LUMC, FrieslandCampina, DSM, Roche, and my fellow PhD students, Dario and Guusje: our biyearly scientific meetings have given me food for thought every time. To my close collaborators from FrieslandCampina, Martine, Jan, André and Thom: thank you for providing all the samples I used in my experiments, but above all thank you for all the discussions, feedback and support. I am grateful for extending me the

opportunity to continue our collaboration beyond my PhD project, and I'm looking forward to the future challenges.

When someone asks me about Hecklab, the first words that come to mind describing it are: it functions as a well-oiled machine! This reflects on the hard work of staff, technicians, PhD students and post-docs that ensure training of every new student and employee, management of the different laboratories, and maintenance and operation of the lab equipment and IT systems. Arjan, you patiently taught me what every bit of hardware in different types of mass spectrometers does, and you were always there every time I needed a helping hand. Dominique, you helped me start up on various protocols in the wet lab, and also helped me integrate in the group in my early time here. Thank you, Mirjam, for always making sure we have the materials we need to carry out our work, and thank you for helping me the many times that I asked you to. Vojta, thank you for the support you've given me in my first couple of years. Sem, I'll never be able to rival your skills in R scripting, or your passion and knowledge for top-down mass spectrometry. I'm grateful for the support you've given me in analysing my intact milk proteins. Geert, thanks for always readily sorting my PC out every time turning it off and on again wouldn't do the trick. Corine, thank you for the planning and organising that ensures everything runs smoothly.

Karli, I kind of wish I'd started doing glycoproteomics since my first year of PhD. But it is still better late than never. You are not only very knowledgeable, but you are also very good at sharing your knowledge and motivating others. Your enthusiasm is contagious, and working with you is inspiring. I am now happy to officially be a member of the glyco group, and I am thankful to my glyco colleagues for all the interesting and insightful discussions, and also for all the fun and laughter. Also, a special thank you to Joshua, for checking and correcting the Dutch translation of my thesis summary.

To all my office mates, past and present: you've made this a fun place to come to work in. Thank you, Suzy, Kelly D., Fleur, Jodie, Kelly G., Fujia and Rensong, for your friendship, serious discussions, light-hearted chit-chat, and all the energising breaks we took together. Kelly D., my fellow milk person, I appreciate all the conversations that we had and the work that we collaborated on, and I admire your perseverance and ambition. Kelly G., we had a similar vibe and we were often on the same page with each other; although seemingly short, I appreciate the time we spent as colleagues, our coffee breaks and the walks through the botanical gardens. Fujia, your name speaks for itself, but you must be one of the nicest and happiest people I have met. It's always a pleasant day if you are also there. Rensong, you are full of

Acknowledgements

enthusiasm and curiosity, and really nice and friendly, I am happy to be sharing the office with you.

Thank you, Inês, for bringing my vision to life in the wonderful design of this booklet!

To all those who have helped and supported me along the way, and whose names I haven't mentioned: thank you!

To my family: *tati*, Elza, you have always been supportive of my decisions, you did the best that you could to help me get where I thought my road was taking me. Adel, my little sister, we are so different and so far apart from each other, and yet so similar and so close; thank you for hearing me out every time I needed a listening ear, and thank you for all the friendly banter to make light of difficult situations. To all my grandparents, *anni*, *ababai*, *mama*, *tata*: I couldn't have asked for more than all the love that you gave me; you supported me morally, financially, and you fuelled my ambition, always believing in me; I am sorry that not all of you are here anymore to share this moment with me and to see what I have achieved with your help.

To my Dutch family: Frans and Marianne, I could not ask for better parents-in-law. You welcomed me with open arms, you always make me feel at home, and you are there and ready to help no matter what I need. And for this I cannot thank you enough.

And finally, lieve Bart: you are my best friend, you are who is there for me all the time, in good times and in bad, you put up with me, you root for me, you cook for me, and most importantly you make me laugh. Never change, I love you.