

# Glycoproteomics incursions into the realm of proteoglycans

Alejandro Gómez Toledo

Department of Clinical Chemistry  
and Transfusion Medicine,  
Institute of Biomedicine  
Sahlgrenska Academy at the  
University of Gothenburg  
Gothenburg, Sweden, 2017



UNIVERSITY OF  
GOTHENBURG

Cover illustration by Alejandro Gómez Toledo

*Glycoproteomics incursions into the realm of proteoglycans*

© 2017 Alejandro Gómez Toledo

[alejandro.gomez.toledo@gu.se](mailto:alejandro.gomez.toledo@gu.se)

ISBN 978-91-629-0091-5

Printed in Gothenburg, Sweden 2017

Ineko AB

This thesis is dedicated to the memory of:



Ammi Grahn (1961-2016)

A most inspiring scientific mentor and a beloved friend

– *“...And when the rose is gone, the garden faded,  
you will no longer hear the nightingale.”*

Rumi



# Abstract

The term proteoglycan encompasses a heterogeneous group of heavily modified metazoan glycoproteins that are involved in fundamental biological processes. They are essential for embryonic development and play important roles in tissue organization and cell haemostasis. Proteoglycans are also linked to pathogenesis and modulate key processes related to microbial infection, cancer behaviour and cardiovascular dysfunction. To understand their impact on human health and disease, thorough studies of their structure-and-function relationships are required. However, this has been hampered by technical difficulties mainly related to the structural complexity of their modifying glycosaminoglycan (GAG) chains.

In this thesis, we developed protocols for the purification and structural characterization of chondroitin sulfate (CS) and heparan sulfate (HS) proteoglycans from complex samples. Our approaches were enclosed within a glycoproteomics framework allowing for the simultaneous identification of the core-proteins and the glycan attachment sites. Additionally, they facilitated the characterization of the proteoglycan linkage region. Our workflows entailed multiple enzymatic degradation steps, chromatographic separation and high-resolution tandem mass spectrometry. Finally, we developed *SweetNET*, a bioinformatics platform to cope with the large amounts of data generated from these high-throughput experiments.

In addition to the limited number of known mammalian proteoglycans (less than 50), we identified 21 novel human core proteins modified with CS chains. We found that several pro-hormones carry CS-modifications, defining them as a novel class of proteoglycans. We also identified novel glycan variations of the proteoglycan linkage region, close to the peptide attachment site, which included fucosylation and sialylation. The examination of the small CS proteoglycan bikunin across different human body fluids revealed an unforeseen heterogeneity of its linkage region, especially in urinary samples. In addition, we could determine the exact macromolecular architecture of the bikunin CS-chain within the inter-alpha trypsin inhibitor complex in serum and cerebrospinal fluid. Finally, we identified placental-type GAGs in induced pluripotent stem cells using a recombinant malaria protein probe. These GAGs displayed a stage-specific dependence and were associated with a heterogeneous group of core-

proteins. The extent and biological implications of these findings for basic stem cell biology need further clarification.

Taken together, we have established preparative and analytical protocols as well as bioinformatics tools for the structural characterization of native proteoglycans in complex samples. This led us to identify novel human proteoglycans as well as novel glycosaminoglycan modifications. Finally, we found that the proteoglycan landscape of pluripotent stem cells changes upon differentiation and can be specifically targeted using a unique protein probe.

## Keywords

Proteoglycans, glycosaminoglycans, mass spectrometry, VAR2CSA, iPS cells, bikunin, inter-alpha trypsin inhibitor, bioinformatics, glycopeptides, pro-hormones

# Sammanfattning på svenska

Proteoglykaner omfattar en heterogen grupp av komplext modifierade glykoproteiner vilka är inblandade i grundläggande biologiska processer. De spelar en viktig roll under embryonal utveckling och är avgörande för upprätthållande av flera basala cellfunktioner. Proteoglykaner är kliniskt intressanta för flera olika sjukdomar genom att de reglerar viktiga mekanismer relaterade till infektion, cancer och fettmetabolism. En detaljerad kartläggning av deras struktur-och-funktion samband är nödvändig för att bättre kunna utnyttja deras diagnostiska och terapeutiska potential. Proteoglykaner är modifierade med olika kolhydrater i långa och heterogena s.k. glykosaminoglykan-kedjor vilka alltid utgjort en svår analytisk utmaning vad avser deras strukturella karakterisering.

I denna avhandling har vi utvecklat preparativa och analytiska protokoll för rening och strukturkarakterisering av kondroitinsulfat (CS) och heparansulfat (HS) innehållande proteoglykaner från komplexa prover. Våra glykoproteomik-liknande metoder möjliggör identifiering av bärarproteinet, lokalisering av den exakta positionen av kolhydratkedjan i proteinsekvensen samt strukturell karakterisering av kolhydratkedjans innersta del, den s.k. länregionen. Dessa metoder innefattar flera enzymatiska nedbrytningssteg, kromatografiska separationer och högupplösande tandem masspektrometri (MS/MS). Slutligen utvecklade vi programvaran *Sweet-NET* som en helt ny och generell bioinformatisk plattform för effektiv hantering och tolkning av tiotusentals masspektra erhållna från MS/MS-analys av olika glykoproteiner.

Utöver det begränsade antalet kända däggdjurs proteoglykaner (<50 st) lyckades vi identifiera 21 st nya humana proteiner modifierade med CS-kedjor. Vi fann att flera pro-hormoner kan bära CS-modifieringar vilket gör dem till en ny klass av proteoglykaner. Vi identifierade också nya modifierationer i länregionen av proteoglykanerna, bland annat fukosylering och sialylering. Dessutom kunde vi bestämma den exakta makromolekylära arkitekturen av inter-alfa-trypsin-inhibitor komplexet i serum och i cerebrospinalvätska. Slutligen identifierade vi uttrycket av placenta-typ GAG kedjor hos inducerade pluripotenta stamceller med hjälp av ett rekombinant malariaprotein, VAR2CSA. Detta uttryck var beroende av stamcellernas differentieringsgrad och kunde förknippas med en heterogen grupp av både kända och okända proteoglykaner.

Sammantaget har vi etablerat preparativa och analytiska protokoll samt bioinformatiska verktyg för strukturell karakterisering av nativa proteoglykaner i komplexa biologiska prover. Detta har lett fram till att vi kunnat identifiera helt nya humana proteoglykaner och nya modifieringar av deras glykosaminoglykan kedjor. Slutligen fann vi även att proteoglykaner på pluripotenta stamceller förändras vid differentiering såväl vad avser GAG-kedjorna som deras bärarproteiner.





# List of publications

This thesis is based on the following studies, referred to in the text by their Roman numerals.

- I. Noborn F, **Gomez Toledo A**, Sihlbom C, Lengqvist J, Fries E, Kjellén L, Nilsson J, Larson G. *Identification of chondroitin sulfate linkage region glycopeptides reveals prohormones as a novel class of proteoglycans*. Mol Cell Proteomics. 2015 Jan;14(1):41-9.
- II. **Gomez Toledo A**, Nilsson J, Noborn F, Sihlbom C, Larson G. *Positive Mode LC-MS/MS Analysis of Chondroitin Sulfate Modified Glycopeptides Derived from Light and Heavy Chains of The Human Inter- $\alpha$ -Trypsin Inhibitor Complex*. Mol Cell Proteomics. 2015 Dec;14(12):3118-31.
- III. Nasir W, **Toledo AG**, Noborn F, Nilsson J, Wang M, Bandeira N, Larson G. *SweetNET: A Bioinformatics Workflow for Glycopeptide MS/MS Spectral Analysis*. J Proteome Res. 2016 Aug 5;15(8):2826-40.
- IV. Noborn F, **Gomez Toledo A**, Green A, Nasir W, Sihlbom C, Nilsson J, Larson G. *Site-specific identification of heparan and chondroitin sulfate glycosaminoglycans in hybrid proteoglycans*. Sci Rep. 2016 Oct 3;6:34537.
- V. **Gomez Toledo A**, Pereira MA, Clausen TM, Simonsson S, Salanti A and Larson G. *The expression of placental-type chondroitin sulfate A is associated with a heterogeneous group of CSPGs in human cancer and IPS cells*. Manuscript

Publications not included in this thesis:

Nilsson J, Noborn F, **Gomez Toledo A**, Nasir W, Sihlbom C, Larson G. *Characterization of Glycan Structures of Chondroitin Sulfate-Glycopeptides Facilitated by Sodium Ion-Pairing and Positive Mode LC-MS/MS*. J Am Soc Mass Spectrom. 2016 Nov 21.

Yu J, Schorlemer M, **Gomez Toledo A**, Pett C, Sihlbom C, Larson G, Westerlind U, Nilsson J. *Distinctive MS/MS Fragmentation Pathways of Glycopeptide-Generated Oxonium Ions Provide Evidence of the Glycan Structure*. Chemistry. 2016 Jan 18;22(3):1114-24.

Hedberg C, **Toledo AG**, Gustafsson CM, Larson G, Oldfors A, Macao B. *Hereditary myopathy with early respiratory failure is associated with misfolding of the titin fibronectin III 119 subdomain*. Neuromuscul Disord. 2014 May;24(5):373-9

**Gomez Toledo A**, Raducu M, Cruces J, Nilsson J, Halim A, Larson G, Rüetschi U, Grahn A. *O-Mannose and O-N-acetyl galactosamine glycosylation of mammalian  $\alpha$ -dystroglycan is conserved in a region-specific manner*. Glycobiology. 2012 Nov;22(11):1413-23.

# Contents











Abstract.....	5
Keywords.....	6
Sammanfattning på svenska.....	7
List of publications.....	10
Contents.....	12
Abbreviations.....	14
1. Introduction.....	17
1.1 Finding the coordinates of the thesis.....	17
1.2 “The reluctant mucopolysaccharides”.....	21
1.2.1 Structural diversity of the glycosaminoglycans.....	22
1.2.2 “The missing link”.....	24
1.2.3 The CS/DS and HS linkage region.....	25
1.2.4 Chemical substitutions and regulation of the GAG linking tetrasaccharide.....	26
1.2.5 Structure and biosynthesis of the GAG backbone.....	27
1.2.6 Proteoglycans in health and disease.....	30
1.2.7 Summary.....	33
2. Methodological considerations.....	35
2.1 Analytical challenge.....	35
2.2 Purification of GAGs and proteoglycans.....	36
2.3 Proteoglycan structural analysis.....	37
2.4 Mass spectrometry.....	37
2.5 MS-based GAG analysis.....	40
2.6 Glycoproteomics.....	41
2.7 Glycoproteomics tools for proteoglycan analysis.....	43
3. Aims.....	44

4. Results and discussion.....	46
4.1 Targeting the data .....	46
4.2 Targeting the protein.....	49
4.3 Targeting the glycan .....	56
4.4 Targeting biology .....	58
5. Conclusions .....	62
6. Future Perspectives .....	64
Acknowledgement .....	66
References.....	68

# Abbreviations

B3GalT6	Beta-1,3-galactosyltransferase 6
B4GalT7	Beta-1,4-galactosyltransferase 7
BMP1	Bone morphogenetic protein 1
CHGA	Chromogranin A
ChondABC	Chondroitinase ABC
ChSy-1, -2, -3	Chondroitin synthase-1, -2, -3
CID	Collisional induced dissociation
CS	Chondroitin sulfate
CSA	Chondroitin sulfate A
CSGalNAcT-1	Chondroitin sulfate N-acetylgalactosaminyltransferase 1
CSGalNAcT-2	Chondroitin sulfate N-acetylgalactosaminyltransferase 2
DC	Direct current
DS	Dermatan sulfate
DS-epi1	Dermatan sulfate epimerase 1
DS-epi2	Dermatan sulfate epimerase 2
ECM	Extracellular matrix
ESC	Embryonic stem cell
ESI	Electrospray ionization
EXT-1	Exostosin-1
EXT-2	Exostosin-2
EXTL3	Exostosin-like 3
FGF-2	Fibroblast growth factor 2
Fuc	Fucose
GAG	Glycosaminoglycans
Gal	Galactose
GalNAc	N-acetylgalactosamine
GlcA	Glucuronic acid
GlcNAc	N-acetylglucosamine
HA	Hyaluronic acid
HCD	Higher-energy collisional dissociation
HexA	Hexuronic acid

HexNAc	N-acetylhexosamine
HS	Heparan sulfate
IdoA	Iduronic acid
IE	Infected erythrocytes
iPS cells	Induced pluripotent stem cells
IaI	Inter-alpha-trypsin inhibitor
KS	Keratan sulfate
LC	Liquid chromatography
MALDI	Matrix-assisted laser desorption/ionization
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NDST1	Heparan sulfate N-deacetylase/N-sulfotransferase 1
NeuAc	N-acetylneuraminic acid
OA	Osteoarthritis
PNA	Peanut agglutinin
PSM	Peptide-spectrum match
RF	Radio frequency
SAX	Strong anion exchange
SRGN	Serglycin
UDP	Uridine diphosphate
UEA-1	Ulex europaeus agglutinin I
UTI	Urinary trypsin inhibitor
VEGF	Vascular endothelial growth factor
XT-1	Xylosyltransferase 1
XT-2	Xylosyltransferase 2
Xyl	Xylose
$\Delta$ HexA	4,5-unsaturated hexuronic acid

 N- acetylgalactosamine (GalNAc)	 Glucuronic acid (GlcA)
 N- acetylglucosamine (GlcNAc)	 Iduronic acid (IdoA)
 Galactose (Gal)	 4,5- unsaturated hexuronic acid ( $\Delta$ HexA)
 Mannose (Man)	 N-acetylneuraminic acid (NeuAc)
 Xylose (Xyl)	 Fucose (Fuc)





# 1. Introduction

## 1.1 Finding the coordinates of the thesis

According to the Oxford English Dictionary, glycobiology is defined as the branch of science concerned with the role of sugars in biological systems. A more rigorous definition would probably include:

*“the study of the structure, biosynthesis, biology and evolution of saccharides (sugar chains or glycans) that are widely distributed in nature, and the proteins that recognize them” (Varki & Sharon, 2009).*

At a first glance, these notions may suggest that glycobiology qualifies as a subdiscipline within the much broader field of biochemical research. However, in a strict historiographical sense, the demarcation between these two disciplines is less clear than what it is usually realized. In fact, it could be equally argued that the genesis of modern biochemistry was profoundly shaped by fundamental developments in early carbohydrate research.

Louis Pasteur counts among the European pioneers that embarked into the systematic study of alcoholic fermentation during the 19th century (Fig 1). After several years of experimental research on yeast cells, Pasteur finally arrived to the conclusion that a “vital force” called “ferment” was responsible for the conversion of sugars into alcohols (Pasteur, 1860). This notion was in line with popular ideas about the “vitalistic” nature of the world and found therefore an easy way into the heart of most European scientific circles. However, soon after that, a German chemist by the name of Justus von Liebig started one of the most interesting controversies in the history of modern chemistry: the Liebig-Pasteur dispute (Hein, 1961). This dispute had many angles but it could be summarized as Pasteur holding that fermentation was a biological process (exclusively depending on living organisms) while Liebig claimed that it was a mechanical process (caused by vibrations from the decomposition of organic matter). To make things even more complicated, both chemists had different experimental approaches and different ideas about where fermentation took place in an organism. Later on, and building onto these concepts, Eduard Buchner demonstrated that living cells were not an absolute pre-requisite for the chemical conversion. In a series of seminal reports published in 1897, Buchner cleverly undermined Pasteur’s assumptions by proving that cell-free yeast extracts were equally suffi-

cient to induce sugar fermentation (Buchner, 1897). However, Buchner's results also pointed to the existence of a substance, similar to the Pasteurian "ferment" that was required for the reaction to take place. For the modern reader, the far-reaching consequences of these specialized controversies may be difficult to grasp but from a philosophical point of view, the impact of Buchner's research should not be underestimated. Buchner's results provided some of the first experimental counter-evidences to vitalistic hypotheses claiming that living organisms were fundamentally different to non-living entities. The mysterious substances required for fermentation were eventually denoted enzymes but their biochemical nature remained for long unknown.



*Figure 1. Louis Pasteur was a scientific pioneer in the study of fermentation, in which yeast cells break down sugars into ethanol and carbon dioxide. "Louis Pasteur in his laboratory", reproduction of a painting from 1885 by the Finnish artist Albert Gustaf Aristides Edelfelt.*

Equally enigmatic was the nature of the interaction between enzymes and their target substances. Reflecting upon the fact that yeasts fermented the D- but

not the L-form of the monosaccharides mannose, glucose and galactose, Emil Fischer hypothesized that structural determinants are required for the enzymes to act on their substrates:

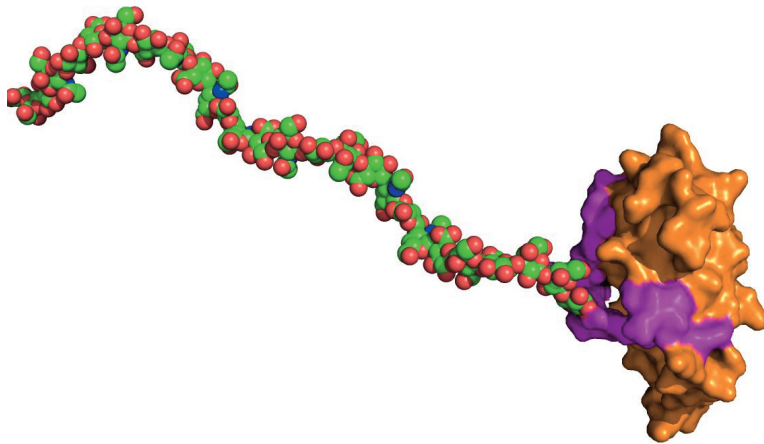
*“To use a metaphor, I would like to say that enzyme and glucoside have to fit together like lock and key in order to exert a chemical effect on each other” (E. Fischer, 1894).*

Fischer’s image of lock and key is a beautiful scientific metaphor that has paved the way for our current understanding of basic enzymology. These few examples among many others, serve to illustrate that the intellectual genealogies of biochemistry and glycobiology are indeed deeply intertwined.

The presence of cell surface glycosylation is a common denominator in virtually all living cells (Varki, 2011). In a sense, it can be compared to the universality of having a genetic code or to the fact that biological membranes are build-up of lipids. Most of the focus on glycans in biological systems has traditionally been put on metabolism. However, the accumulated data to this day links glycosylation to every fundamental cellular process. Glycans may appear in multiple cellular locations, and not only as free sugars but more often as glycoconjugates (i.e. covalently linked to other biomolecules as glycoproteins, glycolipids etc.), which makes them versatile tools for mediating, regulating and fine-tuning cellular functions. As expected, their biological roles may then span from subtle to essential for the development and survival of living organisms. In addition to that, glycosylation is by no way devoid of medical interest. Disturbances in glycan biosynthesis and catabolism are the main causatives of a plethora of pathophysiological processes (Freeze et al., 2014; Hennet & Cabalzar, 2015). Glycans constitute the first cellular barrier encountered by microbial pathogens during infection and their potential as therapeutic targets has been widely acknowledged (Fuster & Esko, 2005). Having said that, how does it come that the study of such important biomolecules has become the subject of a surprisingly small scientific community? What are the reasons for singling out glycobiology as an independent discipline? (There is not such a thing as proteo-biology or lipido-biology). The answer to those questions is not straightforward but one important clue lies in the nature of the subject itself (Lindahl, 2014; Roseman, 2001). Glycans (and glycoconjugates) are without exaggeration the most diverse, structurally complex, and analytically challenging biomolecules of the entire biological realm.

Glycans are secondary gene products and their biosynthesis is a well-orchestrated process. In general, glycan biosynthesis is considered to take place in a non-template driven fashion, which is dependent on the interplay of multiple glycosyltransferases, glycosidases, and other glycan modifying enzymes. It also

requires the presence of activated sugar nucleotides, transporters and chaperones located along the secretory pathway in the endoplasmic reticulum and Golgi compartments. Compared to other biomolecules such as proteins or nucleic acids, glycans are composed of a larger number of building blocks called monosaccharides that can be connected to each other at different positions and with different linkage configurations. Glycans might be linear or branched and additional chemical substitutions such as acetylation, sulfation or phosphorylation are common, which exponentially increases the diversity of glycan structures in nature. The presence of glycoconjugates adds new levels of complexity and functionality. In many cases, the glycan moiety accounts for a significant part of the total molecular weight of a glycoconjugate and can play a major role in its biosynthesis, processing, location, trafficking and context-dependent function. Finally, most glycans enjoy a tremendous flexibility in aqueous solution due to their intrinsic structural properties.



*Figure 2. The proteoglycans are complex glycoconjugates composed of a core protein and one or more acidic glycosaminoglycan chains. The large sizes and distinct chemical nature of the glycan chains have a great impact on the overall physicochemical properties of the proteoglycans. In this picture, the crystal structure of the small human proteoglycan bikunin (PDB: 1BIK) has been depicted in orange using a surface layout. A twenty amino acid long stretch containing the GAG-attachment at the very N-terminus of the protein is colored in pink. A seventeen monosaccharide long CS chain (sphere representation) was modelled and attached to the core-protein. Picture was created using PyMol.*

This thesis is devoted to the study of a special class of glycoconjugates, the proteoglycans (Fig 2), which are the carriers of the largest and most structurally complex of all glycan modifications, the glycosaminoglycans. In spite of their ubiquitous presence in every cell membrane and in the extracellular matrix (ECM), advances towards deciphering the relationships between the structure and function of the native proteoglycans have been exceptionally slow-paced. In fact, obtaining the complete structure of an intact heparan sulfate proteoglycan, still an unrealized dream, including all of its substitutions, the glycosaminoglycan attachment site on the protein and the glycan domain architecture, is possibly one of the most difficult structural challenges in modern biochemistry.

## 1.2 “The reluctant mucopolysaccharides”

The nature of the amorphous “ground substance” that surrounds the cells in connective tissues received great attention during the 19th century. Most of the earlier descriptions of this substance were restricted to anatomic or microscopic examinations. Nevertheless, histologists became quickly acquainted with the anionic character of its major chemical components, a property that was exploited to develop several staining procedures for microscopic visualization. It was not until the development of novel extraction procedures when the presence of a distinct family of related acidic polysaccharides, now denoted glycosaminoglycans, was finally demonstrated in cartilage (G. a. B. Fischer, C., 1861). The biochemical analysis of these polysaccharides, at that time known as “mucopolysaccharides”, constituted the main target for several pioneering studies that laid the basis for our understanding of the composition, structural diversity and physicochemical properties of the glycosaminoglycans. In this context, the unique work of Karl Meyer and his co-workers stands out as an extraordinary scientific accomplishment entailing, among other things, the discovery of hyaluronic acid (Palmer, 1934), dermatan sulfate (K. Meyer, and Chaffee, E., 1941), keratan sulfate (K. Meyer, Linker, A., Davidson, E. A., and Weissmann, B, 1953) and the initial differentiation between multiple classes of chondroitin sulfate (K. Meyer, Davidson, E. A., Linker, A., and Hoffman, P, 1956). A different line of research with a primary focus on blood coagulation led to the isolation of heparin (McLean, 1916), and eventually, to the discovery of another important class of glycosaminoglycans, the heparan sulfate (Jorpes, 1948).

## 1.2.1 Structural diversity of the glycosaminoglycans

The term glycosaminoglycan (GAGs) denotes a group of linear polysaccharides whose repetitive building blocks consist of an amino sugar and an uronic acid (or a galactose). GAGs are usually divided into four major groups: chondroitin /dermatan sulfate (CS/DS), heparan sulfate (HS), hyaluronic acid (HA) and keratan sulfate (KS); based on their monosaccharide compositions and structures (Fig 3). Hitherto, GAG expression has only been demonstrated in metazoan organisms. HA has a more recent evolutionary history and appears restricted to vertebrates. Some species-specific patterns in the fine structure of GAGs have also been described. For example, the CS chains synthesized by invertebrates such as *Drosophila melanogaster* or *Caenorhabditis elegans* are mostly devoid of sulfate substitutions, in contrast to the vast majority of CS chains produced by vertebrate organisms. However, very recent findings have started challenging these widely-held notions (Dierker et al., 2016; Izumikawa et al., 2016).

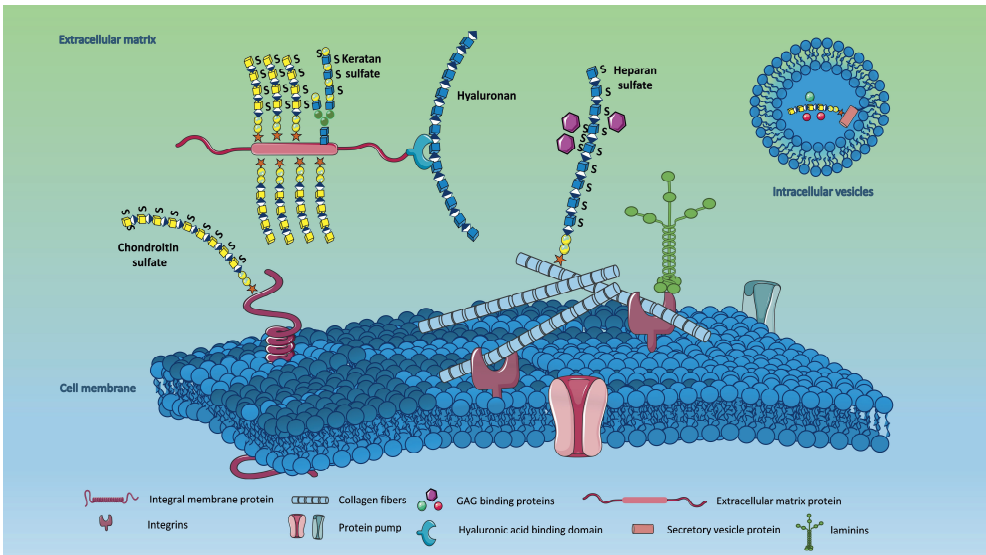


Figure 3. Native glycosaminoglycans display a broad structural diversity. They are traditionally divided into 4 major groups. Chondroitin/ dermatan sulfate, heparan sulfate, and keratan sulfate appear glycosidically bound to a polypeptide chain in the form of a proteoglycan. Hyaluronic acid is secreted as a free glycan. Glycosaminoglycans are present in different locations such as the cell membrane, the extracellular matrix, and inside secretory vesicles.

The main repetitive disaccharide unit of CS is (GalNAc $\beta$ 4GlcA $\beta$ 3) while HS chains are predominantly built of (GlcNAc $\alpha$ 4GlcA $\beta$ 4). Enzymatical elongation of these basic units can generate very large structures and GAG chains of 20-50 kDa are common. A high degree of further variation is achieved by chemical substitutions with sulfates (and sometimes phosphate) groups attached to different monosaccharide residues and at different positions. A certain level of epimerization of glucuronic acid (GlcA) into iduronic acid (IdoA) is also present in CS giving rise to the related polysaccharide DS. IdoA is also present in HS structures within modular distributions of sulfated and epimerized domains. Both CS/DS and HS are long and polydisperse polysaccharides. The carboxylic groups of the uronic acids and the sulfate substituents confer them with a high density of negative charges that helps to shape their acidic nature. During biosynthesis, both CS/DS and HS become glycosidically linked to a polypeptide chain via a common tetrasaccharide structure. A core protein carrying a GAG chain is thus, per definition, a proteoglycan. Due to the large sizes of CS/DS and HS chains, the properties of the GAGs tend to dominate the physicochemical properties of the proteoglycans. On the other hand, the relevance of the core-proteins for particular functions cannot be dismissed and is often context-dependent. Finally, heparin is a subtype of HS that is only produced by connective-tissue-type mast cells on a specific proteoglycan called serglycin. Compared with HS, Heparin chains undergo more extensive epimerization and sulfation, including rare sulfation events that are essential for its well-known anticoagulant activity (Lever & Page, 2002).

The disaccharide composition of HA (GlcNAc $\beta$ 4GlcA $\beta$ 3) is similar but not identical to CS/DS and HS. HA is devoid of further substitutions such as sulfate groups and its biosynthesis does not occur in the secretory pathway. HA chains are directly synthesized by a group of integral membrane proteins called hyaluronan synthases and extruded via ABC transporters through the cell membrane into the extracellular space. The HA chains appear as free glycans or covalently linked to the heavy chains of the inter-alpha trypsin inhibitor (Chen et al., 1996).

KS constitutes a distinct subclass of GAGs composed of sulfated poly-N-acetyllactosamine (GlcNAc $\beta$ 3Gal $\beta$ 4) chains, which are in principle identical to those found in other common glycoproteins. These GAG chains are also distinguished by being synthesized onto N-glycan core structures (KS-I) or O-glycan core 2 structures (KS-II). Keratan sulfate can be further decorated with fucose and sialic acids and both the GlcNAc and the Gal residues may serve as targets for sulfate substitutions.

The large diversity among GAG structures, the presence of different GAG types on the same carrier protein (hybrid proteoglycans) and the serendipitous

history of their discovery in different labs and at different time points, has largely contributed to the difficulties in developing a standardized system for GAG (and proteoglycan) nomenclature and classification. In fact, there is recent experimental evidence for proteins carrying GAG-like polysaccharides but their classification as proteoglycans is disputable, despite showing striking similarities regarding saccharide compositions and structure (Inamori et al., 2016; Yoshida-Moriguchi & Campbell, 2015). On the other hand, the structural and biosynthetic reasons for certain proteins to acquire proteoglycan status are not always clear and the demarcation between a glycoprotein and a proteoglycan can sometimes become confounding and rather artificial. More importantly, what may appear as a theoretical issue has been paralleled by the establishment of a niche within glycobiology research exclusively dedicated to the study of GAGs and proteoglycans, singling them out from the other glycoproteins. The justification for this development deserves historical, structural, and methodological considerations. For example, the discovery and structure determination of the major GAG classes were rapidly accomplished during the first decades of the 20th century but it was not until 1958 when their proteinaceous nature was fully realized (Muir, 1958). That means that a great part of the major methodological and theoretical challenges of the GAG field were already conceptualized (but of course not completely resolved) by the time when their protein counterparts were finally discovered. One of the main points of this thesis is that the transfer of knowledge and technical experience from basic glycoprotein research, such as the development of glycoproteomics technologies, can generate new insights into GAG and proteoglycan structure and function.

### 1.2.2 “The missing link”

The structural characterization of GAGs was accelerated by the development of isolation and precipitation methods for their extraction from different tissues. Extensive proteolytic digestion was a common step during early GAG extraction procedures. The invariable presence of amino acids in these preparations led eventually to the hypothesis that GAG chains might be tightly associated with a protein component. The first report of such a covalent carbohydrate-protein link was based on the finding that serine was exceptionally enriched in CS preparations from hyaline cartilage after proteolytic digestion (Muir, 1958). Other amino acids were also overrepresented, particularly glutamic acid, proline and glycine. The well-known alkali lability of GAG chains added to these observations by specifically implicating the hydroxyl group of serine as the linking group. In a number of elegant studies, Lindahl and Rodén finally solved the



complete structure of the linkage region in 1966 and showed that both CS/DS and HS proteoglycans share the same tetrasaccharide precursor (GlcA $\beta$ 3Gal $\beta$ 3Gal $\beta$ 4Xyl $\beta$ 1-O-) onto which the GAG chains are further extended (Helting & Roden, 1968; Lindahl, 1966, 1968; Lindahl & Roden, 1966). The linking monosaccharide turned out to be an aldopentose (five-carbon sugar) called xylose. Xylose is widely found across the plant kingdom where it constitutes the main component of the xylan group of hemicelluloses that build up the plant cell walls. Otherwise, xylose is a rare monosaccharide in vertebrate glycans although some notable exceptions have been described (Haltom & Jafar-Nejad, 2015).

### 1.2.3 The CS/DS and HS linkage region

The biosynthesis of CS/DS and HS chains takes place in the secretory pathway starting with the stepwise assembly of the linkage region. This tetrasaccharide is synthesized by the sequential addition of individual monosaccharides rather than being transferred “en bloc” as a precursor oligosaccharide. The xylose unit is always attached to the hydroxyl group of specific serine residues through a beta glycosidic linkage. Similar to other biological glycosylation reactions, the transfer of the linking xylose to the polypeptide requires the presence of a donor sugar nucleotide (UDP-xylose). The first reports of xylose transfer to endogenous protein acceptors dates back to the mid-1960s making the peptide- O-xylosyltransferase one of the first glycosyltransferases to be described (Wilson, 2004). In vertebrates, two active xylosyltransferase isoforms have been reported so far (XT-1 and XT-2). In addition to their slightly different tissue distributions, the analysis of their enzyme activities has revealed that both enzymes can transfer xylose to similar peptide acceptors but with different efficiency (Roch et al., 2010).

Given that transfer of xylose is a rate-limiting step in the biosynthesis of GAGs, the amino acid preferences of the xylosyltransferases have deserved considerable attention. Despite some early discrepant results pointing to the possibility of threonines becoming xylosylated (Mann et al., 1990), most of the accumulated data *in vitro* and *in vivo*, indicates that serine residues are the true targets for the XT-1 and XT-2 activities. Confirming this notion, the presence of endogenous proteoglycans carrying GAG chains at threonine residues has so far not been described. In addition, a consensus sequence for GAG attachment has been formulated (Brinkmann et al., 1997; Esko & Zhang, 1996; L. J. Zhang et al., 1995). GAGs are usually found attached to a Ser-Gly dipeptide flanked by a cluster of acidic amino acid residues. The assembly of HS chains in particular,

occurs preferentially in region with repetitive Ser-Gly sequences (L. J. Zhang et al., 1995). This may indicate that, in certain cases, the xylosyltransferases can act in a processive fashion. Interestingly, this “sequon” has been confirmed in later studies but a note of caution is pertinent. The primary structure of the acceptor does not seem to be the only determinant because the protein conformation has also an impact on the efficiency of transfer. The well-recognized existence of “part-time” proteoglycans raises additional problems given that the mere presence of a GAG sequon does not necessarily imply that the site will always be glycosylated. Last but not the least, compared to other classes of protein glycosylation, the number of known naturally occurring GAG attachments sites is remarkably scarce. Novel methodologies, including the strategies developed in this thesis, are expected to increase our knowledge of the true endogenous targets for the xylosyltransferases and thereby, of their amino acid preferences.

After the transfer of the xylose moiety, the assembly of the linkage region is completed by the addition of two Gal and one GlcA. The transfer of the Gal moieties is catalysed by two independent galactosyltransferases: B4GALT7 (GalT-I) and B3GALT6 (GalT-II). In turn, the GlcA is transferred by the action of B3GAT3 which is not involved in the extension of the repetitive disaccharide units of the CS or HS backbone and it is therefore classified as having GlcAT-I activity. The attachment of the fifth monosaccharide is a key checkpoint that determines if CS (addition of GalNAc $\beta$ 4) or HS (addition of GlcNAc $\alpha$ 4) will be synthesized. The presence of a specific set of dedicated enzymes and a well-defined structure for the linkage region indicates that its assembly could be a potential “hot spot” for regulation. On top of that, CS and HS are not functionally equivalent, despite sharing identical linkage regions, which also raises questions about the regulatory determinants for this biosynthetic bifurcation. Notably, the linkage region tetrasaccharide is the target for several substitutions; some of them having a regulatory function.

#### 1.2.4 Chemical substitutions and regulation of the GAG linking tetrasaccharide

Previous structural studies of the GAG-protein linkage have revealed the presence of several substitutions. Phosphorylation at the C-2 position of the linking xylose is probably the best studied of these modifications and occurs both on CS/DS and on HS chains. Phosphorylation is catalysed by *FAM20b*, a sugar kinase located in the secretory pathway (Koike et al., 2009). This event is proposed to occur transiently during elongation from Gal-Xyl to Gal-Gal-Xyl followed by rapid dephosphorylation after the addition of the GlcA residue

(Moses et al., 1997). Bringing further support to that notion, the existence of a 2-phosphoxylose phosphatase has recently been reported (Koike et al., 2014). The C-2 phosphorylation is known to modulate the assembly of the linkage region by dramatically increasing the activity of GalT-II (Wen et al., 2014). A certain role in regulating the GlcAT-I activity has also been suggested as this glucuronyl-transferase has an increased preference for phosphorylated substrates (Gulberti et al., 2005; Tone et al., 2008). Cells lacking *FAM20b* are unable to extend the tetrasaccharide linkage region and produce instead immature GAG chains, capped with a sialic acid modification (Neu5Aca2,3Gal $\beta$ 4Xyl $\beta$ -O-). Intriguingly, the phosphorylated linkage structures do not support further polymerization *in vitro* (Koike et al., 2014). At the same time, mature phosphorylated GAGs have been identified in several cell systems at substoichiometrical levels. If this presence is due to incomplete phosphorylation or dephosphorylation is currently unknown. It could be possible that *in vitro* assays do not fully recapitulate the true biosynthetic conditions. It is also likely that other factors such as the core protein dependence or differences between cellular systems might be important.

Contrary to the 2-O-phosphorylation, sulfate substitutions at the C-4 and/or C-6 positions of the Gal residues have been observed in CS/DS but not in HS chains. Therefore, it has been suggested that Gal sulfation might be involved in priming the linkage region towards CS/DS biosynthesis. In support for such a role, the sulfation of the Gal residues increases the initiation activity and catalytic efficiency of the CSGalNAcT-1, a critical enzyme that transfers GalNAc onto the linkage region and directs the biosynthesis towards CS/DS (Gulberti et al., 2012). In addition to that, the amino acid sequence of the protein acceptor seems also to be critical (Izumikawa & Kitagawa, 2015). Unfortunately, the mechanistic relationships between the specific GAG attachment site, the modifications of the linkage region and the final GAG-type have not been fully elucidated. A more systematic analysis of all of these modifications (as well as the discovery of potential novel ones) is thus required to understand the structural diversity of the proteoglycan linkage region and its regulatory role during GAG biosynthesis.

### 1.2.5 Structure and biosynthesis of the GAG backbone

#### *The CS/DS backbone*

After completion of the linkage region, the CS/DS backbone is synthesized by the alternate addition of GlcA and GalNAc residues. As mentioned before, the transfer of the first GalNAc is mediated by the action of the CSGalNAcT-1. Another enzyme, CSGalNAcT-2 is primarily involved in the regulation of the chain

length and/or the number of chains (Izumikawa et al., 2011). At least three other enzymes catalyze the polymerization of the CS/DS backbone: ChSy-1, -2 and -3 (chondroitin synthase-1, -2 and -3). These enzymes possess dual catalytic activity (GlcAT-II and GalNAcT-II) but cannot synthesize the chondroitin chain by themselves. Nevertheless, co-expression of any of two of these proteins or their individual expression together with the chondroitin polymerization factor (ChPF), leads to effective polymerization. The length of the resulting chains varies then depending on the enzyme combination (Izumikawa et al., 2008). Another important modification of the CS/DS chains is the epimerization of GlcA into IdoA. This reaction occurs at the polymer level and not by conversion of UDP-GlcA into UDP-IdoA as it was originally proposed. Two epimerases, DS-epi1 and DS-epi2 have been described in humans (Malmstrom et al., 2012)

Sulfation of the CS/DS backbone is probably the most important modification of these GAG chains. The sulfation patterns of a given CS/DS chain are usually instrumental for their ability to engage protein ligands and mediate biological functions. They are equally important for the ability of CS/DS chains to bind water and create hydrated matrices that absorb compressive loads in cartilage. Sulfation can occur on the C-4 and/or the C-6 positions of the GalNAc as well as on the C-2 position of the GlcA/IdoA residues. CS/DS disaccharide types are classified based on their sulfation patterns (Fig 4).

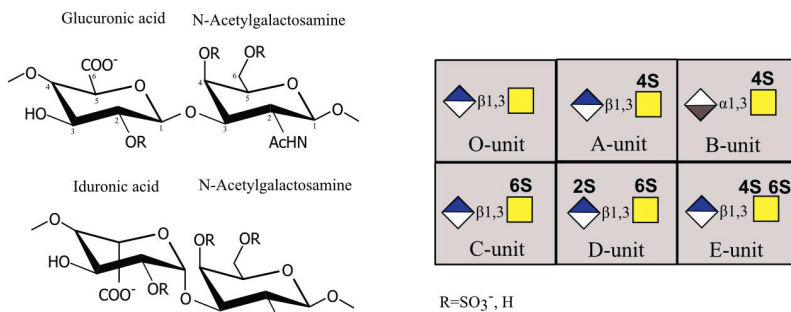


Figure 4. CS/DS disaccharides are usually classified based on their sulfation patterns and uronic acid stereochemistry. Some of the most common examples are shown here.

To date, seven CS/DS specific sulfotransferases have been characterized (Fukuta et al., 1995; Kusche-Gullberg & Kjellen, 2003). All of them utilize the same

activated sulfate donor, 3'-phosphoadenosine 5'-phosphosulfate (PAPS) for the sulfation reaction.

### *The HS/ Heparin backbone*

At least three enzymes, EXT1, EXT2 and EXTL3 are known to be involved in the polymerization of the backbone of HS/Heparin polysaccharides. EXTL3 is proposed to be the main enzyme catalyzing the addition of the first GlcNAc onto the tetrasaccharide linkage. This is a key step that commits the nascent GAG chain towards HS biosynthesis. After initiation, the EXT1/EXT2 polymerase complex starts the transfer of alternating GlcNAc and GlcA onto the growing polymer. In similarity to CS chains, the HS backbone serves as the target for several modifications. *N*-deacetylation/*N*-sulfation of the GlcNAc is mediated by the action of several NDST enzymes, leading to a typical distribution of distinct *N*-sulfated (NS), *N*-acetylated (NA) and hybrid (NS/NA) domains. The GlcNAc units may also undergo further sulfate substitution at position C-3 and/or C-6, a process that is mediated by at least three 6-*O*-sulfotransferases and seven 3-*O*-sulfotransferases. Epimerization of GlcA into IdoA is catalyzed by a single C-5 epimerase that it is not related to the CS epimerases. Furthermore, the great majority of the IdoA units become sulfated by a 2-*O*-sulfotransferase.

In general, discrete patterns of sulfate and uronic acid epimers give rise to binding sequences for interaction with different ligands. These motifs constitute the structural basis for HS-mediated biological functions. Interestingly, despite the lack of a template, HS biosynthesis is known to be tightly regulated during development (Poulain & Yost, 2015). More importantly, several studies have also suggested the existence of tissue-specific HS compositions, which seem to imply that the fine structures of HS chains are not randomly generated (Kato et al., 1994; Shi & Zaia, 2009). To explain these observations, several regulatory models have been proposed, including the potential existence of the HS biosynthetic machinery as multiprotein complexes in the Golgi (GAGosomes) (Esko & Selleck, 2002). A “coding and decoding” hypothesis has also been suggested where certain primary sulfation events (sulfate code) can be recognized by enzymes involved in downstream modification of the HS chains (X. Zhang et al., 2016). This is an area of intense research.

### 1.2.6 Proteoglycans in health and disease

Extensive research on the biological roles of GAGs (and proteoglycans) has shed new light on their crucial and sometimes unexpected involvement in numerous biological processes. Due to the original development of GAG research, two main functions were initially ascribed to these polysaccharides. Firstly, a structural role in connective tissue organization was rapidly acknowledged, particularly in cartilage. Additionally, an important role in controlling blood homeostasis was also realized due to the discovery of heparin and its anticoagulant properties.

Proteoglycans are now known to constitute the major structural component of the ECM of cartilaginous tissues. Cartilage ECM is primarily composed of aggregates of HA, aggrecan and the link protein. Together, these molecules form a gel-like matrix that absorbs compressive mechanical loads through water re-sorption and desorption. Aggrecan is a high molecular weight proteoglycan that contains around 100 GAG chains (Kiani et al., 2002). The GAG chains are mostly of the CS type but the presence of KS has also been demonstrated. Altered degradation of aggrecan is a hallmark of progressive and degenerative joint disorders such as osteoarthritis (OA), which leads to deterioration of the cartilage and fibrosis of the synovium and the joint capsule (Huang & Wu, 2008).

Heparin is a highly-modified variant of HS attached to serglycin and stored in the secretory granules of mast cells from connective tissues. The endogenous physiological function of heparin is still largely unknown. On the other hand, its potent anticoagulant activity has been readily exploited in clinical settings to prevent venous thrombosis and pulmonary embolism as well as for the management of coronary artery disease (Hirsh et al., 2001). The anticoagulant activity of heparin is ascribed to a well-defined pentasaccharide that binds to antithrombin and thrombin in a ternary complex (Li et al., 2004). This binding leads to inactivation of thrombin and factor Xa, and thus prevents fibrin formation and thrombin-mediated activation of blood platelets.

New additional knowledge about the biological importance of GAGs has been acquired through lessons from relatively rare genetic disorders where mutations in the GAG biosynthetic machinery have been reported (Mizumoto et al., 2013). Accumulating data clearly indicates that GAGs are essential for normal bone development and for the maintenance of skin integrity in humans. In addition, defects in GAG catabolism are the main cause of a number of inherited lysosomal storage disorders collectively known as mucopolysaccharidoses (Muenzer, 2011). Advances in genetic manipulation techniques and new experimental models have enabled the exploration of additional roles for these versatile molecules in both health and disease. Surprisingly, the data to date seems to implicate GAGs, in a complex manner, in a plethora of fundamental processes such

as development, stem cell biology, immunity, wound healing as well as in pathogen infection and cancer. As a comprehensive covering of these areas is clearly out of the scope of this thesis, only a few examples will be selected to illustrate general principles related to GAG biology and pathogenesis. Some of these examples will also be revisited in the next sections in the light of the specific results generated during this thesis.

### *Perlecan in cancer and angiogenesis*

Perlecan is a large basement membrane proteoglycan and a key component of the vascular ECM. It consists of a high molecular weight core protein (~470Da) that it is heavily decorated with multiple glycans including the attachment of three HS chains at the N-terminus. The first indications of perlecan being involved in cancer and angiogenesis came from reports describing a dramatic expression and deposition of perlecan in highly invasive melanomas (Cohen et al., 1994). Later on, it was found out that perlecan acts as a co-receptor for the fibroblast growth factor (FGF2), a well-established pro-angiogenic factor (Aviezer et al., 1994). More importantly, suppression of perlecan in melanoma cells from both human and murine origins blocks the autocrine and paracrine functions of FGF2 and suppresses proliferation and invasion (Aviezer et al., 1997). In similarity with other HS proteoglycans, perlecan interacts with many different growth factors through its GAG chains. Interestingly, transgenic animals lacking the perlecan domain where the GAG chains are attached show impaired tumor growth and FGF2 mediated angiogenesis (Zhou et al., 2004). Furthermore, the 6-O-sulfation of HS chains has been implicated in the regulation of the angiogenic response of endothelial cells to FGF2 and to the vascular endothelial growth factor (VEGF) (Ferrerias et al., 2012). Endothelial-targeted deletions of NDST1, a key enzyme in the N-deacetylation and N-sulfation of HS, impair angiogenesis and reduce tumor growth in murine experimental systems (Fuster et al., 2007). Collectively, the data points to a major pro-angiogenic role for perlecan in tumor neovascularization and highlights the importance of the HS fine structure for these processes.

As is certainly the case for other proteoglycans, perlecan is also the target for proteolytic events that liberate biologically active fragments from its core-protein. Endorepellin is a C-terminal fragment of perlecan that it is generated by the action of an endopeptidase called cathepsin L (Cailhier et al., 2008; Mongiat et al., 2003). In turn, the bone morphogenetic protein 1 (BMP1) can act on endorepellin leading to the proteolytic release of a soluble LG3 domain (Gonzalez et al., 2005). Notably, both endorepellin and its LG3 domain can trigger a signaling cascade in endothelial cells that prevents them from establishing capillary

morphogenesis in several angiogenic assays (Bix et al., 2004). Recent studies have conclusively established an anti-angiogenic role for these two bioactive fragments through the modulation of specific integrin signaling pathways (Douglass et al., 2015). These interesting findings underscore the importance of proteoglycans and their GAG chains for modulation of key processes, such as tumor angiogenesis, through their binding to growth factors. They also exemplify how proteolytic events can generate protein fragments with a distinct biological activity, sometimes completely different from the precursor core protein. However, if GAG glycosylation itself could be a modulating factor for the proteolytic processing, as it has been shown for other types of glycan modifications (Goth et al., 2015), has not been fully addressed.

### *VAR2CSA and pregnancy-associated malaria infection*

Many GAGs can be targeted by microbial pathogens (and virulence factors) for adhering and invading host cells. Several studies have shown that GAGs can also modulate systemic microbial dissemination and immune evasion (Bartlett & Park, 2010). One example of such a GAG-pathogen interaction is exemplified by the acquired ability of malaria-infected erythrocytes (IEs) to evade immune clearance by engaging GAG receptors in the placenta during pregnancy-associated *Plasmodium falciparum* infection (Fried & Duffy, 1996). The adhesion and sequestration of IEs in the placenta is mediated by the expression of the malaria VAR2CSA protein in the membrane of IEs (Reeder et al., 1999). The placental receptor for the VAR2CSA protein is a distinct subset of chondroitin-4-sulfate (CSA) specifically attached to the core-protein of syndecan-1 in the intervillous space and the syncytiotrophoblasts (Ayres Pereira et al., 2016). CSA is expressed in many other locations of the microvasculature but Var2CSA-binding is only supported in the placenta. This suggests unique structural features for placental CSA but the structure of the exact binding sequence has not been fully determined. Pregnancy-associated malaria is highly correlated with maternal anemia, spontaneous abortion, and stillbirth (Brabin et al., 2004).

Surprisingly, it has been recently shown that placental-type CSA is also expressed in a broad array of human malignant cells but not in healthy tissues (Salanti et al., 2015). This CS-signature has been re-named to oncofetal-CS and can be specifically targeted by recombinant VAR2CSA constructs. Contrary to the placental setting, oncofetal-CS expression in cancer appears to be heterogeneous as to the core-protein identities but further studies are needed to address this issue. Still, the presence of a common GAG-signature in cancer cells opens new avenues for diagnosis and therapeutic interventions. As it will be discussed later, some of the results in this thesis indicate that oncofetal CS expression is



also present in pluripotent stem cells and might be a signature for highly proliferative and undifferentiated cellular states.

### *Proteoglycans in development and stem cell biology*

The importance of GAGs for development was initially recognized since mutations in enzymes involved in *Drosophila* HS biosynthesis were linked to impaired signaling through the WNT, FGF and Hedgehog pathways (Bornemann et al., 2004; Haerry et al., 1997). This led to severe defects in cell differentiation and morphogenesis (Hacker et al., 2005). Later, several lines of inquiry have established HS chains as essential regulators of morphogen gradient formation (Hayashi et al., 2012; Kleinschmit et al., 2010; Yan & Lin, 2009). In vertebrates, HS polysaccharides play an important role in modulating early developmental processes such as left-right patterning and neuronal and cardiovascular development (Arrington et al., 2013; Pan et al., 2014). Lineage-specific expression of the HS biosynthetic machinery has also been demonstrated (Nairn et al., 2007; Yabe et al., 2005). These findings, together with the tight developmental regulation of the HS fine structures, suggest the existence of an instructive HS “sugar code” for vertebrate development (Poulain & Yost, 2015). However, this concept remains controversial and needs further confirmation.

The impact of CS on similar developmental processes is currently less well understood. However, recent reports have provided evidence for a critical role under early embryogenesis. For example, CS proteoglycans are required for embryonic cell division and cytokinesis in both *C. elegans* and mice (Hwang et al., 2003; Izumikawa et al., 2010; Mizuguchi et al., 2003). They are also indispensable for pluripotency and self-renewal of embryonic stem cells (ESCs) (Izumikawa et al., 2014). Interestingly, HS becomes critically relevant during exit of self-renewal by facilitating FGF signaling (Kraushaar et al., 2013). These results emphasize the notion that HS and CS are not functionally equivalent, and that the spatiotemporal regulation of different GAG types is an important requirement for normal early development. The mechanistic basis for these processes is still poorly defined.

### 1.2.7 Summary

Initially ascribed with an exclusive structural role, the GAGs (and the proteoglycans) have slowly found their way into the mainstream of biological research. Their impact on fundamental biological processes has brought together researchers from different fields, ranging from biomaterial engineers to oncologists and

developmental biologists. Their role in widespread diseases has become obvious and led to the exploration of novel diagnostic and therapeutic strategies. On the other hand, the current lack of high throughput methodological tools has hampered a faster development for the field and many difficult challenges remain.

# 2. Methodological considerations

## 2.1 Analytical challenge

From a scientific-philosophical point of view, I will argue that modern biochemistry is embedded in a theoretical framework, which can be described as “a structure and function paradigm”. In other words, our current understanding of the biological world is based on the premise that *structural determinants* mediate *chemical interactions*, which in turn, are the main drivers of *biological functions*. Translated to the setting of glyco-analysis, most methodological approaches in glycobiology are generally concerned with one of these three aspects of glycosylation: i) the characterization of glycan structures and the mechanisms regulating their assembly, ii) the detailed understanding of glycan-protein interactions and iii) the assessment of glycan-mediated biological functions as they correlate to specific glycan structures.

This thesis is primarily concerned with the partial structural characterization of proteoglycans. Currently, no single technique can uncover all the structural details of a given glycoconjugate and the structural glycobiologist must rely on a broad array of methodologies to acquire different layers of structural understanding. The choice of methodology is also dictated by the amounts and purity of the available material. Ideally, the complete structural characterization of a proteoglycan will include at least five different variables:

- The identification of the core-protein and the GAG attachment site(s)
- The establishment of the glycan composition (i.e. number and type of monosaccharides and substituents)
- The establishment of the glycan sequence (i.e. the order of constituent monosaccharides and their substituents as well as their linkage configurations and positions)
- A measurement of the macroheterogeneity (glycan occupancy)
- A measurement of the microheterogeneity (diversity of glycoforms)

## 2.2 Purification of GAGs and proteoglycans

Before structural analysis is conducted, the proteoglycans need to be isolated from different biological sources, typically from cell cultures, tissues, organs or body fluids. As discussed before, many proteoglycans constitute scaffolds for organizing ECMs and engage in binding to multiple proteins. For that reason, some of them are resistant to extraction in solvents that preserve molecular associations. Tough dissociative conditions are typically required to break down such macromolecular interactions (Fedarko, 1994; Whitelock & Iozzo, 2002). Given that some proteoglycans are also membrane-bound, the use of detergents is usually recommended to achieve optimal extraction. In general, the protocols for proteoglycan purification take advantage of the physicochemical properties that distinguish proteoglycans from other biomolecules. These include a high density of negative charges, large hydrodynamic sizes, and high buoyant density.

Anion exchange chromatography is a suitable step for general proteoglycan enrichment. Typically, a positively charged matrix is used to interact with negatively charged groups such as the uronic acids and the sulfate substituents of the GAG chains. Bound molecules are then displaced by increasing salt concentration or changing the pH. Although this method is quite unspecific, (all types of GAGs are enriched), it allows for both purification and concentration in a single step. In fact, this technique alone can provide crude proteoglycan preparations of sufficient purity for further analysis. On the other hand, some acidic proteins or nucleic acid may also adhere to the column, requiring endonuclease treatment and/or further purification steps. The presence of salts in the elution buffers may also interfere with downstream analysis, especially if mass-spectrometric detection is desired.

Anion exchange chromatography can be combined with other purification methods to improve the isolation of proteoglycans. Most proteoglycans have large hydrodynamic sizes due to the contribution of the GAG chains. Size exclusion chromatography (SEC) has traditionally been used for fractionation of intact molecules as well as for separating oligosaccharide products of GAG depolymerization reactions (Zaia, 2009). Another advantage is that SEC can provide an estimation of the average weight of the fractionated components. Additionally, most proteoglycans possess high buoyant densities in CsCl equilibrium density gradients (Meselson et al., 1957). This facilitates their sedimentation by ultracentrifugation but this approach only works for proteoglycans with high GAG- protein ratios. In general, there is no single purification protocol that works for all proteoglycans and the conditions must be empirically determined for each experimental system. Other variables such as the biological source, the starting amounts of material and the downstream analysis must also be taken into consideration.

## 2.3 Proteoglycan structural analysis

Proteoglycans are complex molecules and their detailed characterization demands the analysis of several structural levels. One way of simplifying the task is separating the GAG chains from the core protein either by proteolytic digestion or through chemical release. Exhaustive treatment with mixtures of proteases such as pronase has been used to digest the proteoglycan core-proteins into glycopeptides consisting of an intact GAG chain attached to the linking serine residue. These glycopeptides can be further isolated and concentrated by anion exchange chromatography or similar purification approaches. An alternative to this method is the chemical release of the GAG chains by base-catalyzed  $\beta$ -elimination under reducing conditions. This well-established procedure generates a xylitol residue at the GAG reducing end (Seno & Sekizuka, 1978).

Irrespective of which approach is applied, the released GAGs (or glycopeptides) are usually digested by GAG-specific lyases to depolymerize the glycan chain into a pool of constituent disaccharides. This is a powerful way of reducing the heterogeneous GAG populations into sequence representatives of the average structure. An increasing number of studies have shown that the analysis of digested GAG chains can be used to derive structure-and-function relationships in relevant biological contexts (Ly et al., 2010). Partial depolymerization can also be achieved to generate oligosaccharides that are suitable for glycan sequencing. The analysis of intact GAGs and depolymerization products can be conducted using different glyco-analytical techniques (Bielik & Zaia, 2010). However, biological mass spectrometry (MS) has largely emerged as the preferred platform for glycan structural analysis.

## 2.4 Mass spectrometry

Mass spectrometry (MS) is one of the most important micro-analytical techniques in modern glyco-analysis. In principle, MS is based on the generation of gas-phase ions of a compound (the analyte) followed by the accurate measurement of its mass-to-charge ( $m/z$ ) ratio. In this way, the elemental composition of the analyte can be determined. Most modern MS instruments support dissociative techniques to induce fragmentation of selected ions to generate a secondary mass spectrum ( $MS^2$ ). The product ion pattern in  $MS^2$  often contains new information about the composition and structure of the molecule of interest. Multi-stage fragmentation ( $MS^n$ ) is currently possible, increasing the degree of structural insight that can be derived from such experiments. One of the reasons for the total dominance of MS in biochemical analysis has to do with the fact that no other technique is capable of producing more structural information per unit

quantity of an analyte. The advent of “soft” ionization methods such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) has made possible the direct analysis of femtomole quantities of non-volatile and thermally labile compounds such as proteins and glycans, even in relatively complex biological samples. As ESI was used throughout this thesis, a short description of its basic principles is pertinent.

### *Electrospray ionization*

ESI is achieved in the ion source by spraying a solution of the analyte through a small capillary into an electric field. An electric potential is typically established between the tip of the capillary and a counter electrode. The electrospraying leads to the formation of an aerosol of highly charged droplets at atmospheric pressure and the polarity of the electric potential will determine if positively or negatively charged ions are formed. These initial droplets are made to shrink by a drying gas through solvent evaporation. As droplets evaporate and diminish in size, they become unstable due to an increase in the electric charge density on their surfaces. Repulsive forces between equal charges make the droplets “to explode” generating even smaller droplets. Although several mechanisms have been proposed for the last step of ESI (Wilm & Mann, 1994), it always results in the complete evaporation of the solvent yielding a continuous beam of gas-phase ions. Modern nano-ESI generates ion droplets 10-fold smaller than traditional ESI eliminating certain stages of droplet division and facilitating greater ionization efficiency and stronger signals (Karas et al., 2000).

Compared to older ionization techniques, ESI imparts relatively low energy to the molecules during desolvation. For that reason, it is considered a soft ionization method that prevents “in source” metastable decay. As the ions are generated directly from solution, ESI constitutes a perfect interface for liquid chromatography (LC) coupled to MS detection. ESI tends also to produce ions carrying multiple charges and the  $m/z$  ratios of most biomolecules fall within the detection range of commonly used MS analyzers. On the other hand, the presence of contaminants such as salts and plastic polymers can severely affect the ionization process, which makes sample preparation for ESI more demanding than for other alternative techniques such as MALDI.

## *MS analysis*

After ESI, the ions are led into a mass analyzer device to allow for accurate  $m/z$  measurements, and in certain applications, for controlled fragmentation of selected precursor ions. There is a large variety of MS instruments (including hybrid types) that are routinely used for biomolecular analysis. The reliable identification of biomolecules in complex mixtures requires robust and sensitive MS instrumentation with high resolving power, mass accuracy, and wide dynamic range. Recently, the Orbitrap analyzer has proven to be a popular high-resolution instrument due to its relatively low cost, its flexibility in hybrid configurations and its exceptional high-performance, mass accuracy and resolution (Hu et al., 2005).

One of the central features of the Orbitrap is its specific geometry. In the Orbitrap, a DC voltage is applied between a spindle-like inner electrode and a barrel-like outer electrode. This static electric field creates stable rotatory ion trajectories around the central electrode as well as oscillations in the axial  $z$ -direction. The axial frequency is used to derive the  $m/z$  ratio of an ion package. The axial oscillations are independent of both the initial properties of the ions and the orbiting motions around the central electrode. This independence is the main parameter responsible for the high accuracy and mass resolving power of the Orbitrap (Makarov, 2000). The axial oscillations generate an image current in the outer electrode that is acquired as a time-domain transient and converted into frequencies (and  $m/z$ ) by Fourier transformation. Commercially available Orbitraps can achieve mass accuracies of 2-5 ppm, a linear dynamic range of up to four orders of magnitude, and a mass resolution of up to 150,000.

Orbitraps can also be combined with other devices such as a quadrupole mass analyzer to create a composite instrument that allow for a more flexible ion manipulation. In typical hybrid configurations, such as the Q Exactive instrument used during this thesis work, the quadrupole can be used as a mass filter device (Michalski et al., 2011; Scheltema et al., 2014). As the name already indicates, the quadrupole consists of four parallel metal rods where RF and DC voltages are established. Ions travelling through the space between the rods can be manipulated by changing the voltages so that only molecules of a specific  $m/z$  ratio can move in stable trajectories while all the other ions will collide with the rods. The selected ions can then be transferred to the storage or fragmentation devices before Orbitrap analysis.

### *Collisional dissociation*

Although accurate mass measurements alone can sometime lead to the identification of a compound, it is more often the case that fragmentation of the analyte is additionally required to generate structural information. Dissociation of molecular ions in the gas-phase can be achieved by different fragmentation techniques (Sleno & Volmer, 2004). Independently of the approach, the main goal of such experiments is to achieve a controlled decomposition of a selected molecule, following predictable fragmentation rules and yielding product ions that can aid in establishing the identity and the structure of the targeted compound. The fragment ions are then collected and their  $m/z$  ratio is determined. Some instruments support multistage fragmentation where selected fragment ions can undergo a new round of fragmentation ( $MS^n$ ).

Collisional-induced dissociation (CID) is a common technique to achieve molecular ion dissociation in the gas-phase. The ions are accelerated and allowed to collide repeatedly with neutral molecules, usually an inert gas, inside a collision chamber. The kinetic and collisional energy is converted into internal energy and dissipated across the molecule leading to bond dissociation and the generation of fragment ions and neutral losses. In the case of glycans, the fragmentation of glycosidic bonds is often favored but exactly which product ions are formed is also dependent on the collisional energy. Higher-energy dissociation (HCD) is a beam-type CID specific to the Orbitrap. Fragmentation takes place in a HCD cell outside the Orbitrap and the ions pass through a curved shaped ion trap (C-trap). The C-trap is critical for Orbitrap measurements as it squeezes electrostatically the ion package in time and space followed by pulse-injection into the Orbitrap. Otherwise, the static electric field of the Orbitrap would not result in efficient trapping of the ions and they would just pass through the device. In some configurations, the fragmentation can actually take place inside the C-trap itself (Olsen et al., 2007).

## 2.5 MS-based GAG analysis

GAGs are long and intrinsically polydisperse molecules. After release of the intact glycan from the core protein, GAGs are often subjected to complete or partial enzymatic depolymerization. Commercially available CS and HS degrading enzymes act through an eliminative mechanism (lyases) to cleave the HexNAc-HexA bonds. This cleavage generates 4,5-unsaturated HexA ( $\Delta$ HexA) at the non-reducing end of the disaccharides and of the hexasaccharide containing the linkage region. The specificity of individual enzymes can be exploited to



gain access into the polysaccharide structures (Ernst et al., 1995). Chemical treatments for achieving cleavage of N-sulfated domains of HS/Heparin are also commonly used. The released disaccharide pool can be further fractionated and subjected to quantitative analytical methods including several derivatization schemes combined with chromatographic separation and fluorescence detection (Zaia, 2009). However, MS approaches have become increasingly popular, especially in combination with a wide range of liquid chromatography (LC) strategies.

The MS analysis of GAG disaccharides is often conducted in negative mode ionization but positive mode with ion-pairing reagents has also been reported (Juhasz & Biemann, 1994, 1995). In general, the sulfate groups tend to dissociate as the internal energy of the molecule increases, which demands low desolvation energies to prevent metastable decay during ionization. Additionally, during fragmentation experiments in negative mode, the dissociation of the sulfate groups is largely dependent on the charge state of the precursor ion and the amount of collisional energy (Staples & Zaia, 2011; Zaia & Costello, 2001). Currently, ESI is considered the gentlest ionization method for disaccharide analysis but fine-tuning of the MS operating settings remains crucial for success. Given that ion efficiency differs between the expected disaccharides, the use of defined standards is also paramount to enable quantification in unknown samples. Differentiation of isobaric structures, for example  $\Delta$ HexA-GalNAc(6S) vs  $\Delta$ HexA-GalNAc(4S) can be achieved by MS/MS fragmentation and analysis of the product ion pattern (Desaire & Leary, 2000). The analysis of larger GAG oligosaccharides is also possible but the complexity of the resulting MS spectra is far more challenging. On the other hand, while disaccharide determination is a good way of deriving the average composition of a GAG chain, direct oligosaccharide sequencing allows for domain characterization and provides an overview of biologically relevant sequences. Notably, the sequencing of the entire GAG chain of bikunin and decorin has recently been reported (Ly et al., 2011; Zhao et al., 2013). The remarkable results from those studies point to the existence of surprisingly well-defined GAG sequences in native proteoglycans, challenging our current understanding of their biosynthesis and their structure-and-function relationships.

## 2.6 Glycoproteomics

Although detailed structural information can be obtained by releasing or separating the GAG chains from the proteins, such approaches cannot provide the identity of the carrier protein and the site(s) of glycan attachment. This information is

critical to assign specific functions to particular proteoglycans and to evaluate the impact of site-specific GAG modification on biologically relevant contexts. In addition, glycans appear as arrays of related structures attached to the same position on the aglycone (microheterogeneity) and their occupancy (macroheterogeneity) may vary, which cannot be effectively addressed without maintaining the connection between the glycan and the protein. In principle, these methodological concerns are not only important for proteoglycans but also for the analysis of all N- and O-glycoproteins. The protein information has traditionally been obtained through standard biochemical techniques such as western blots and/or site-specific mutagenesis or inferred from mRNA expression data. The problem here is that such methods cannot be applied in a system-wide fashion and require a-priori knowledge of the relevant glycoproteins before analysis.

During the last years, new MS-based approaches have emerged that specifically address these difficult issues. Glycoproteomics is a new field that focuses on the system-wide site-specific study of the glycoproteome (Nilsson et al., 2013; Thaysen-Andersen & Packer, 2014). The main molecular targets for glycoproteomics are usually glycopeptides generated by protease-digestion of glycoproteins. Many different strategies and workflows have been reported during the last two decades including the engineering of the glycosylation by genome editing or through metabolic labeling, the development of new enrichment protocols and LC-separation schemes, and the use of new fragmentation techniques (Hinneburg et al., 2016; Nilsson et al., 2009; Steentoft et al., 2011; Vester-Christensen et al., 2013; H. Zhang et al., 2003). In similarity to MS-based glycan analysis, glycoproteomics approaches are not general in nature and need customization that takes carefully into account the biological source, the glycan structures and the desired level of structural information. Currently, glycoproteomics data can account for the composition of the glycan chain, the glycosylation site(s) and the identity of the carrier core protein. Unfortunately, other “high-resolution” glycan information such as the glycosidic linkage positions and  $\alpha/\beta$ -configurations has not been demonstrated at the glycopeptide level. In the same line, it has been widely assumed that the identity of isobaric monosaccharides cannot be determined by these approaches although new data, including the findings from this thesis, have started to challenge that notion (Halim et al., 2014; Hofmann et al., 2015; Sarbu et al., 2015; J. Yu et al., 2016).

## 2.7 Glycoproteomics tools for proteoglycan analysis

During the course of this thesis, we started developing MS-based glycoproteomics approaches for the global analysis of CS/DS- and HS- proteoglycans from biological fluids and cell cultures. To our knowledge, only one single study has been reported so far by another group where the potential of high-throughput biological MS has been directed to answer questions regarding the identity and attachment sites of CS-proteoglycans (Olson et al., 2006). In that study, nine novel CS-proteoglycan core-proteins were simultaneously identified in *C. elegans* by releasing the GAG chains through beta elimination and tagging the glycosylation sites with DTT, followed by LC-MS/MS analysis. Compared to traditional approaches, where a combination of several complicated biochemical techniques were required to identify one single proteoglycan, the possibility of getting high-throughput data in a single experiment is quite an achievement. Obviously, these experimental setups were applied at the expense of losing information on the fine-details of the GAG chains, and can be considered complementary to other glyco-analytical tools. However, the number of currently known proteoglycans is very small compared to other glycosylation events. This fact could probably reflect the limited protein preferences for GAG modification but also technical difficulties in obtaining the protein identities of novel proteoglycans and/or a large bias towards disregarding the core proteins as mere presenters of GAG chains.

In the present work, we took advantage of the fact that commercially available GAG-degrading enzymes, in addition to disaccharides, leave a hexasaccharide structure covering the linkage region still attached to the protein. We set up a simple protocol using proteolytic digestion, anion exchange chromatography and GAG depolymerization, to obtain what it is referred throughout this thesis as proteoglycan linkage region glycopeptides (Fig.5). These glycopeptides turned out to be suitable targets for different LC-MS/MS analytically strategies yielding information on completely novel core-proteins and allowing for the site-specific analysis of hybrid proteoglycans. This analysis also gave insights into the combinatorial possibilities of the proteoglycan linkage region, including the identification of novel glycan modifications. We also developed *SweetNET*, a robust bioinformatics workflow to handle hundreds of thousands of MS/MS spectra generated by these glycoproteomics experiments and showed its general applicability not only to the linkage-region glycopeptides but also to N- and O-glycopeptides. Finally, we also demonstrate how these new tools can be used to address biologically relevant problem such as the detailed analysis of the expression of oncofetal CS in cancer and iPS cells.

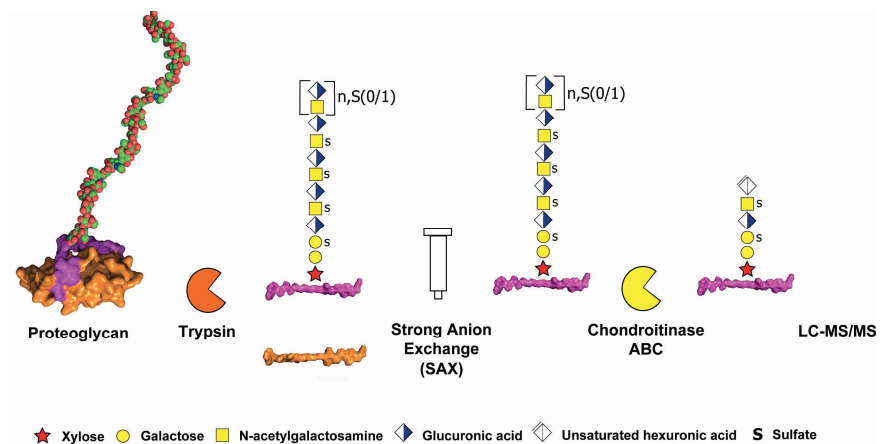


Figure 5. General workflow developed in this thesis to generate and analyze proteoglycan linkage region glycopeptides. During the analysis of heparan sulfate proteoglycans the Chondroitinase ABC step is substituted with heparinase I-III treatment.

## 3. Aims

The general aim of this thesis was to develop a comprehensive glycoproteomics workflow for the global characterization of CS/DS- and HS- proteoglycans from complex biological samples. Our main hypothesis was that traditional approaches for glycosaminoglycan characterization can be combined with global proteomics/glycoproteomics techniques to gain insights into multiple aspects of the proteoglycan structures that are not covered by current methodologies. If successful, we hypothesized that such approaches should facilitate addressing generally unresolved issues regarding glycosaminoglycan attachment sites, glycan structural variations and also eventually expand the number and types of proteoglycans found in nature. For overcoming the technical difficulties, we hypothesized that starting with the simplest and most easily available proteoglycans and successively, in a step-by step manner, expanding into the various classes of proteoglycans would be the most efficient way to work.

The specific aims were to:

- Develop methods for the generation and LC-MS/MS structural characterization of linkage region glycopeptides derived from CS- and HS- proteoglycans
- Develop a robust bioinformatics platform to effectively handle high-throughput glycoproteomics data
- Apply glycoproteomics tools to assess the site-specific expression of CS in iPS and cancer cells

# 4. Results and discussion

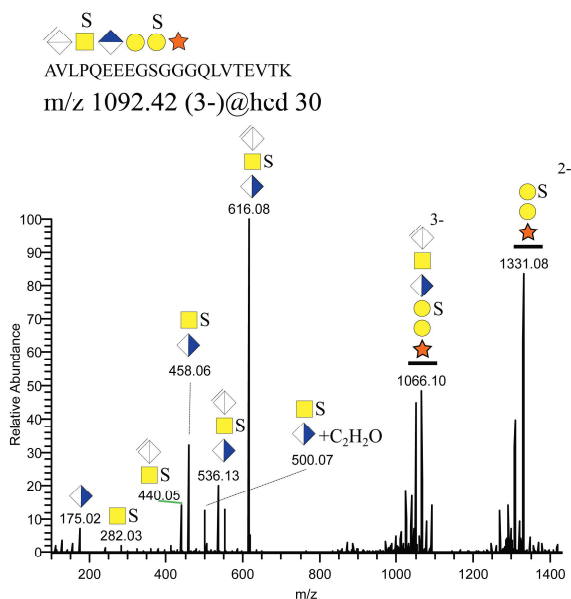
## 4.1 Targeting the data

### *Choosing optimal MS conditions*

Glycopeptides are hybrid molecules composed of a glycan and a peptide moiety and both components will simultaneously contribute to their average physico-chemical properties. As glycans and peptides are chemically different, it is often hard to find optimal conditions for the chromatographic and MS analysis of glycopeptides. Usually a combination of approaches is required. Furthermore, proteoglycan-derived glycopeptides are substituted with labile sulfate groups. A typical problem during collisional fragmentation of sulfated glycopeptides is that structurally informative bonds have different kinetics of dissociation, in the order from weaker to stronger: sulfate ester bonds => glycosidic bonds => peptide bonds. The gas-phase fragmentation of glycosylated peptide ions is also complex. In fact, several dissociation pathways may co-exist and their balance is dependent on the method chosen as well as on the composition and charge of the analyte (Dodds, 2012).

In Paper I and II, we explored MS conditions to obtain structural information from proteoglycan linkage region glycopeptides. Instead of combining orthogonal fragmentation techniques we decided to stick to HCD as it yielded both glycan and peptide fragmentation in a single step. At the same time, the level of the input energy can be easily fine-tuned which turned out to have a major impact on the fragmentation of our target analytes. For example, in Paper II we found that HCD fragmentation at low energies facilitates dissociation of the glycosidic bonds at the expense of the peptide fragmentation. From such experiments, the composition and sequence of the modifying glycan could be established. Increasing the energy resulted in the opposite outcome, yielding good peptide fragmentation but the glycans were completely lost. We developed then a strategy where each precursor ion underwent sequential or parallel fragmentation at different energy levels. As it is shown in Paper II, this resulted in improved structural determination of the linkage-region glycopeptides but the substituents, especially sulfates, were still dissociated as neutral losses. Full deprotonation can compensate for the fragility of sulfate groups e.g. by conducting the fragmentation in negative mode or by using ion-pairing reagents. As discussed before, negative mode is a natural analytic choice for GAG

oligosaccharides as it increases the stability of the sulfate groups. Unfortunately, fragmentation of linkage glycopeptides in negative mode does not produce enough peptide fragments, which hinders a high-throughput strategy for the identification of the core-proteins (Fig 6).



*Figure 6. MS/MS spectrum of a CS-linkage region glycopeptide derived from human urinary bikunin. The analysis was conducted in negative mode ionization and illustrates the absence of informative peptide fragmentation.*

For that reason, most of the experiments in this thesis were conducted in positive mode, despite the obvious loss of the sulfate substituents. However, we also show in Paper II that cation complexation through  $\text{Na}^+/\text{H}^+$  exchange is a good way of achieving stabilization of the sulfates. Such parallel experiments were conducted for pinpointing their position on the glycan chain. Summarizing, the hybrid nature of our targeted glycopeptides required the parallel or sequential use of different techniques. The amount and complexity of the output data also demanded the development of robust bioinformatics tools for data handling.

The bioinformatics treatment of glycoproteomics data poses many challenges. One major difficulty compared to other molecules lies on an exponential increase in the search space as the combinatorial possibilities depend on variations regarding both the peptide sequences and the glycan units. Currently, there is a lack of glycopeptide spectral libraries for structural matching and common proteomics-oriented peptide-spectrum match (PSM) algorithms tend to underestimate the confidence of modified peptides. These shortcomings are usually compensated for by time-consuming “manual verification” being strictly dependent on high competence of single individuals. During the last few years, some steps have been taken to develop glyco-bioinformatics tools to address these problems. In Paper III, we developed *SweetNET*, a bioinformatics platform for handling high-throughput glycopeptide MS/MS data in a unified semi-automatic environment. Compared to current approaches, *SweetNET* contributes with the idea that the organization of the data based on glycopeptide-specific spectral features and molecular networking, facilitates MS-based structural assignment. The workflow was tested on data from different enrichment protocols showing that not only proteoglycan-derived glycopeptides but also typical N- and O- glycopeptides, can be effectively addressed in our platform.

One pivotal aspect that *SweetNET* considers is the presence of glycan-derived oxonium ions generated through the collisional dissociation of glycoconjugates. These oxonium ions constitute fingerprints for the presence of glycopeptides and the *SweetNET* software exploits them for data filtering and grouping. We have previously shown that decomposition of GalNAc or GlcNAc yields different oxonium ion intensity profiles (Halim et al., 2014; J. Yu et al., 2016). These profiles are automatically computed in the workflow to provide the saccharide identity of the N-acetylhexosamines. This information is used to distinguish between various types of glycans and core-structures, taking a further step into obtaining “high resolution” glycan information from glycoproteomics data. More specifically, in Paper III, we show that separation of N-glycans from O-glycans or even different O-glycan core-structures can be established by this method, even before the identity of the whole glycopeptide is obtained.

Another characteristic of glycopeptides is that the glycan moiety often appears as an array of glycoforms. This inherent microheterogeneity is difficult to tackle, as it requires predicting all possible saccharide permutations, which often translates into the need for a certain pre-knowledge of the expected glycan modifications in the sample. *SweetNET* handles this problem by incorporating the concept of molecular networking (Bandeira, 2007; Guthals et al., 2012; Nguyen et al., 2013). In such approaches, spectral similarities between structur-



ally related compounds are exploited to cluster the MS/MS data. Consensus identification is then derived from sets of similar spectra facilitating the discovery of unexpected modifications. Integration of oxonium ion intensity profiles and spectral clustering are thus the main novel features of *SweetNET*. The power of this bioinformatics workflow is exemplified in Paper III where the analysis of hundreds of thousands of glycopeptide MS/MS spectra was completed during a few hours resulting in the structural characterization of 165 N-, 215 O- and 23 CS-glycosylation sites from around 200 human core proteins. Additionally, it also led to the identification of novel fucose modifications in the linkage region of several human urinary CSPGs. However, one important weakness of *SweetNET* in its current format is that it lacks a robust method for estimating the false discovery rate. This will therefore be addressed in future developments of the software.

## 4.2 Targeting the protein

### *Identification of novel proteoglycan core-proteins*

One of the major findings of this thesis is the identification of novel core-proteins carrying GAG modifications. Several attempts to develop a classification system for proteoglycans have been reported before (Iozzo & Murdoch, 1996; Iozzo & Schaefer, 2015). In contrast to the GAG chains, the core-proteins are far more divergent and encompass gene products with different structures, cellular locations and functions. A careful review of the literature also reveals that the evidence for GAG attachment on specific proteins ranges from solid to circumstantial, depending on the choice of methods and biological systems. Hitherto, the most recent nomenclature for vertebrate proteoglycans encompasses 43 core-proteins. This classification is based on three different criteria: cellular localization, gene/protein homology and domain architecture. Based on the literature and the data presented in this thesis, the number of vertebrate proteoglycans can be expanded to 77 genes (Fig 7). The development of more sensitive technologies and the systematic exploration of new biological sources will most likely increase this number in the future. In this context, the confirmation of 23 previously established proteoglycans and the identification of 21 novel ones, show that glycoproteomics is a powerful tool for such an endeavour.

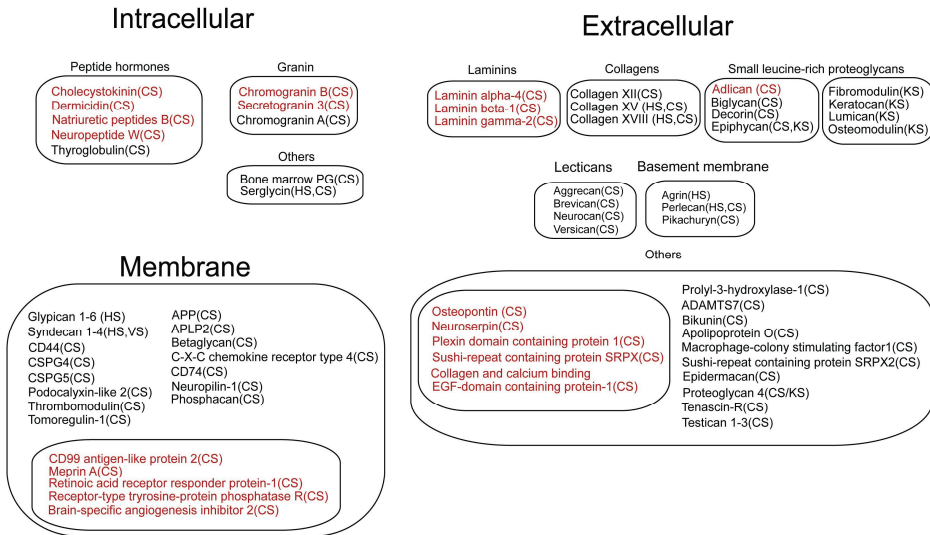


Figure 7. Summary of all known vertebrate proteoglycan core proteins. All novel proteoglycans identified during this thesis are marked in red.

During several years, serglycin (SRGN) was largely believed to be the only true intracellular proteoglycan. Several groups have also reported that Chromogranin A (CHGA), a member of the granin family located in the secretory vesicles of neurons and endocrine cells, is also a potential carrier of CS modifications (Barbosa et al., 1991; Gowda et al., 1990). As it is shown in Paper I, we could confirm the proteoglycan status of CHGA and provided, for the first time, the exact site for GAG attachment. Two other members of the granin family, Chromogranin B and Secretogranin 3 were also identified carrying GAG modifications. Apart from being precursors to various bioactive peptides, the granins are key players in the sorting and aggregation of secretory products in the trans-Golgi network, thereby having a significant impact on vesicle biogenesis (Bartolomucci et al., 2011). Similarly, SRGN promotes electron-dense core formation of the granules of mucosa mast cells (Braga et al., 2007). It has been advocated that the acidic GAG chains could play an active role in these processes. In fact, several lines of evidence implicate GAGs in proper granule formation as well as in the storage of basic components in both endocrine and exocrine cells (Abrink et al., 2004; Aroso et al., 2015; Ringvall et al., 2008).

As mentioned before, CHGA is the precursor of several bioactive hormone peptides such as vasostatin -1 and -2, catestatin and parastatin, with a broad spectrum of biological activities. Human thyroglobulin, the precursor of the thyroid hormones is also a CS proteoglycan (Conte et al., 2006). In Paper I and V, we found that other human pro-hormones (and hormone peptides) are carriers of CS chains, for example cholecystokinin (CCK), a hormonal regulator of the gastrointestinal system, and the brain natriuretic peptides (BNP), a cardiac hormone and a clinical marker for heart failure. Intriguingly, in both cases the GAG sites were located at the N-terminus of the proteins in predicted pro-peptide regions. It is tempting to speculate that the GAG chains may be important during storage and/or processing but are absent from the functional protein after maturation (i.e. after removal of the pro-peptide regions). That would explain why such well-known proteins have never been observed carrying GAG modifications. Additionally, some of novel GAG sites are not conserved across mammalian species (Fig 8), which highlights the importance of carefully choosing relevant experimental systems.

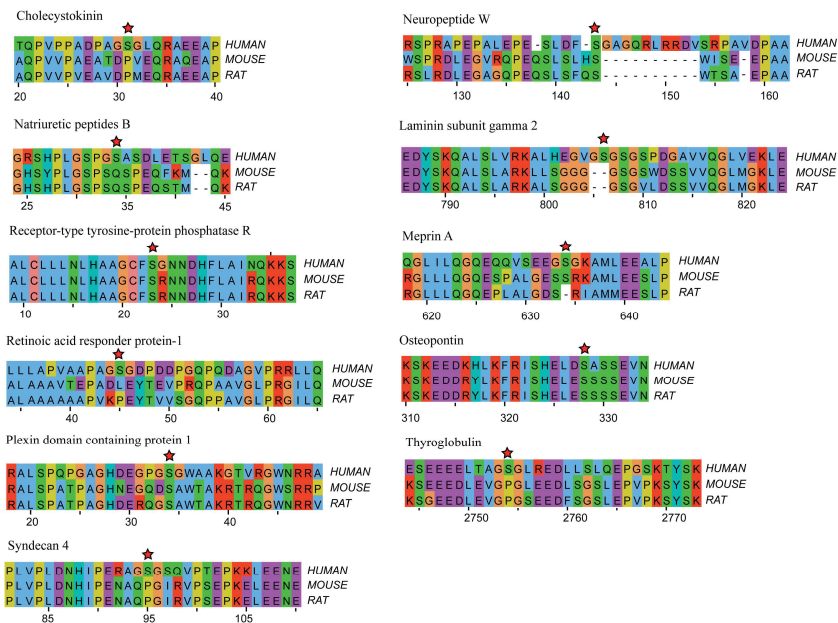


Figure 8. Some of the novel GAG sites identified in this thesis are not well conserved between humans and other mammals (illustrated for mice and rats).

As our methodology relies on proteolytically digested glycopeptides, the location of the GAG sites in relation to the intact proteins cannot be assessed. This

problem becomes critical when dealing with proteins that are targeted for endogenous proteolytic events or that undergo alternative splicing. Nevertheless, the identification of GAG modifications on several proteins associated with secretory vesicles indicates that glycosylation may play an important role in this context. It remains to be seen if other pro-hormones can be equally modified.

As it is shown in Fig 7 some other novel proteins were also found modified with GAGs. One of the general functions of the proteoglycans is to organize the basement membrane /ECM by interacting with collagens and laminins. Some collagens are known to be “de facto” proteoglycans but we also found that some laminins carry GAG chains as well. This finding underscores the important role of acidic GAG chains for tissue organization. Finally, the small-leucine rich proteoglycans (SLRPs) count among the most abundant proteoglycans in the extracellular space of connective tissue. In Paper I, we found that the Matrix-Remodelling Associated protein 5 (MXRA5) also known as adlcan, carries a GAG chain and should therefore be included among the known CS-SLRPs together with decorin, biglycan and epiphygan.

#### *Revisiting the sequence context of the GAG attachment sites*

Despite our findings, the proteoglycans appear to constitute a relatively small number of proteins, probably less than 100 genes. This limited number indicates that the requirements for GAG attachment are restricted and/or regulated by specific structural determinants on the acceptor core-proteins. One aspect already addressed in several studies is the presence of semi-consensus sequences that facilitate the addition of the linking xylose to the polypeptide chain. The context of the primary sequence appears to be critical, as GAGs are reportedly attached to serine residues flanked by glycine. The presence of acidic amino acids near the modification site also seems to be important. To date, most of these conclusions have been drawn on small datasets because the number of known attachment sites has been difficult to obtain. One advantage of our glycoproteomics approach is that the exact position of the attachment site may now be readily assessed.

We analysed the sequence of 48 experimentally determined glycosylation sites derived from 44 core-proteins (Fig 9). As is clearly observed from the WebLogo analysis, the emerging pattern largely confirmed what has been proposed before, although some differences were also obvious. In addition to glycine, the presence of alanine and serine residues in position +1 and -1 shows that they too are feasible possibilities for several endogenously glycosylated GAG sites. In line with that, changing the flanking glycine to the chemically equivalent alanine or to serine does not disturb xylose transfer to the acceptor site of

bikunin (Roch et al., 2010). Based on our data, a tentative [EPA]-[GSA]-S-[GA] motif near acidic amino acids could be formulated although not all the sites seem to comply with it.

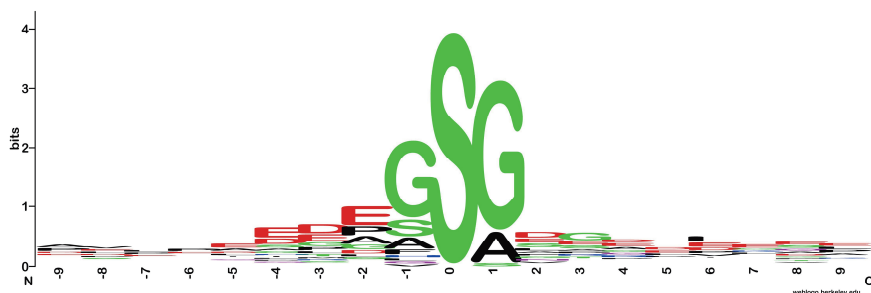


Figure 9. WebLogo analysis of 48 experimentally determined GAG sites derived from 44 human core proteins.

Interestingly, it has been suggested that the presence of SG repeats promotes the specific assembly of HS chains over CS/DS chains. In Paper IV, we set up a method to analyze linkage-region glycopeptides derived from mouse Perlecan, a typical basement membrane proteoglycan. Perlecan is known to carry both HS and CS chains but how these different GAG-types relate to the glycosylation sites is currently unknown. It should be mentioned that all potential GAG sites on Perlecan are well conserved across mammals. By digesting the glycans with HS- or CS-specific lyases in combination with oxonium ion analysis, we determined the HS sites in Perlecan to be located on the N-terminal domain of the protein. The amino acid sequence **ASGDGLGSGDVGSGD** for the HS modifications agreed with the postulated requirement of SG repeats. Interestingly, we also found a novel hybrid C-terminal site **DWHPEGSGGN** that can be modified with either HS or CS. The human homologous peptide was also identified in cerebrospinal fluid and urine (Paper I). Although the absence of SG-repeats may suggest that this site is preferentially targeted for CS assembly, the presence of bulky aromatic residues near a GAG site (a tryptophan in this case) has been associated with HS rather than with CS (L. J. Zhang & Esko, 1994). We have also identified a hybrid site on human Collagen XVIII derived from human induced pluripotent stem cells (Fig 10). This attachment site looks like a classic CS-site but has an aromatic residue (phenylalanine) only a few amino acids away from the modified serine. As the data is not enough to draw any conclusions on this subject, it will be interesting to see if other hybrid proteoglycans follow the same trend.

LTFIDMEGSGFGDLEALR\_Collagen alpha-1(XVIII)

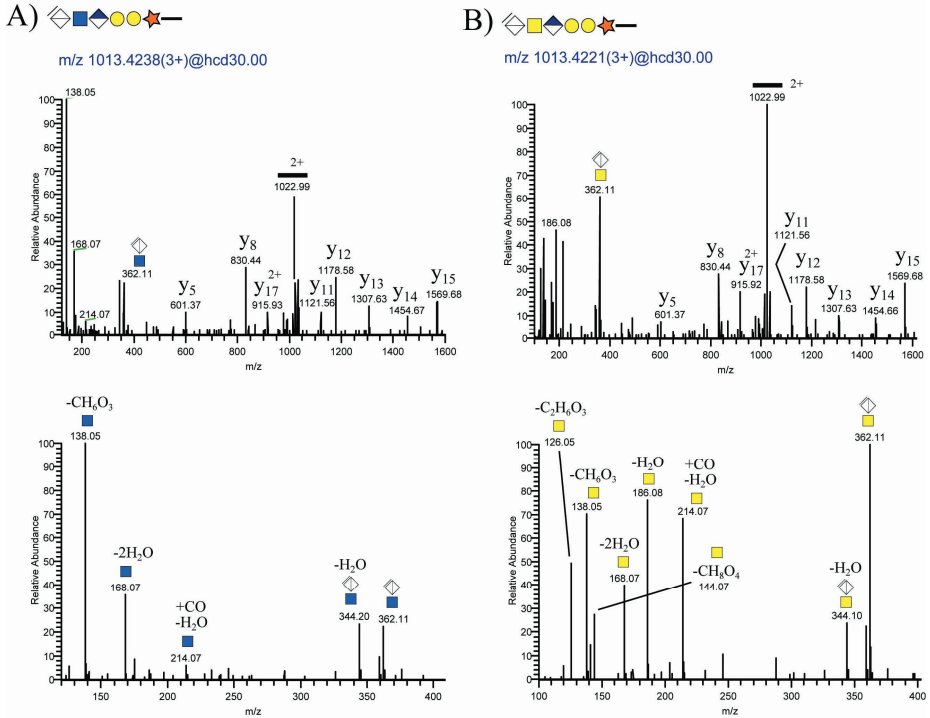


Figure 10. HCD MS/MS fragmentation spectra of HS (A) and CS (B) glycopeptides derived from Collagen alpha 1 (XVIII) from human iPS cells. Linkage region hexasaccharides carrying HS and CS modifications are found at the same attachment site (top panels) but differ in their oxonium ion intensity profiles confirming the simultaneous presence of GlcNAc (HS) and GalNAc (CS) modifications (bottom panels).

GAGs as potential allosteric factors for proteolytic processing of the proteoglycan core-proteins

Interestingly, the novel hybrid site on perlecan was found to be located only four amino acids away from the endogenous BMP1-cleavage site that liberates the LG3 domain of Endorepellin (Gonzalez et al., 2005). As discussed before, the LG3-domain is a proteolytic fragment of perlecan with anti-angiogenic properties. The close proximity between the cleavage- and the novel GAG- sites suggests that BMP1 may physically interact with the glycan (Fig 11) which in turn implies that the GAG chain could serve as an allosteric element for the enzymatic cleavage. This possibility becomes even more fascinating in the light of recent

findings showing that BMP1 is responsible for the proteolytic processing of pro-decorin and pro-biglycan leading to the complete removal of their pro-peptide regions (Scott et al., 2000; von Marschall & Fisher, 2010). In both cases, the BMP1 cleavage sites are also located only 4-5 amino acids away from the GAG-attachment sites.

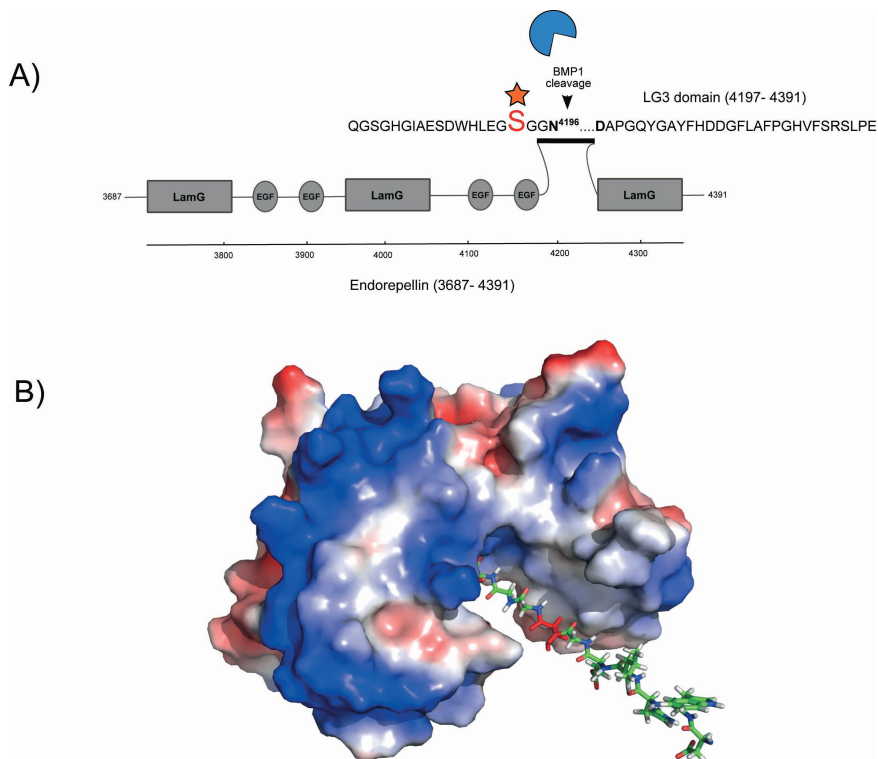


Figure 11. Endorepellin is cleaved by BMP1 only three amino acids away from the C-terminal hybrid site (A) releasing the LG3 domain, one of the laminin-G-domains (LamG) of perlecan. In B, the crystal structure of the proteolytic domain of BMP1 is depicted with a modelled peptide inside the catalytic cleft corresponding to the sequence around the endorepellin cleavage site. The GAG-site serine is coloured in red. The closeness between the proteolytic site and the GAG site suggests that the GAG-chain may physically interact with the protease and modulate the cleavage possibly by binding to the rim of positively charged amino acids of the BMP1 surface (positive=blue, neutral=white, negative=red).

To the best of our knowledge, an allosteric role for the proteoglycan GAG chains during proteolytic processing of their own core-proteins has never been suggested before. However, the fact that some proteoglycans are processed in complex ways in both health and disease, and that some proteolytic events take place near glycosylation sites should encourage an exploration of this possibility through experimental means in relevant systems. The methodology described in Paper IV could help in addressing this fundamental question.

### 4.3 Targeting the glycan

#### *Revisiting the structure of the proteoglycan linkage region*

The assembly of the proteoglycan linking tetrasaccharide constitutes an important checkpoint for CS and HS biosynthetic regulation. Several glycan substitutions have been reported so far, some of them having a regulatory function. The analysis of glycopeptides as conducted in Paper I allowed for characterization of the glycan structure of the linkage region in addition to pinpointing the GAG attachment sites. In Paper II, we conducted a detailed characterization of bikunin CS glycopeptides across different human body fluids. Bikunin is the light chain of the inter-alpha-trypsin inhibitor (I $\alpha$ I) protein family, which is highly abundant in body fluids. The I $\alpha$ I complex is composed of a unique macromolecular arrangement of several heavy chains (H1, H2 and H3) proteins cross-linked to the CS-GAG chain of bikunin through Asp to GalNAc ester bonds (Salier et al., 1996). The free form of bikunin is known as the urinary trypsin inhibitor (UTI). Bikunin is probably one of the best-studied vertebrate proteoglycans and its GAG chain is apparently simple. It carries a single low-sulfated CS chain attached at Ser-10 with a uniform GlcA stereochemistry. The linkage region has been structurally addressed in previous studies concluding a uniform 4-O- sulfation of the second Gal (Chi et al., 2008; Yamada et al., 1995). The heavy chains are crosslinked towards the non-reducing end of the CS chain but their exact arrangement has not been completely established (Enghild et al., 1989).

With this information in mind, it was very surprising to find in Paper II that the linkage region of bikunin was unprecedentedly heterogeneous. In addition to Gal sulfation, we were also able to identify a subpopulation carrying xylose phosphorylation, specifically in urinary samples. Similarly, several other phosphorylated CSPGs were also detected in urine and CSF as shown in Paper I. In general, xylose phosphorylation in bikunin and in the other native proteoglycans occurred at substoichiometrical levels. Unfortunately, as partial glycan de-



polymerization was conducted before MS analysis it was not possible to assess if the phosphorylated GAGs differed in any respect from their non-sulfated counterparts. However, during enrichment, the glycopeptides carrying intact GAG chains were separated by SAX chromatography and their elution profiles thus correlate with their length and/or sulfation degree. We found that GAGs with unmodified or phosphorylated linkage regions were mostly recovered in the low-salt elution fractions indicating shorter or less sulfated chains. Also in line with previous studies, we identified some degree of xylose phosphorylation in the linkage regions of HS proteoglycans as reported in Paper IV.

Beside the well-studied xylose phosphorylation, we show in Paper II that the innermost Gal of bikunin can also be decorated with a novel NeuAc modification. A NeuAc $\alpha$ 2,3 capping of the Gal has been described before in CSPG linkage regions derived from *FAM20B* knocked cells and in xyloside-primed GAGs from skin fibroblasts (Takagaki et al., 1996; Wen et al., 2014). However, given that the NeuAc modifications on bikunin was observed on structures extended with at least one disaccharide beyond the linking tetrasaccharide, the attachment of the NeuAc moiety should be different from those previously reported (and most likely a 2,6-linkage). Additionally, novel fucose modifications were also detected in urinary bikunin (Paper II) as well as in other native human CSPGs (Paper III and V). Based on the MS data, the fucose moiety was invariably attached to the linking xylose. As discussed in the method section, the glycan linkage position is not possible to obtain from a glycoproteomics experiment but there is an intriguing possibility that the fucosylation may compete with the C-2 phosphorylation of xylose. The biosynthesis of fucosyl-branched CS-structures has been demonstrated in some echinoderms such as sea cucumbers but not in higher organisms (Myron et al., 2014). Taken together, the findings presented in this thesis indicate that the structure of the CS linkage region of bikunin is highly heterogeneous, contrary to previous studies. We also reported novel NeuAc and Fuc modifications of the linkage region of bikunin and of other human CSPGs, which points to the existence of yet unknown glycosyltransferases involved in the biosynthesis of CS chains. The extent and biological importance of these novel modifications remains to be determined.

### *Solving the IaI structural puzzle*

The CS chain of bikunin is also unique in the sense that it is often cross-linked to several heavy chain proteins to give rise to the IaI complexes (Salier et al., 1996). The IaI complex is composed of bikunin cross-linked through its CS chain to either H1 or H2 or both.

The combination of bikunin and H3 is known as the pre- $\alpha$ -trypsin inhibitor. By digesting these molecules with a cocktail of CS-specific lyases (Chond ABC), we obtained both bikunin linkage region glycopeptides and C-terminal glycopeptides derived from the crosslinked regions of H1 and H2 from human CSF and plasma samples. In most cases, both heavy chains were simultaneously attached to the same CS-stretch and near each other. Additionally, we could also identify H1 cross-linking glycopeptides at the very non-reducing end of the CS chain. A summary of all our findings for the linkage and cross-linkage region of native bikunin are presented in Fig 12.

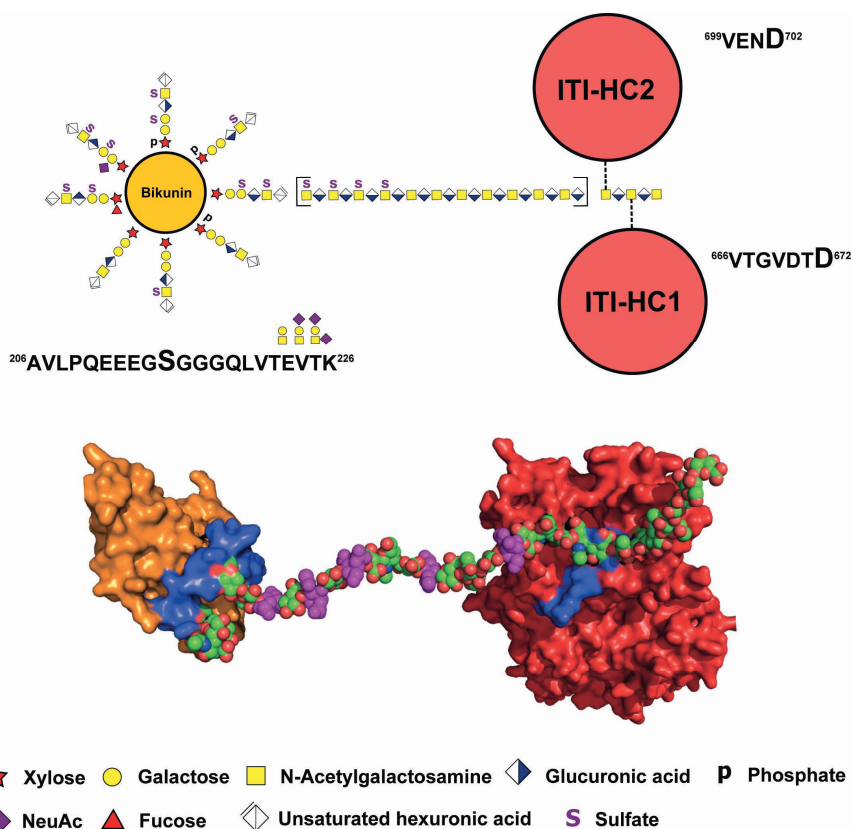


Figure 12. Summary of all the findings presented in this thesis about the structure of the inter- $\alpha$ -trypsin inhibitor complex derived from human plasma and cerebrospinal fluid. The light chain bikunin displays a broad heterogeneity of its linkage region. The heavy chains 1 and 2 are located towards the non-reducing end in close vicinity to each other.

## 4.4 Targeting biology

### *Identification of specific remodelling of the CS-glycosylation of iPS cells using a unique malaria protein probe*

Proteoglycans are deeply involved in key developmental processes and as mentioned before, a specific instructional role for HS during vertebrate development has been suggested (Kraushaar et al., 2013; Poulain & Yost, 2015). Equally, HS GAG chains have been implicated in the regulation of stem cell behaviour and differentiation (Mikami & Kitagawa, 2016; Tamm et al., 2012). The role of CS during these processes has been less well explored. Recent findings indicate that deregulation of normal stem cell functions is tightly linked to tumorigenesis and malignant cellular transformation (Shackleton, 2010). As acidic GAG polysaccharides are important for shaping both the stem cell niches and the tumour microenvironments we hypothesized that specific proteoglycan signatures might be shared between particular stem cell populations and tumour cells.

In Paper V, we used a unique recombinant protein probe (rVAR2) derived from the malaria VAR2CSA protein that specifically interacts with distinct CSA populations in the placenta as well as in a broad array of human tumour cells. We demonstrated that rVAR2 affinity purification in conjunction with a glyco-proteomics workflow is a feasible way of obtaining site-specific glycan information on the selected CSA-populations targeted by the probe. In sharp contrast to the placental setting where rVAR2-reactive CS-chains are specifically assembled on Syndecan-1, we found that a heterogeneous group of core proteins are the targets for CSA-attachment in tumour cells of different origins. As the placenta-type CSA modifications in cancer cells are involved in tumour motility (Clausen et al., 2016), it is possible that an increase in GAG presentation by mobilizing several core-proteins directly mediates or facilitates the acquirement of a more aggressive and invasive tumour phenotype. In line with our initial hypothesis, we also observed that rVAR2-reactive CSPGs are present in iPS cells. Furthermore, the absence of reactivity in the parental somatic cell line indicated that this expression was induced by cellular reprogramming and might therefore be under stage-specific regulation. Similar to the cancer setting, the assembly of the placenta-type CSA in iPS cells occurred on multiple core proteins but their specific functional roles remain to be determined.

The occurrence of dramatic changes in glycosylation during cellular transitions has been reported. These changes are not only restricted to proteoglycans but also encompass other glycoconjugates such as N- and O-glycoproteins and glycolipids. For example, the cell surface glycosphingolipids SSEA-3 and -4 are

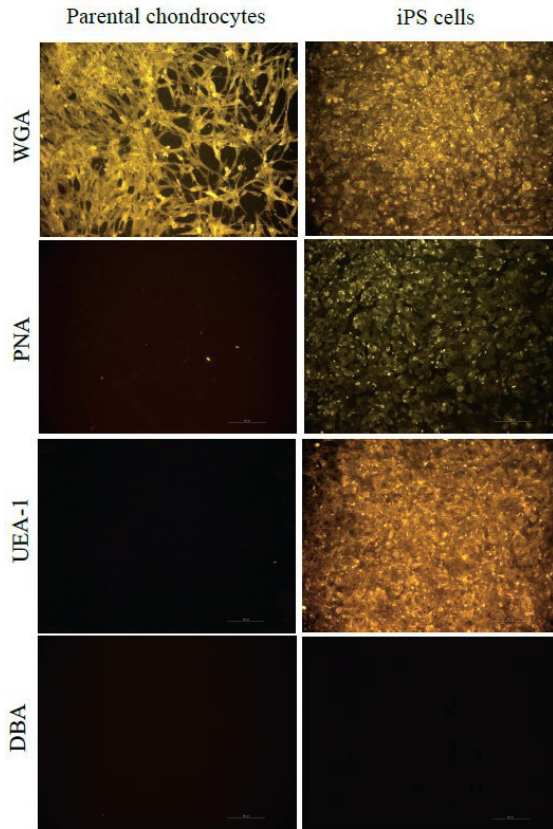


Figure 13. Lectin staining of a human chondrocyte cell line and iPS cells derived therefrom showing distinct differences in glycophenotypes. WGA: Wheat germ agglutinin, PNA: Peanut agglutinin, UEA-1: *Ulex europaeus* agglutinin 1, and DBA: *Dolichos biflorus* agglutinin.

stage-specific carbohydrate antigens that are useful for the identification of cell populations with stem cell traits (A. L. Yu et al., 2016). Pluripotent stem cells are also characterized by an increase in high mannose N-glycan structures compared to fully differentiated cells, whereas complex type N-glycans are much more abundant in differentiated cells (Fujitani et al., 2013; Furukawa et al., 2016). In some cases, this glycan remodelling has been shown to follow differential transcriptional changes in the glycosylation machinery (Nairn et al., 2012). By using a panel of carbohydrate-binding lectins to stain chondrocyte-derived iPS cells and their parental articular chondrocytes, we have also confirmed that major changes in the expression of alpha1,2-fucose (as recognized by UEA-1)

and the T-antigen Gal $\beta$ 1,3GalNAc $\alpha$ 1 -Ser/Thr (as recognized by PNA) are also associated with pluripotent cellular states (Fig 13). The remodelling of the cellular glycosylation landscape during cellular transitions, and particularly during differentiation, is the result of the complex interplay between dynamic signalling and transcriptional waves. If these resulting “glycowaves” are a mere reflection of the transcriptional state of the cell or if they are contributing factors that actively facilitate overcoming kinetic barriers during cellular transitions, is an important question that remains to be elucidated.

## 5. Conclusions

The general aim of this thesis was to develop comprehensive and high-throughput methods for the structural characterization of native CS/DS- and HS-proteoglycans from complex biological samples. Given the chemical complexity of these molecules not all structural aspects could be simultaneously addressed. Since the discovery of the proteoglycans, a broad array of different techniques has been developed to obtain the average structure of GAG chains as well as to get insights into their sulfation patterns. These advances have facilitated our understanding of the biological importance of the proteoglycans and their impact in human health and disease. However, the number of known proteoglycans appears to be limited and the GAG attachment sites on the core proteins have been difficult to access. In this thesis, we have focused on these two aspects by developing new glycoproteomics concepts and tools.

Taken together, our results indicate that GAG-attachment occurs on a broader spectrum of core proteins than was previously known and expected. We found that several human pro-hormones are *de facto* CS-proteoglycans, suggesting a general function for acidic GAGs in the storage and/or processing of some cargo proteins within secretory vesicles. By analysing multiple, experimentally determined GAG sites, we also show that the human xylosyltransferases act on semi-consensus motifs although these sequences are not as well defined as previously assumed. Additionally, some specific GAG sites display a hybrid nature, allowing for the simultaneous presentations of different GAGs types at the same amino acid position.

We also discovered that the proteoglycan linkage region is the target for unexpected glycan modifications such as sialylation and fucosylation. This finding indicates that there are other glycosyltransferases involved in the biosynthesis of these complex glycans, which might have implications for their biosynthetic regulation. Finally, we demonstrated that placental-type CS-chains are expressed in iPS cells in a stage specific manner. By using the methods described in this thesis, it was also possible to address their GAG attachment site-specificity in a high throughput fashion. It is our view that these glycoproteomics methods and techniques are to be considered complementary to more established glycan-oriented approaches. Together, they will help the biochemical community to understand all the intricate structural complexity of the proteoglycans and hopefully, facilitate the full exploration of the role of these glyconju-

gates in basic biology and stem cell research as well as in clinical settings, serving as novel biomarkers or as relevant therapeutic targets.

## 6. Future Perspectives

The way natural sciences have evolved in western tradition has shaped our idea of science as an experimental method, rather than as a well-defined intellectual product. “The scientific method”, this iterative procedure of collecting and measuring observations about nature to test and modify hypothesis, has also all the ingredients of a truly intellectual process. Therefore, it is very common that sound scientific work revolves around finding new answers to old problems as well as raising new questions from new findings.

Given our finding that some pro-hormones are carriers of GAG modification, it would be interesting to assess if other pro-hormones are also decorated with GAG chains. The assessment of the extent and biological function of such modifications in human endocrine cells can give new mechanistic insights into multiple diseases, ranging from diabetes to endocrine tumours. GAG chains are highly acidic and may aid in the storage of positively charged cargo molecules within secretory vesicles. The importance of such glycan modification for the specific function of the granins during vesicle biogenesis has so far not been addressed, but it is expected to play an important role. The methods developed during this thesis in conjunction with genetic deconstruction of the GAG biosynthesis may help to address these fundamental questions in relevant cell systems.

The presence of oncofetal CS-modifications in iPS cells also needs further clarification. To determine if this GAG expression is a general feature of pluripotent stem cells, a broader number of iPS and ES cells need to be tested. A closer evaluation of the expression of oncofetal CS in relation to pluripotency and differentiation markers is also required as well as a deep mechanistic understanding of their function for self-renewal and lineage specification. There is an appealing possibility that this GAG expression may constitute a molecular window to deliver specific signals to the stem cells to generate homogenous clinical-grade stem cell cultures for safe clinical purposes. Additionally, this finding should be put in relation to the broad expression of these CS-chains in multiple human tumours. How exactly deregulation of stem cell functions is connected to carcinogenesis is currently not completely understood. However, new insights may come from the study of how changes in the glycosylation affect the kinetics of cellular transitions.

There are still some basic questions regarding the structural diversity and regulation of mammalian proteoglycans that need further exploration. Our ob-



servation of novel glycan modifications at the proteoglycan linkage region clearly revealed the possible existence of yet unknown glycosyltransferases involved in GAG biosynthesis. Their identities, enzymatic activities and potential roles in the regulation of GAG assembly should also be addressed in future studies. Finally, the prospect of achieving a global site-specific analysis of full-length proteoglycans in complex samples is a long-term goal that still remains a challenge. It is also possible that thematic variations of the methodologies presented in this thesis may help to address structural studies of other classes of proteoglycans e.g. the keratan sulfate and glycosylphosphatidylinositol anchored proteoglycans. However, high throughput exploration of CS/DS- and HS- proteoglycans in biological samples can already, as described here, be accurately performed on generally available, sensitive and robust analytical platforms.

# Acknowledgement

Every intellectual journey is unique. Although uncustomary, before paying my respects and gratitude to some of the key persons that have facilitated this personal voyage, I am obliged to acknowledge that intimate circle that constitutes my “raison d'être”: my beloved wife and my beautiful boy. Before you, my dear **Maria**, my garden was a silent and rocky desert. You moved in with all your flowers and fertilized my soil with the sweet fragrance of a conjugal dialogue. Thanks for tangling your life to mine, and for connecting your beginning to my end, like an indissoluble rope. All my love to my son **Gabriel** that moved into our garden like a nightingale. I hope to be there by the rosebush enjoying your melodious singing until my last day. My infinite gratitude to my parents-in law **Juan** and **Maria** for building a protective wall around my garden. Thanks for the water and the shadow that made me grow in harmony, for the very first time in my life.

Special thanks to my supervisor **Göran Larson** for showing me the woods of knowledge beyond the garden wall. For teaching me how to discern between the sweet tree and the bitter shrubs. For drawing a map so I didn't go astray, for support and guidance. I don't know how to repay my debt except by offering you life-lasting friendship.

Thanks to my co-supervisor **Stina Simonsson** for sharing knowledge and enthusiasm. For introducing me to the fascinating realm of cellular metamorphoses. For showing me that it is possible to grow flowers from solid stones like bone from stem cells. Also, my gratitude to **Anders Lindahl** and **Anders Oldfors** for excellent co-supervision and mentorship. Thanks to **Camilla Brantsing** and **Alma Forsman** for using their talent and skills to help me with cells and mRNA-analysis.

My intellectual escapades wouldn't have been half so exciting without a constant scientific exchange with my colleagues and friends at the Glycobiology group: **Waqas Nasir**, **Jonas Nilsson**, and **Fredrik Noborn** as well as **Adnan Halim**, **Inger Johansson** and **Angelika Kunze**. Thanks for every word that became a seed and every seed that flourished into scientific work. My best wishes to **Inga Rimkute** and **Mahnaz Nikpour**, the new “bees” of our group. I am sure you will continue pollinating intellectual flowers and spreading scientific knowledge.

I also owe a big debt of gratitude to my “dreamcatchers”, colleagues and friends: **Camilla Hesse** and **Erika Lindberg**. Thanks for giving me a free pass to fulfill an old dream: teaching in the gardens of the Holy Land. Thanks to you I be-

friended two wonderful gardeners of souls: **Dr. Hatem Khaled** and **Dr. Yumna Shehadeh** as well as all the talented biomedical students at Al-Quds University, that as the rest of the Palestinian people still struggle against an unjust and shameful colonial situation.

Many thanks to all my collaborators and co-authors: **Marina B. Ayres Pereira**, **Thomas Clausen** and **Ali Salanti** at the University of Copenhagen; **Andrea Persson**, **Katrin Mani** and **Ulf Ellervik** at Lund University; **Lena Kjellén** and **Erik Fries** at Uppsala University; **Mingxun Wang** and **Nuno Bandeira** at the University of California, San Diego (UCSD) as well as **Carola Hedberg**, **Bertil Macao** and **Petra Bergström** here at Gothenburg University. Thanks for the opportunities, inspiration and splendid exchange of material and thoughts.

It is impossible to do gardening without tools so the very skeleton of my thesis has only been possible thanks to the efforts of extraordinary talented scientists at the Proteomics Core Facility at Gothenburg University: **Carina Sihlbom**, **Egor Vorontsov**, **Annika Thorsell** and **Johan Lenggqvist**. Thanks for your ideas, assistance and co-authorship.

Special love to my old chemistry teacher, friend and adoptive grandpa: **Bo Karlsson**. You are the very reason I embarked into research!

Also, my gratitude to my new mentor **Jeffrey D Esko** at UCSD for showing me that if you close your eyes in the woods you can hear the ruminant waters of the Pacific Ocean. Thanks for inviting me to join your group and to explore the sweet side of the ocean under your guidance.

Thanks to all the agencies that have provided funding to this research: The Swedish Research Council and Governmental grants to the Sahlgrenska University Hospital.

Finally, a last thought to my home country Cuba, to my family and friends there, and to that entire generation of Cubans that are now spread around the globe. I really hope that the inevitable flowing back of the tide can gather all our experiences and dreams and make them converge into a better future for our country, as the water return to the sea.

# References

- Abrink, M., Grujic, M., & Pejler, G. (2004). Serglycin is essential for maturation of mast cell secretory granule. *Journal of Biological Chemistry*, 279(39), 40897-40905.
- Aroso, M., Agricola, B., Hacker, C., & Schrader, M. (2015). Proteoglycans support proper granule formation in pancreatic acinar cells. *Histochem Cell Biol*, 144(4), 331-346.
- Arrington, C. B., Peterson, A. G., & Yost, H. J. (2013). Sdc2 and Tbx16 regulate Fgf2-dependent epithelial cell morphogenesis in the ciliated organ of asymmetry. *Development*, 140(19), 4102-4109.
- Aviezer, D., Hecht, D., Safran, M., Eisinger, M., David, G., *et al.* (1994). Perlecan, Basal Lamina Proteoglycan, Promotes Basic Fibroblast Growth Factor-Receptor Binding, Mitogenesis, and Angiogenesis. *Cell*, 79(6), 1005-1013.
- Aviezer, D., Iozzo, R. V., Noonan, D. M., & Yayon, A. (1997). Suppression of autocrine and paracrine functions of basic fibroblast growth factor by stable expression of perlecan antisense cDNA. *Mol Cell Biol*, 17(4), 1938-1946.
- Ayres Pereira, M., Mandel Clausen, T., Pehrson, C., Mao, Y., Resende, M., *et al.* (2016). Placental Sequestration of Plasmodium falciparum Malaria Parasites Is Mediated by the Interaction Between VAR2CSA and Chondroitin Sulfate A on Syndecan-1. *PLoS Pathog*, 12(8), e1005831.
- Bandeira, N. (2007). Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications. *Biotechniques*, 42(6), 687, 689, 691 passim.
- Barbosa, J. A., Gill, B. M., Takiyyuddin, M. A., & Oconnor, D. T. (1991). Chromogranin-a - Posttranslational Modifications in Secretory Granules. *Endocrinology*, 128(1), 174-190.
- Bartlett, A. H., & Park, P. W. (2010). Proteoglycans in host-pathogen interactions: molecular mechanisms and therapeutic implications. *Expert Rev Mol Med*, 12.
- Bartolomucci, A., Possenti, R., Mahata, S. K., Fischer-Colbrie, R., Loh, Y. P., *et al.* (2011). The Extended Granin Family: Structure, Function, and Biomedical Implications. *Endocr Rev*, 32(6), 755-797.
- Bielik, A. M., & Zaia, J. (2010). Historical overview of glycoanalysis. *Methods Mol Biol*, 600, 9-30.
- Bix, G., Fu, J., Gonzalez, E. M., Macro, L., Barker, A., *et al.* (2004). Endorepellin causes endothelial cell disassembly of actin cytoskeleton and focal adhesions through alpha 2 beta 1 integrin. *Journal of Cell Biology*, 166(1), 97-109.

- Bornemann, D. J., Duncan, J. E., Staatz, W., Selleck, S., & Warrior, R. (2004). Abrogation of heparan sulfate synthesis in *Drosophila* disrupts the Wingless, Hedgehog and Decapentaplegic signaling pathways. *Development*, *131*(9), 1927-1938.
- Brabin, B. J., Romagosa, C., Abdelgalil, S., Menendez, C., Verhoeff, F. H., *et al.* (2004). The sick placenta-the role of malaria. *Placenta*, *25*(5), 359-378.
- Braga, T., Grujic, M., Lukinius, A., Hellman, L., Abrink, M., *et al.* (2007). Serglycin proteoglycan is required for secretory granule integrity in mucosal mast cells. *Biochemical Journal*, *403*, 49-57.
- Brinkmann, T., Weilke, C., & Kleesiek, K. (1997). Recognition of acceptor proteins by UDP-D-xylose proteoglycan core protein beta-D-xylosyltransferase. *The Journal of biological chemistry*, *272*(17), 11171-11175.
- Buchner, Eduard. (1897). Alkoholische Gährung ohne Hefezellen (Vorläufige Mitteilung). *Berichte der Deutschen Chemischen Gesellschaft*(30), 117-124.
- Cailhier, J. F., Sirois, I., Laplante, P., Lepage, S., Raymond, M. A., *et al.* (2008). Caspase-3 activation triggers extracellular cathepsin L release and endorepellin proteolysis. *Journal of Biological Chemistry*, *283*(40), 27220-27229.
- Chen, L., Zhang, H., Powers, R. W., Russell, P. T., & Larsen, W. J. (1996). Covalent linkage between proteins of the inter-alpha-inhibitor family and hyaluronic acid is mediated by a factor produced by granulosa cells. *Journal of Biological Chemistry*, *271*(32), 19409-19414.
- Chi, L. L., Wolff, J. J., Laremore, T. N., Restaino, O. F., Xie, J., *et al.* (2008). Structural analysis of bikunin glycosaminoglycan. *J Am Chem Soc*, *130*(8), 2617-2625.
- Clausen, T. M., Pereira, M. A., Al Nakouzi, N., Oo, H. Z., Agerbaek, M. O., *et al.* (2016). Oncofetal Chondroitin Sulfate Glycosaminoglycans are Key Players in Integrin Signaling and Tumor Cell Motility. *Mol Cancer Res*.
- Cohen, I. R., Murdoch, A. D., Naso, M. F., Marchetti, D., Berd, D., *et al.* (1994). Abnormal Expression of Perlecan Proteoglycan in Metastatic Melanomas. *Cancer Res*, *54*(22), 5771-5774.
- Conte, M., Arcaro, A., D'Angelo, D., Gnata, A., Mamone, G., *et al.* (2006). A single chondroitin 6-sulfate oligosaccharide unit at Ser-2730 of human thyroglobulin enhances hormone formation and limits proteolytic accessibility at the carboxyl terminus - Potential insights into thyroid homeostasis and autoimmunity. *Journal of Biological Chemistry*, *281*(31), 22200-22211.
- Desaire, H., & Leary, J. A. (2000). Detection and quantification of the sulfated disaccharides in chondroitin sulfate by electrospray tandem mass spectrometry. *J Am Soc Mass Spectr*, *11*(10), 916-920.
- Dierker, T., Shao, C., Haitina, T., Zaia, J., Hinas, A., *et al.* (2016). Nematodes join the family of chondroitin sulfate-synthesizing organisms: Identification of an active chondroitin sulfotransferase in *Caenorhabditis elegans*. *Scientific reports*, *6*, 34662.

- Dodds, E. D. (2012). Gas-phase dissociation of glycosylated peptide ions. *Mass Spectrom Rev*, 31(6), 666-682.
- Douglass, S., Goyal, A., & Iozzo, R. V. (2015). The role of perlecan and endorepellin in the control of tumor angiogenesis and endothelial cell autophagy. *Connect Tissue Res*, 56(5), 381-391.
- Enghild, J. J., Thogersen, I. B., Pizzo, S. V., & Salvesen, G. (1989). Analysis of inter-alpha-trypsin inhibitor and a novel trypsin inhibitor, pre-alpha-trypsin inhibitor, from human plasma. Polypeptide chain stoichiometry and assembly by glycan. *The Journal of biological chemistry*, 264(27), 15975-15981.
- Ernst, S., Langer, R., Cooney, C. L., & Sasisekharan, R. (1995). Enzymatic Degradation of Glycosaminoglycans. *Crit Rev Biochem Mol*, 30(5), 387-444.
- Esko, J. D., & Selleck, S. B. (2002). Order out of chaos: Assembly of ligand binding sites in heparan sulfate. *Annu Rev Biochem*, 71, 435-471.
- Esko, J. D., & Zhang, L. J. (1996). Influence of core protein sequence on glycosaminoglycan assembly. *Curr Opin Struc Biol*, 6(5), 663-670.
- Fedarko, N. S. (1994). Isolation and purification of proteoglycans. *EXS*, 70, 9-35.
- Ferreras, C., Rushton, G., Cole, C. L., Babur, M., Telfer, B. A., *et al.* (2012). Endothelial Heparan Sulfate 6-O-Sulfation Levels Regulate Angiogenic Responses of Endothelial Cells to Fibroblast Growth Factor 2 and Vascular Endothelial Growth Factor. *Journal of Biological Chemistry*, 287(43), 36132-36146.
- Fischer, Emil. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Ber Dtsch Chem Ges*(27), 2985-2993.
- Fischer, G. and Boedeker, C. (1861). Künstliche Bildung von Zucker aus Knorpel (Chondrogen), und über die Umsetzung des genossenen Knorpels im menschlichen Körper *Ann. Chem. Pharm*, 117, 111-118.
- Freeze, H. H., Chong, J. X., Bamshad, M. J., & Ng, B. G. (2014). Solving glycosylation disorders: fundamental approaches reveal complicated pathways. *Am J Hum Genet*, 94(2), 161-175.
- Fried, M., & Duffy, P. E. (1996). Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta. *Science*, 272(5267), 1502-1504.
- Fujitani, N., Furukawa, J., Araki, K., Fujioka, T., Takegawa, Y., *et al.* (2013). Total cellular glycomics allows characterizing cells and streamlining the discovery process for cellular biomarkers. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6), 2105-2110.
- Fukuta, M., Uchiyama, K., Nakashima, K., Kato, M., Kimata, K., *et al.* (1995). Sulfotransferases in Biosynthesis of Glycosaminoglycan. *Trends Glycosci Glyc*, 7(38), 527-528.
- Furukawa, J., Okada, K., & Shinohara, Y. (2016). Glycomics of human embryonic stem cells and human induced pluripotent stem cells. *Glycoconjugate journal*, 33(5), 707-715.

- Fuster, M. M., & Esko, J. D. (2005). The sweet and sour of cancer: Glycans as novel therapeutic targets. *Nat Rev Cancer*, 5(7), 526-542.
- Fuster, M. M., Wang, L. C., Castagnola, J., Sikora, L., Reddi, K., *et al.* (2007). Genetic alteration of endothelial heparan sulfate selectively inhibits tumor angiogenesis. *Journal of Cell Biology*, 177(3), 539-549.
- Gonzalez, E. M., Reed, C. C., Bix, G., Fu, J., Zhang, Y., *et al.* (2005). BMP-1/tolloid-like metalloproteases process endorepellin, the angiostatic C-terminal fragment of perlecan. *Journal of Biological Chemistry*, 280(8), 7080-7087.
- Goth, C. K., Halim, A., Khetarpal, S. A., Rader, D. J., Clausen, H., *et al.* (2015). A systematic study of modulation of ADAM-mediated ectodomain shedding by site-specific O-glycosylation. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47), 14623-14628.
- Gowda, D. C., Hogueangeletti, R., Margolis, R. K., & Margolis, R. U. (1990). Chromaffin Granule and Pc12 Cell Chondroitin Sulfate Proteoglycans and Their Relation to Chromogranin-A. *Arch Biochem Biophys*, 281(2), 219-224.
- Gulberti, S., Jacquinet, J. C., Chabel, M., Ramalanjaona, N., Magdalou, J., *et al.* (2012). Chondroitin sulfate N-acetylgalactosaminyltransferase-1 (CSGalNAcT-1) involved in chondroitin sulfate initiation: Impact of sulfation on activity and specificity. *Glycobiology*, 22(4), 561-571.
- Gulberti, S., Lattard, V., Fondeur, M., Jacquinet, J. C., Mullier, G., *et al.* (2005). Phosphorylation and sulfation of oligosaccharide substrates critically influence the activity of human beta 1,4-galactosyltransferase 7 (GalT-I) and beta 1,3-glucuronosyltransferase I (GlcAT-I) involved in the biosynthesis of the glycosaminoglycan-protein linkage region of Proteoglycans. *Journal of Biological Chemistry*, 280(2), 1417-1425.
- Guthals, A., Watrous, J. D., Dorrestein, P. C., & Bandeira, N. (2012). The spectral networks paradigm in high throughput mass spectrometry. *Mol Biosyst*, 8(10), 2535-2544.
- Hacker, U., Nybakken, K., & Perrimon, N. (2005). Heparan sulphate proteoglycans: the sweet side of development. *Nature reviews. Molecular cell biology*, 6(7), 530-541.
- Haerry, T. E., Heslip, T. R., Marsh, J. L., & OConnor, M. B. (1997). Defects in glucuronate biosynthesis disrupt wingless signaling in *Drosophila*. *Development*, 124(16), 3055-3064.
- Halim, A., Westerlind, U., Pett, C., Schorlemer, M., Ruetschi, U., *et al.* (2014). Assignment of Saccharide Identities through analysis of oxonium ion fragmentation profiles in LC-MS/MS of glycopeptides. *Journal of proteome research*, 13(12), 6024-6032.
- Haltom, A. R., & Jafar-Nejad, H. (2015). The multiple roles of epidermal growth factor repeat O-glycans in animal development. *Glycobiology*, 25(10), 1027-1042.

- Hayashi, Y., Sexton, T. R., Dejima, K., Perry, D. W., Takemura, M., *et al.* (2012). Glypicans regulate JAK/STAT signaling and distribution of the Unpaired morphogen. *Development*, *139*(22), 4162-4171.
- Hein, G. E. (1961). Liebig-Pasteur Controversy - Vitality without Vitalism. *J Chem Educ*, *38*(12), 614-&.
- Helting, T., & Roden, L. (1968). The carbohydrate-protein linkage region of chondroitin 6-sulfate. *Biochim Biophys Acta*, *170*(2), 301-308.
- Hennet, T., & Cabalzar, J. (2015). Congenital disorders of glycosylation: a concise chart of glycocalyx dysfunction. *Trends Biochem Sci*, *40*(7), 377-384.
- Hinneburg, H., Stavenhagen, K., Schweiger-Hufnagel, U., Pengelley, S., Jabs, W., *et al.* (2016). The Art of Destruction: Optimizing Collision Energies in Quadrupole-Time of Flight (Q-TOF) Instruments for Glycopeptide-Based Glycoproteomics. *J Am Soc Mass Spectr*, *27*(3), 507-519.
- Hirsh, J., Anand, S. S., Halperin, J. L., Fuster, V., & American Heart Association. (2001). AHA Scientific Statement: Guide to anticoagulant therapy: heparin: a statement for healthcare professionals from the American Heart Association. *Arterioscler Thromb Vasc Biol*, *21*(7), E9-9.
- Hofmann, J., Hahm, H. S., Seeberger, P. H., & Pagel, K. (2015). Identification of carbohydrate anomers using ion mobility-mass spectrometry. *Nature*, *526*(7572), 241-+.
- Hu, Q. Z., Noll, R. J., Li, H. Y., Makarov, A., Hardman, M., *et al.* (2005). The Orbitrap: a new mass spectrometer. *J Mass Spectrom*, *40*(4), 430-443.
- Huang, K., & Wu, L. D. (2008). Aggrecanase and Aggrecan Degradation in Osteoarthritis: a Review. *J Int Med Res*, *36*(6), 1149-1160.
- Hwang, H. Y., Olson, S. K., Esko, J. D., & Horvitz, H. R. (2003). *Caenorhabditis elegans* early embryogenesis and vulval morphogenesis require chondroitin biosynthesis. *Nature*, *423*(6938), 439-443.
- Inamori, K. I., Beedle, A. M., de Bernabe, D. B., Wright, M. E., & Campbell, K. P. (2016). LARGE2-dependent glycosylation confers laminin-binding ability on proteoglycans. *Glycobiology*.
- Iozzo, R. V., & Murdoch, A. D. (1996). Proteoglycans of the extracellular environment: Clues from the gene and protein side offer novel perspectives in molecular diversity and function. *Faseb Journal*, *10*(5), 598-614.
- Iozzo, R. V., & Schaefer, L. (2015). Proteoglycan form and function: A comprehensive nomenclature of proteoglycans. *Matrix Biol*, *42*, 11-55.
- Izumikawa, T., Dejima, K., Watamoto, Y., Nomura, K. H., Kanaki, N., *et al.* (2016). Chondroitin 4-O-Sulfotransferase Is Indispensable for Sulfation of Chondroitin and Plays an Important Role in Maintaining Normal Life Span and Oxidative Stress Responses in Nematodes. *The Journal of biological chemistry*, *291*(44), 23294-23304.
- Izumikawa, T., Kanagawa, N., Watamoto, Y., Okada, M., Saeki, M., *et al.* (2010). Impairment of embryonic cell division and glycosaminoglycan



- biosynthesis in glucuronyltransferase-I-deficient mice. *The Journal of biological chemistry*, 285(16), 12190-12196.
- Izumikawa, T., & Kitagawa, H. (2015). Amino acid sequence surrounding the chondroitin sulfate attachment site of thrombomodulin regulates chondroitin polymerization. *Biochemical and biophysical research communications*, 460(2), 233-237.
- Izumikawa, T., Koike, T., Shiozawa, S., Sugahara, K., Tamura, J. I., *et al.* (2008). Identification of chondroitin sulfate glucuronyltransferase as chondroitin synthase-3 involved in chondroitin polymerization - Chondroitin polymerization is achieved by multiple enzyme complexes consisting of chondroitin synthase family members. *Journal of Biological Chemistry*, 283(17), 11396-11406.
- Izumikawa, T., Okuura, Y., Koike, T., Sakoda, N., & Kitagawa, H. (2011). Chondroitin 4-O-sulfotransferase-1 regulates the chain length of chondroitin sulfate in co-operation with chondroitin N-acetylgalactosaminyltransferase-2. *The Biochemical journal*, 434(2), 321-331.
- Izumikawa, T., Sato, B., & Kitagawa, H. (2014). Chondroitin sulfate is indispensable for pluripotency and differentiation of mouse embryonic stem cells. *Scientific reports*, 4, 3701.
- Jorpes, E., and Gardell, S. (1948). On heparin monosulphuric acid. *J. biol. Chem*(176), 267-276.
- Juhasz, P., & Biemann, K. (1994). Mass-Spectrometric Molecular-Weight Determination of Highly Acidic Compounds of Biological Significance Via Their Complexes with Basic Polypeptides. *Proceedings of the National Academy of Sciences of the United States of America*, 91(10), 4333-4337.
- Juhasz, P., & Biemann, K. (1995). Utility of Noncovalent Complexes in the Matrix-Assisted Laser-Desorption Ionization Mass-Spectrometry of Heparin-Derived Oligosaccharides. *Carbohydr Res*, 270(2), 131-147.
- Karas, M., Bahr, U., & Dulcks, T. (2000). Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. *Fresen J Anal Chem*, 366(6-7), 669-676.
- Kato, M., Wang, H. M., Bernfield, M., Gallagher, J. T., & Turnbull, J. E. (1994). Cell-Surface Syndecan-1 on Distinct Cell-Types Differs in Fine-Structure and Ligand-Binding of Its Heparan-Sulfate Chains. *Journal of Biological Chemistry*, 269(29), 18881-18890.
- Kiani, C., Chen, L., Wu, Y. J., Yee, A. J., & Yang, B. B. (2002). Structure and function of aggrecan. *Cell Res*, 12(1), 19-32.
- Kleinschmit, A., Koyama, T., Dejima, K., Hayashi, Y., Kamimura, K., *et al.* (2010). Drosophila heparan sulfate 6-O endosulfatase regulates Wingless morphogen gradient formation. *Dev Biol*, 345(2), 204-214.
- Koike, T., Izumikawa, T., Sato, B., & Kitagawa, H. (2014). Identification of Phosphatase That Dephosphorylates Xylose in the Glycosaminoglycan-Protein Linkage Region of Proteoglycans. *Journal of Biological Chemistry*, 289(10), 6695-6708.

- Koike, T., Izumikawa, T., Tamura, J., & Kitagawa, H. (2009). FAM20B is a kinase that phosphorylates xylose in the glycosaminoglycan-protein linkage region. *The Biochemical journal*, 421(2), 157-162.
- Kraushaar, D. C., Dalton, S., & Wang, L. C. (2013). Heparan sulfate: a key regulator of embryonic stem cell fate. *Biol Chem*, 394(6), 741-751.
- Kusche-Gullberg, M., & Kjellen, L. (2003). Sulfotransferases in glycosaminoglycan biosynthesis. *Curr Opin Struc Biol*, 13(5), 605-611.
- Lever, R., & Page, C. P. (2002). Novel drug development opportunities for heparin. *Nat Rev Drug Discov*, 1(2), 140-148.
- Li, W., Johnson, D. J. D., Esmon, C. T., & Huntington, J. A. (2004). Structure of the antithrombin-thrombin-heparin ternary complex reveals the antithrombotic mechanism of heparin. *Nat Struct Mol Biol*, 11(9), 857-862.
- Lindahl, U. (1966). Further characterization of the heparin-protein linkage region. *Biochim Biophys Acta*, 130(2), 368-382.
- Lindahl, U. (1968). Glucuronic acid- and glucosamine-containing oligosaccharides from the heparin-protein linkage region. *Biochim Biophys Acta*, 156(1), 203-206.
- Lindahl, U. (2014). A personal voyage through the proteoglycan field. *Matrix Biol*, 35, 3-7.
- Lindahl, U., & Roden, L. (1966). The chondroitin 4-sulfate-protein linkage. *The Journal of biological chemistry*, 241(9), 2113-2119.
- Ly, M., Laremore, T. N., & Linhardt, R. J. (2010). Proteoglycomics: Recent Progress and Future Challenges. *OmicS*, 14(4), 389-399.
- Ly, M., Leach, F. E., Laremore, T. N., Toida, T., Amster, I. J., et al. (2011). The proteoglycan bikunin has a defined sequence. *Nat Chem Biol*, 7(11), 827-833.
- Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Anal Chem*, 72(6), 1156-1162.
- Malmstrom, A., Bartolini, B., Thelin, M. A., Pacheco, B., & Maccarana, M. (2012). Iduronic Acid in Chondroitin/Dermatan Sulfate: Biosynthesis and Biological Function. *Journal of Histochemistry & Cytochemistry*, 60(12), 916-925.
- Mann, D. M., Yamaguchi, Y., Bourdon, M. A., & Ruoslahti, E. (1990). Analysis of Glycosaminoglycan Substitution in Decorin by Site-Directed Mutagenesis. *Journal of Biological Chemistry*, 265(9), 5317-5323.
- McLean, J. (1916). The thromboplastic action of cephalin. *Am. J. Physiol.* (41 ), 250-257.
- Meselson, M., Stahl, F. W., & Vinograd, J. (1957). Equilibrium Sedimentation of Macromolecules in Density Gradients. *Proceedings of the National Academy of Sciences of the United States of America*, 43(7), 581-588.
- Meyer, K., and Chaffee, E. (1941). The mucopolysaccharides of skin. . *J. biol. Chem*(138), 491-499.
- Meyer, K., Davidson, E. A., Linker, A., and Hoffman, P. (1956). The acid mucopolysaccharides of connective tissue. *Biochim. biophys. Acta*(21), 506-518.

- Meyer, K., Linker, A., Davidson, E. A., and Weissmann, B. (1953). The mucopolysaccharides of bovine cornea. *J. Biol. Chem*(205 ), 611-616.
- Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., *et al.* (2011). Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol Cell Proteomics*, 10(9).
- Mikami, T., & Kitagawa, H. (2016). Sulfated glycosaminoglycans: their distinct roles in stem cell biology. *Glycoconjugate journal*.
- Mizuguchi, S., Uyama, T., Kitagawa, H., Nomura, K. H., Dejima, K., *et al.* (2003). Chondroitin proteoglycans are involved in cell division of *Caenorhabditis elegans*. *Nature*, 423(6938), 443-448.
- Mizumoto, S., Ikegawa, S., & Sugahara, K. (2013). Human Genetic Disorders Caused by Mutations in Genes Encoding Biosynthetic Enzymes for Sulfated Glycosaminoglycans. *Journal of Biological Chemistry*, 288(16), 10953-10961.
- Mongiat, M., Sweeney, S. M., San Antonio, J. D., Fu, J., & Iozzo, R. V. (2003). Endorepellin, a novel inhibitor of angiogenesis derived from the C terminus of perlecan. *Journal of Biological Chemistry*, 278(6), 4238-4249.
- Moses, J., Oldberg, A., Cheng, F., & Fransson, L. A. (1997). Biosynthesis of the proteoglycan decorin--transient 2-phosphorylation of xylose during formation of the trisaccharide linkage region. *European journal of biochemistry / FEBS*, 248(2), 521-526.
- Muenzer, J. (2011). Overview of the mucopolysaccharidoses. *Rheumatology*, 50, V4-V12.
- Muir, H. (1958). Nature of the Link between Protein and Carbohydrate of a Chondroitin Sulphate Complex from Hyaline Cartilage. *Biochemical Journal*, 69, 195-204.
- Myron, P., Siddiquee, S., & Al Azad, S. (2014). Fucosylated chondroitin sulfate diversity in sea cucumbers: A review. *Carbohydr Polym*, 112, 173-178.
- Nairn, A. V., Aoki, K., dela Rosa, M., Porterfield, M., Lim, J. M., *et al.* (2012). Regulation of Glycan Structures in Murine Embryonic Stem Cells Combined Transcript Profiling of Glycan-related Genes and Glycan Structural Analysis. *Journal of Biological Chemistry*, 287(45).
- Nairn, A. V., Kinoshita-Toyoda, A., Toyoda, H., Xie, J., Harris, K., *et al.* (2007). Glycomics of proteoglycan biosynthesis in murine embryonic stem cell differentiation. *Journal of proteome research*, 6(11), 4374-4387.
- Nguyen, D. D., Wu, C. H., Moree, W. J., Lamsa, A., Medema, M. H., *et al.* (2013). MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences of the United States of America*, 110(28), E2611-2620.
- Nilsson, J., Halim, A., Grahn, A., & Larson, G. (2013). Targeting the glycoproteome. *Glycoconjugate journal*, 30(2), 119-136.
- Nilsson, J., Ruetschi, U., Halim, A., Hesse, C., Carlsohn, E., *et al.* (2009). Enrichment of glycopeptides for glycan structure and attachment site identification. *Nature Methods*, 6(11), 809-U826.

- Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., *et al.* (2007). Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9), 709-712.
- Olson, S. K., Bishop, J. R., Yates, J. R., Oegema, K., & Esko, J. D. (2006). Identification of novel chondroitin proteoglycans in *Caenorhabditis elegans*: embryonic cell division depends on CPG-1 and CPG-2. *The Journal of cell biology*, 173(6), 985-994.
- Palmer, Karl Meyer and John W. (1934 ). The polysaccharide of the vitreous humor. *J. Biol. Chem.*(107), 629.
- Pan, Y., Carbe, C., Kupich, S., Pickhinke, U., Ohlig, S., *et al.* (2014). Heparan sulfate expression in the neural crest is essential for mouse cardiogenesis. *Matrix Biol*, 35, 253-265.
- Pasteur, Louis. Mémoire sur la fermentation alcoolique. *Annales de Chimie et de Physique*(58), 323-426.
- Poulain, F. E., & Yost, H. J. (2015). Heparan sulfate proteoglycans: a sugar code for vertebrate development? *Development*, 142(20), 3456-3467.
- Reeder, J. C., Cowman, A. F., Davern, K. M., Beeson, J. G., Thompson, J. K., *et al.* (1999). The adhesion of *Plasmodium falciparum*-infected erythrocytes to chondroitin sulfate A is mediated by P. *falciparum* erythrocyte membrane protein 1. *Proceedings of the National Academy of Sciences of the United States of America*, 96(9), 5198-5202.
- Ringvall, M., Roennberg, E., Wernersson, S., Duelli, A., Henningsson, F., *et al.* (2008). Serotonin and histamine storage in mast cell secretory granules is dependent on serglycin proteoglycan. *J Allergy Clin Immun*, 121(4), 1020-1026.
- Roch, C., Kuhn, J., Kleesiek, K., & Gotting, C. (2010). Differences in gene expression of human xylosyltransferases and determination of acceptor specificities for various proteoglycans. *Biochemical and biophysical research communications*, 391(1), 685-691.
- Roseman, S. (2001). Reflections on glycobiology. *Journal of Biological Chemistry*, 276(45), 41527-41542.
- Salanti, A., Clausen, T. M., Agerbaek, M. O., Al Nakouzi, N., Dahlback, M., *et al.* (2015). Targeting Human Cancer by a Glycosaminoglycan Binding Malaria Protein. *Cancer Cell*, 28(4), 500-514.
- Salier, J. P., Rouet, P., Raguenez, G., & Daveau, M. (1996). The inter-alpha-inhibitor family: from structure to regulation. *The Biochemical journal*, 315 (Pt 1), 1-9.
- Sarbu, M., Zhu, F. F., Peter-Katalinic, J., Clemmer, D. E., & Zamfir, A. D. (2015). Application of ion mobility tandem mass spectrometry to compositional and structural analysis of glycopeptides extracted from the urine of a patient diagnosed with Schindler disease. *Rapid Commun Mass Sp*, 29(21), 1929-1937.
- Scheltema, R. A., Hauschild, J. P., Lange, O., Hornburg, D., Denisov, E., *et al.* (2014). The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High-performance Quadrupole and an Ultra-high-field Orbitrap Analyzer. *Mol Cell Proteomics*, 13(12), 3698-3708.

- Scott, I. C., Imamura, Y., Pappano, W. N., Troedel, J. M., Recklies, A. D., *et al.* (2000). Bone morphogenetic protein-1 processes probiglycan. *Journal of Biological Chemistry*, 275(39), 30504-30511.
- Seno, N., & Sekizuka, E. (1978). Quantitative Beta-Elimination-Reduction of O-Glycosyl Linkages in Chondroitin Sulfates. *Carbohyd Res*, 62(2), 271-279.
- Shackleton, M. (2010). Normal stem cells and cancer stem cells: similar and different. *Semin Cancer Biol*, 20(2), 85-92.
- Shi, X. F., & Zaia, J. (2009). Organ-specific Heparan Sulfate Structural Phenotypes. *Journal of Biological Chemistry*, 284(18), 11806-11814.
- Sleno, L., & Volmer, D. A. (2004). Ion activation methods for tandem mass spectrometry. *J Mass Spectrom*, 39(10), 1091-1112.
- Staples, G. O., & Zaia, J. (2011). Analysis of Glycosaminoglycans Using Mass Spectrometry. *Curr Proteomics*, 8(4), 325-336.
- Stentoft, C., Vakhrushev, S. Y., Vester-Christensen, M. B., Schjoldager, K. T. B. G., Kong, Y., *et al.* (2011). Mining the O-glycoproteome using zinc-finger nuclease-glycoengineered SimpleCell lines. *Nature Methods*, 8(11), 977-982.
- Takagaki, K., Nakamura, T., Shibata, S., Higuchi, T., & Endo, M. (1996). Characterization and biological significance of sialyl alpha 2-3galactosyl beta 1-4Xylosyl beta 1-(4-methylumbelliferone) synthesized in cultured human skin fibroblasts. *J Biochem-Tokyo*, 119(4), 697-702.
- Tamm, C., Kjellen, L., & Li, J. P. (2012). Heparan sulfate biosynthesis enzymes in embryonic stem cell biology. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, 60(12), 943-949.
- Thaysen-Andersen, M., & Packer, N. H. (2014). Advances in LC-MS/MS-based glycoproteomics: Getting closer to system-wide site-specific mapping of the N- and O-glycoproteome. *Bba-Proteins Proteom*, 1844(9), 1437-1452.
- Tone, Y., Pedersen, L. C., Yamamoto, T., Izumikawa, T., Kitagawa, H., *et al.* (2008). 2-O-phosphorylation of xylose and 6-O-sulfation of galactose in the protein linkage region of Glycosaminoglycans influence the glucuronyltransferase-I activity involved in the linkage region synthesis. *Journal of Biological Chemistry*, 283(24), 16801-16807.
- Varki, A. (2011). Evolutionary forces shaping the Golgi glycosylation machinery: why cell surface glycans are universal to living cells. *Cold Spring Harb Perspect Biol*, 3(6).
- Varki, A., & Sharon, N. (2009). Historical Background and Overview. In A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, & M. E. Etzler (Eds.), *Essentials of Glycobiology* (2nd ed.). Cold Spring Harbor (NY).
- Wen, J. Z., Xiao, J. Y., Rahdar, M., Choudhury, B. P., Cui, J. X., *et al.* (2014). Xylose phosphorylation functions as a molecular switch to regulate

- proteoglycan biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 111(44), 15723-15728.
- Vester-Christensen, M. B., Halim, A., Joshi, H. J., Steentoft, C., Bennett, E. P., *et al.* (2013). Mining the O-mannose glycoproteome reveals cadherins as major O-mannosylated glycoproteins. *Proceedings of the National Academy of Sciences of the United States of America*, 110(52), 21018-21023.
- Whitelock, J. M., & Iozzo, R. V. (2002). Isolation and purification of proteoglycans. *Methods Cell Biol*, 69, 53-67.
- Wilm, M. S., & Mann, M. (1994). Electrospray and Taylor-Cone Theory, Doles Beam of Macromolecules at Last. *Int J Mass Spectrom*, 136(2-3), 167-180.
- Wilson, I. B. (2004). The never-ending story of peptide O-xylosyltransferase. *Cell Mol Life Sci*, 61(7-8), 794-809.
- von Marschall, Z., & Fisher, L. W. (2010). Decorin is processed by three isoforms of bone morphogenetic protein-1 (BMP1). *Biochemical and biophysical research communications*, 391(3), 1374-1378.
- Yabe, T., Hata, T., He, J., & Maeda, N. (2005). Developmental and regional expression of heparan sulfate sulfotransferase genes in the mouse brain. *Glycobiology*, 15(10), 982-993.
- Yamada, S., Oyama, M., Yuki, Y., Kato, K., & Sugahara, K. (1995). The uniform galactose 4-sulfate structure in the carbohydrate-protein linkage region of human urinary trypsin inhibitor. *European journal of biochemistry / FEBS*, 233(2), 687-693.
- Yan, D., & Lin, X. (2009). Shaping morphogen gradients by proteoglycans. *Cold Spring Harb Perspect Biol*, 1(3), a002493.
- Yoshida-Moriguchi, T., & Campbell, K. P. (2015). Matriglycan: a novel polysaccharide that links dystroglycan to the basement membrane. *Glycobiology*, 25(7), 702-713.
- Yu, A. L., Hung, J. T., Ho, M. Y., & Yu, J. (2016). Alterations of Glycosphingolipids in Embryonic Stem Cell Differentiation and Development of Glycan-Targeting Cancer Immunotherapy. *Stem Cells Dev*, 25(20), 1532-1548.
- Yu, J., Schorlemer, M., Gomez Toledo, A., Pett, C., Sihlbom, C., *et al.* (2016). Distinctive MS/MS Fragmentation Pathways of Glycopeptide-Generated Oxonium Ions Provide Evidence of the Glycan Structure. *Chemistry*, 22(3), 1114-1124.
- Zaia, J. (2009). On-Line Separations Combined with Ms for Analysis of Glycosaminoglycans. *Mass Spectrom Rev*, 28(2), 254-272.
- Zaia, J., & Costello, C. E. (2001). Compositional analysis of glycosaminoglycans by electrospray mass spectrometry. *Anal Chem*, 73(2), 233-239.
- Zhang, H., Li, X. J., Martin, D. B., & Aebersold, R. (2003). Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nature Biotechnology*, 21(6), 660-666.

- Zhang, L. J., David, G., & Esko, J. D. (1995). Repetitive Ser-Gly Sequences Enhance Heparan-Sulfate Assembly in Proteoglycans. *Journal of Biological Chemistry*, 270(45), 27127-27135.
- Zhang, L. J., & Esko, J. D. (1994). Amino-Acid Determinants That Drive Heparan-Sulfate Assembly in a Proteoglycan. *Journal of Biological Chemistry*, 269(30), 19295-19299.
- Zhang, X., Wang, F. S., & Sheng, J. Z. (2016). "Coding" and "Decoding": hypothesis for the regulatory mechanism involved in heparan sulfate biosynthesis. *Carbohydr Res*, 428, 1-7.
- Zhao, X., Yang, B., Solakyilidirim, K., Joo, E. J., Toida, T., *et al.* (2013). Sequence Analysis and Domain Motifs in the Porcine Skin Decorin Glycosaminoglycan Chain. *Journal of Biological Chemistry*, 288(13), 9226-9237.
- Zhou, Z. J., Wang, J. M., Cao, R. H., Morita, H., Soininen, R., *et al.* (2004). Impaired angiogenesis, delayed wound healing and retarded tumor growth in perlecan heparan sulfate-deficient mice. *Cancer Res*, 64(14), 4699-4702.

