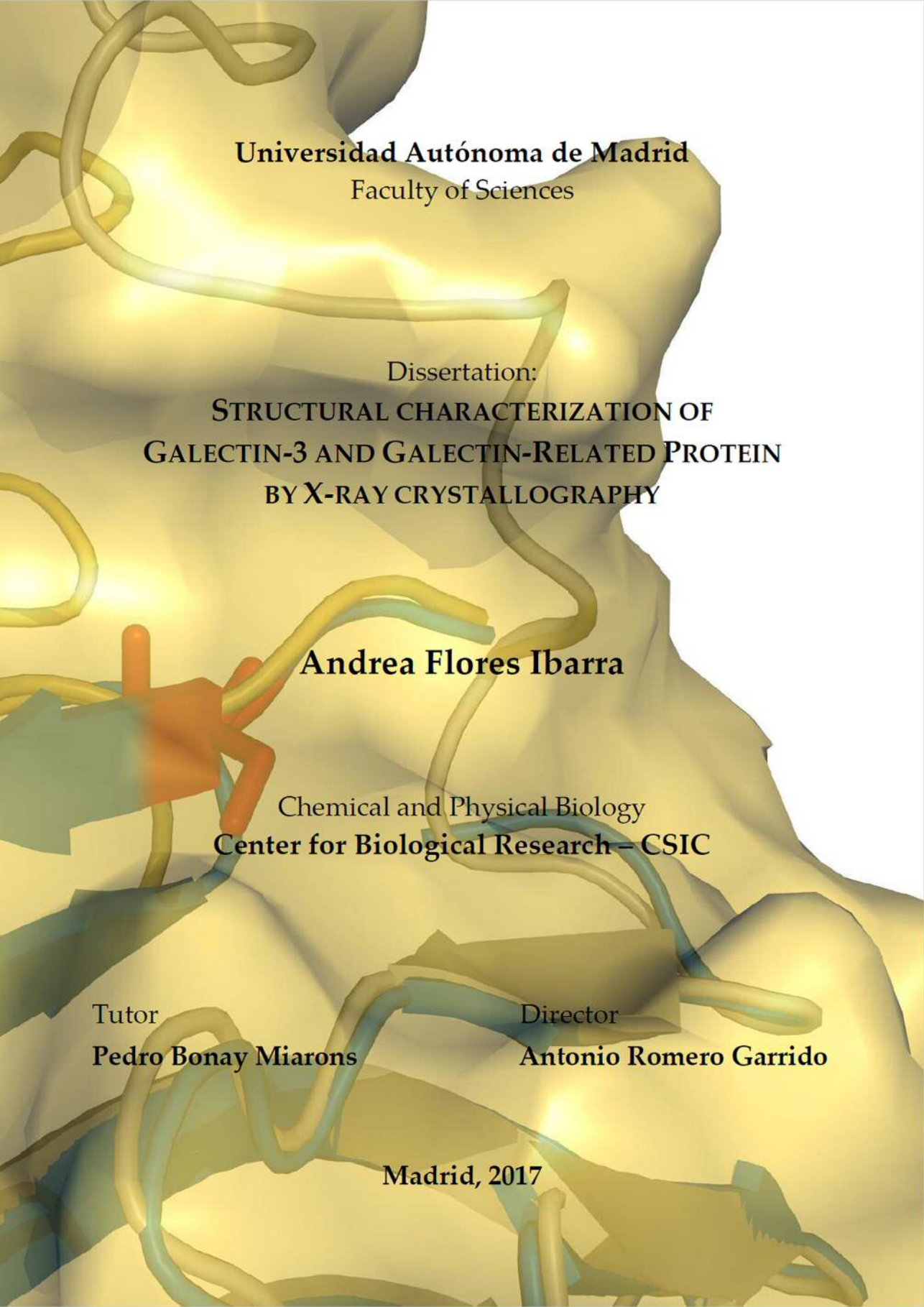


Universidad Autónoma de Madrid
Faculty of Sciences

Dissertation:

STRUCTURAL CHARACTERIZATION OF GALECTIN-3 AND GALECTIN-RELATED PROTEIN BY X-RAY CRYSTALLOGRAPHY

Andrea Flores Ibarra

Chemical and Physical Biology
Center for Biological Research – CSIC

Tutor
Pedro Bonay Miarons

Director
Antonio Romero Garrido

Madrid, 2017

¿Qué somos y cómo realizaremos eso que somos?
Octavio Paz, Nobel Prize of Literature 1980

"Te enterramos ayer.
Ayer te enterramos.
Te echamos tierra ayer.
Quedaste en la tierra ayer.
Estás rodeada de tierra
desde ayer.
Arriba y abajo y a los lados,
por tus pies y por tu cabeza,
está la tierra desde ayer.
Te metimos en la tierra,
te tapamos con tierra ayer.
Perteneces a la tierra
desde ayer.
Ayer te enterramos
en la tierra, ayer".

Jaime Sabines.

DEDICO ESTA TESIS A MI ABUE, POR SER MI PIEDRA DE TOQUE
Y LA MUJER QUE MÁS HE ADMIRADO
Cecilia Caballero Jiménez (1924-2014)

ACKNOWLEDGEMENTS

- To all my lab-mates in the Department of Physical and Chemical Biology (CIB), for your company through the years. To Antonio, for your guidance.
- To the people at the Institute of Chemistry (UNAM), for receiving me as part of your group.
- To all the people in the Glycopharm Marie Curie ITN, for your inspiration in my scientific research, I wish to have an everlasting relation with you.
- To my students, for your admiration and your will to never stop learning.
- To my best friends, scattered around the globe, I love you. To Sandy and Silvia in particular.
- To my siblings, and my nephew and niece, Alan and Talia. For your presence in my life. I love you deeply.
- And to all that were a part of my path in the obtention of this PhD, *para ustedes*.

OUTLINE

OUTLINE

List of abbreviations v

Glosary vi

ABSTRACT ix

RESUMEN xi

INTRODUCTION 1

1. CELL GLYCOSYLATION 1

2. GLYCAN-BINDING PROTEINS 3
   2.1. Types of lectins 3
      2.1.1. *R-type lectins* 4
      2.1.2. *L-type lectins* 4
      2.1.3. *P-type lectins* 5
      2.1.4. *C-type lectins* 5
      2.1.5. *I-type lectins* 6
      2.1.6. *Galectins* 6

3. THE GALECTIN FAMILY 7
   3.1. Galectins clasification and function 9

4. STRUCTURAL CHARACTERIZATION OF GALECTINS AND GRP 14
   4.1. Galectin-3 15
   4.2. Galectin-Related Protein 17

OBJECTIVES 21

MATERIAL AND METHODS 23

1. PROTEIN EXPRESSION AND PURIFICATION 23
   1.1. *E. coli* expression of C-GRP-C 23
      1.1.1. *DNA cloning* 23
      1.1.2. *DNA ligation* 24

1.1.3. *DNA transformation* ……. 25
1.1.4. *Heterologous protein over-expression* ….…… 26
1.2. Protein purification …..……..……... 26
GAL-3[N VII-IX] …..……... 27
C-GRP-C ……..……... 27

2. PROTEIN CRYSTALLIZATION …… 29
2.1. Nucleation and crystal growth ……..……..…….. 29
2.2. Protein solubility …….....…….…….……... 30
GAL-3[N VII-IX] ……..……... 30
C-GRP-C …...……..…… 31

3. PROTEIN CRYSTALLOGRAPHY ……..…… 32
3.1. Principles of crystalline symmetry ...…….…….. 32
3.2. Solvent content in crystals ….....…….…… 32
3.3. Synchrotron radiation …….....…….…… 32
3.4. Data collection …….....…….…… 33
3.5. Rational mathematical representation of X-ray diffraction ..……... 34
3.6. X-ray scattering by a crystal .....…….……... 35
3.7. The Patterson function and Friedel's law ……..…….. 36
3.8. Softwares for crystallographic resolution …… 37
3.9. Principles of the self-rotation function …… 38
GAL-3[N VII-IX] ……..…... 39
C-GRP-C ……..……... 40

4. PROTEIN BIOPHYSICAL CHARACTERIZATION ...…….. 41
4.1. Small-Angle X-ray Scattering ...……. 41
GAL-3[N VII-IX] ……..……... 41
C-GRP-C ……. 42
4.2. Analytical ultracentrifugation ……...……. 42
C-GRP-C …….…… 43

RESULTS 45

GAL-3[N VII-IX] 45

Protein purification 45

X-ray diffraction resolution 47

Small-Angle X-ray Scattering 59

C-GRP-C 62

E. coli expression and protein purification 62

X-ray diffraction resolution 64

Small-Angle X-ray Scattering 67

Analytical ultracentrifugation 69

DISCUSSION 71

CONCLUSIONS 87

CONCLUSIONES 87

REFERENCES 89

PUBLICATIONS xiii

Flores-Ibarra et al 2015

García, Flores-Ibarra, Michalak (first authors) et al 2016

APENDICES

List of abbreviations

Gal-3[N VII-IX] – Human Galectin-3 deletion mutant Δ23-75
C-GRP-C – Chicken Galectin-Related Protein deletion mutant Δ1-36

ASF – Asialofentuin
AUC – Analytical Ultracentrifugation
Bcl-2 – B-cell lymphoma 2
bFGF.- Basic Fibroblast Growth Factor
BME – β-Mercaptoethanol
CG – Chicken Galectin
CKI – Protein Kinase I (formerly known as Casein Kinase I)
CRD – Carbohydrate Recognition Domain
DTT – Dithiothreitol
ECM – Extracellular Matrix
EGP – ER/Golgi Pathway
ERK – Extracellular-Regulated Kinase
FRET – Fluorescence Resonance Energy Transfer
FT – Fourier Transform
GAG – Glycosaminoglycan
Gal – Galectin
Gal-3$_{FL}$ – Full-length Galectin-3
GBP – Glycan-Binding Protein
HSPC159 – Human Stem Progenitor Cell 159
IL-5 – Interleukin-5
IPTG – Isopropyl β-D-1-thioglycopyranoside
LMU – Ludwig-Maximilians University
MMP – Metalloproteinase
NCS – Non-Crystallographic Symmetry
N-LD – Gal-3 N-terminal Lead Domain
N-PG – Gal-3 poly-Proline/Glycine repeats of the N-terminal chimeric module
NWGR – Asn-Trp-Gly-Arg protein motif
PBS – Phosphate Buffer Saline
PI3K – Phosphoinositol-3-Kinase
[PSI] BLAST – [Position-Specific Iterated] Basic Local Alignment Search Tool
RAF1 – *Raf* proto-oncogene Serine/Threonine Protein Kinases
SAXS – Small-Angle X-ray Scattering
TNF – Tumor Necrosis Factor
TRAIL – TNF-Related Apoptosis Inducing Ligand
VEGFR – Vascular Endothelial Growth Factor Receptor

Glossary

Asymmetric unit.- The asymmetric unit of a space group is that part of the crystallographic unit cell which can be used to generate the complete unit cell by the symmetry of the space group.

Miller indices.- A set of numbers that quantify the intercepts and thus may be used to uniquely identify the plane or surface in which the crystal is oriented.

Biochemical lattice.- A cluster formed by ligands into lipid raft microdomains required for optimal transmission of signals relevant to cell function.

Crystal lattice.- Describes the periodicity of a crystal by mathematical points at specific coordinates in space. There is a simple inverse relationship between the spacing of unit cells in the crystallographic real lattice and the spacing of reflections on the detector, which because of its inverse relationship to the real lattice, is called **reciprocal**.

Epimerization.- The chemical conversion of one epimer to its stereoisomer that differs in the arrangement of groups on a single asymmetric carbon atom.

Glycosidic linkage.- Linkage of a monosaccharide to another residue via the anomeric hydroxyl group. The linkage generally results from the reaction of a hemiacetal with an alcohol to form an acetal. Glycosidic linkage between two monosachharides have defined regiochemistry and stereochemistry.

Glycosidases.- An enzyme that catalyzes the hydrolysis of a bond joining a sugar of a glycoside to an alcohol or another sugar unit.

Glycosyltransferases.- Enzymes involved in glycan biosynthesis and modification, including aspects of substrate specificity, primary sequence relationships, structures and enzyme mechanisms.

Great obstetrical syndromes.- refer to conditions with the following characteristics: (1) multiple etiologies; (2) a long preclinical period; (3) adaptive in nature; (4) fetal involvement and (5) the result of complex interactions between the maternal and fetal genome and the environment.

Hematopoiesis.- The formation of mature blood cellular components from stem cells and other undifferentiated blood cells.

Integrins.- Heterodimeric transmembrane receptors that, upon ligand binding, activate signal transduction pathways that mediate cellular signals such as adhesion, regulation of the cell cycle, organization of the intracellular cytoskeleton, and movement of new receptors to the cell membrane.

Jelly-roll folding.- This motif describes a particular topology for arranging antiparallel β-strands in protein structures.

Lectins.- Members of a superfamily of (glyco)proteins with the capacity to bind carbohydrates which lack enzymatic activity and are distinct from antibodies and oligosaccharide sensor/transport proteins.

Monochromator.- An optical device that comprises a dispersive element, an entrance slit and mirrors to create a parallel beam similar to sunlight, and an exit slit and mirrors to extract the monochromatic (single-wavelength) light.

Non-crystallographic symmetry.- A symmetry operation that is not compatible with the periodicity of a crystal pattern (in two or three dimensions). In biological crystallography, the term 'non-crystallographic symmetry' is often used to indicate a symmetry relationship between similar subunits within the crystallographic asymmetric unit. This use comes from the fact that the operation required to superimpose one subunit on another is similar to a space group operation, but it operates only over a local volume.

Synchrotron radiation.- The name given to the radiation which occurs when charged particles are accelerated in a curved path or orbit. This radiated energy is proportional to the fourth power of the particle speed and is inversely proportional to the square.

Unit cell.- Minimum volume cell corresponding to a single lattice point of a structure with discrete translational symmetry.

ABSTRACT

ABSTRACT

Galectins are a family of β-galactoside-binding proteins located in a wide range of organisms, from mammals to fungi and prokaryotes, where they exert functions that can be mediated by both carbohydrate-protein and protein-protein interactions. Galectins were classified based on their structural features in prototype, tandem-repeat and chimeric type. Prototype galectins contain one CRD per subunit and can form homodimers whereas tandem-repeat galectins are heterodimers of two non-identical CRDs linked by a peptide. The chimeric galectins have a distinctive N-terminal domain comprised of a lead peptide and nine poly-Pro/Gly repeats (namely, N-PG) linked to the C-terminal CRD. In humans, the only chimeric type is Galectin-3 (Gal-3). The N-PG module had eluded structural resolution so far, albeit the many reports on the CRD structure. Using a protein engineering approach, Gal-3 constructs with different lengths of the N-PG were produced and set to crystallize. A particular construct, with repeats VII to IX and the lead domain (Gal-3[N VII-IX]) crystallized with high-reproducibility. The crystals belonged to the orthorhombic space group $P2_12_12_1$ and had a resolution of 2.2 Å. Further studies were performed in this construct and the full-length protein by SAXS. Taken together, the results on the structural characterization of Gal-3[N VII-IX] allow to ascertain the position of key amino acids in the N-terminal that play fundamental roles in Gal-3 function, most importantly the position of an apoptosis- and metastasis-related phosphorylation site, **Ser6**. This new structure with the N-terminal section sets the path for the elucidation of full-length Gal-3 structure and its pharmacological applications.

To date, a total of 15 members have been identified in the galectin family plus a few related proteins that closely share their sequence. A comprehensive genetic mapping of this family identified one particular Galectin-Related Protein (formerly known as HSPC159, now GRP) located in bone marrow that has a high degree of identity with galectins sequences, sharing their *jelly-roll* characteristic topology. This GRP is present only in vertebrates, unlike its ubiquitous relatives. Interestingly, GRP does not bind β-galactosides. Human GRP structure has been reported twice with no further information on its function. In order to study in detail the entirety of the galectin family for identifying themes of divergence and recognition within a limited number of proteins, *Gallus gallus* GRP was chosen as model. Crystals of this protein belonged to the body-centered orthorhombic space group $I2_12_12_1$ and had a resolution of 1.5 Å. Their crystallographic resolution showed the particular hallmarks that hinder β-galactoside binding. This structure and its properties were further studied by SAXS and AUC.

RESUMEN

La familia de las galectinas comprende proteínas de unión a β-galactósidos presentes en una amplia variedad de organismos, de mamíferos a hongos y procariotes, donde cumplen funciones que pueden ser activadas por interacciones con carbohidratos o con otras proteínas. Las galectinas fueron clasificadas, con base en sus características estructurales, en prototípicas, tándem y quiméricas. Las prototípicas pueden formar homodímeros y las tándem forman heterodímeros del dominio de reconocimiento de carbohidratos (CRD), en tanto que las galectinas quiméricas tienen un domino N-terminal distintivo, que comprende un péptido líder y nueve repeticiones Pro/Gly (N-PG), unido al CRD en el C-terminal. En humanos, la única galectina quimérica es Galectina-3 (Gal-3). La resolución estructural del módulo N-PG no había sido conseguida, a pesar de la existencia de numerosos reportes del CRD de Gal-3. Haciendo uso de la ingeniería proteica, se produjeron construcciones de Gal-3 con distintas longitudes del N-PG para ensayos de cristalización. Una construcción que contiene el péptido líder y las repeticiones VII-IX del N-PG (Gal-3[N VII-IX]) cristalizó con alta reproducibilidad. Los cristales pertenecen al grupo espacial ortorrómbico $P2_12_12_1$ y tienen una resolución de 2.2 Å. Estudios posteriores fueron realizados por dispersión de ángulo pequeño (SAXS). Los resultados de los experimentos para la caracterización estructural de Gal-3[N VII-IX] permiten localizar la posición de aminoácidos que juegan un papel fundamental en la función de Gal-3, en especial la posición de un sitio de fosforilación que media procesos de apoptosis y metástasis, **Ser6**. Esta nueva estructura con el [N VII-IX] es el primer paso en la elucidación de la estructura completa de Gal-3 y de sus aplicaciones farmacológicas.

En total, 15 miembros han sido identificados en la familia de las galectinas, junto con otras proteínas relacionadas. Un estudio genético completo identificó en particular una proteína de la médula ósea asociada a galectinas (GRP, antes conocida como HSPC159), con un alto grado de identidad en su secuencia y que comparte el mismo plegamiento *jelly-roll*. GRP está presente únicamente en vertebrados, a diferencia de las galectinas. Cabe resaltar que GRP no une β-galactósidos. La estructura de GRP humana ha sido reportada dos veces sin profundizar en sus funciones. Para estudiar en detalle a la familia entera de galectinas e identificar puntos de divergencia y reconocimiento en un número limitado de proteínas, se escogió como modelo a la GRP de *G. gallus*. Los cristales de esta proteína pertenecen al grupo ortorrómbico $I2_12_12_1$ y tienen una resolución de 1.5 Å. La estructura de GRP mostró las características específicas que impiden la unión de β-galactósidos. Esta estructura y sus propiedades fueron confirmadas por medio de experimentos de SAXS y AUC.

INTRODUCTION

INTRODUCTION

Defined in a broad sense, glycobiology is the study of carbohydrates, including the chemistry and structure of saccharides, the enzymology of glycan formation and degradation, the recognition of glycans by specialized proteins, the formation of glycoconjugates, the roles of glycans in complex biological systems, and their analysis and manipulation. Taken together with the fact that they encompass some of the major post-translational modifications of proteins themselves, glycoconjugate-forming carbohydrates help to explain how the relatively small number of genes in the typical genome can generate the enormous biological complexities inherent in the development, growth and functioning of intact organisms. Microscopic studies of cells revealed that these glycoconjugates, with extensive glycosylation both in frequency and chain length, coat the cell surface in a complex network termed glycocalyx (Varki et al 2009; Lichtenstein and Rabinovich 2013). Certain changes in the cell expression of glycans and glycoconjugates are often found in the course of transformation and progression to malignancy, metastasis-associated processes including angiogenesis and immune escape, as well as in other pathological situations such as inflammation and autoimmunity, the precise mechanisms of which are understood in only a few cases (Rabinovich and Thijssen 2014; Cagnoni et al 2016).

With the growing awareness that the significance of glycosyl residues is to impart a discrete recognition role to receptors, glycans are now accredited to have information-coding ability previously assigned exclusively to proteins. Actually, in terms of coding capacity, the theoretical number of all possible oligosaccharides isomers built from monosaccharides is several orders of magnitude larger (1.44×10^{15}) than *peptides* (6.4×10^{6}) and *oligonucleotides* (4096). Therefore, for pharmacological and biotechnological reasons, the study of both carbohydrates and their recognition proteins is becoming increasingly relevant (Gabius et al 2011).

1. CELL GLYCOSYLATION

Glycosylation of the cell occurs by ensemble of monosaccharide building blocks with each other and with other molecules. Monosaccharides are imported into the cell, salvaged from degraded glycans, or derived from other sugars within the cell. Although most glycosylation reactions occur in the Golgi apparatus, precursor activation and interconversion occur mostly in the cytoplasm. Regardless of their localization, most glycosylation reactions use activated forms of monosaccharides as donor for reactions that are catalyzed by enzymes such as *glycosyltransferases* and *glycosidades*. These enzymes link monosaccharides moieties into linear and branched

glycan chains. They catalyze a group-transfer reaction, with either inversion or retention of stereochemistry at the anomeric carbon atom of the donor substrate, in which the monosaccharide moiety of a single activated sugar donor substrate is transferred to the acceptor substrate. Generally, strict donor-acceptor and linkage specificity are exhibited by most glycosyltransferases, a property that serves to define and limit the number and type of glycan structures observed in a given organism. Substrates include other saccharides, lipids, small organic molecules, DNA and proteins. The main carbohydrate constituents of cellular glycans differ only in the relative positioning of one or two hydroxyl groups, thus *epimerization* must have substantial consequences for protein recognition of the sugars. The two general classes of protein-bound glycans in glycoconjugates are *N*-linked (to the **N** atom of **Asn** side chain), and *O*-linked (to the **O** atom of **Ser**/**Thr** side chains) (Varki et al 2009).

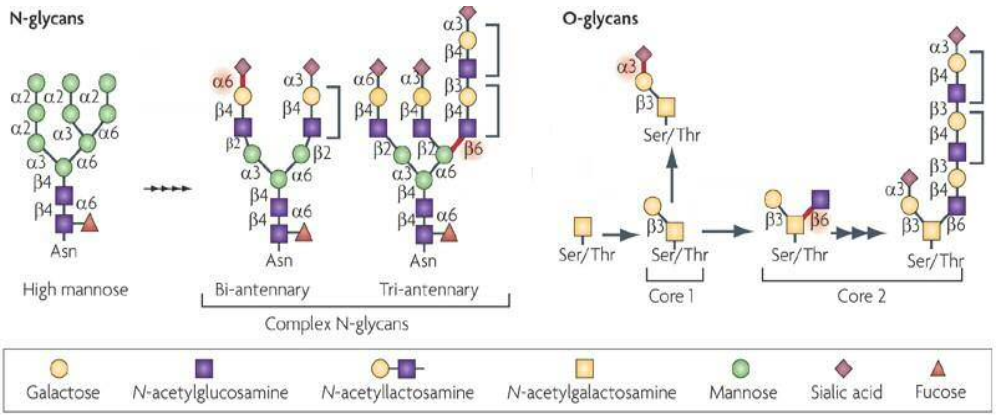


Fig 1. The two general classes of protein-bound glycans in glycoconjugates are *N*-linked and *O*-linked. Multiple LacNAcs may be presented on the branches of *N*-glycans or occur as polyLacNAc chains on either *N*- or *O*-linked glycans. Generation of polyLacNAc sequences is regulated in part by the family of core 2 β-1,6-*N*-acetylglucosaminyltransferase branching enzymes for *O*-glycans and β-1,6-*N*-acetylglucosaminyltransferase V branching enzyme for *N*-glycans (Hernández and Baum 2002; Rabinovich and Toscano 2009).

Certainly nature appears to have taken full advantage of the vast diversity of glycans expressed in organisms by evolving protein modules to recognize discrete glycans that mediate specific physiological or pathological processes. Conjugation of sugars to proteins occurs throughout the entire phylogenetic spectrum, with a known total of 13 monosaccharides and 8 amino acids forming at least 41 types of glycosidic linkages (**Fig 1**). Interestingly, a group of 250-500 genes is devoted to the synthesis and remodeling of glycan chains. Many glycans and glycoconjugates are recognized with higher specificity via Glycan-Binding Proteins (**GBP**s): in fact, there are no

living organisms in which GBPs have not been found (Varki et al 2009; Gabius et al 2011; Voet et al 2013).

2. GLYCAN-BINDING PROTEINS

Glycan-binding proteins are often glycoproteins themselves, synthesized in the ER/Golgi pathway (**EGP**). However, a significant subset of soluble GBPs, such as galectins, heparin-binding growth factors and some cytokines, are synthesized on free ribosomes in the cytoplasm and then delivered directly to the exterior of the cell. This makes functional sense, since several of these lectins can recognize biosynthetic intermediates that occur in the EGP, *e.g.* galactosides and high-mannose oligosaccharides. Once secreted, glycan-binding sites of GBPs tend to be of a relatively low affinity though they can exhibit high specificity. The ability of such low affinity sites to mediate biologically relevant interactions in the intact system thus usually requires multivalency (Rabinovich et al 2002a).

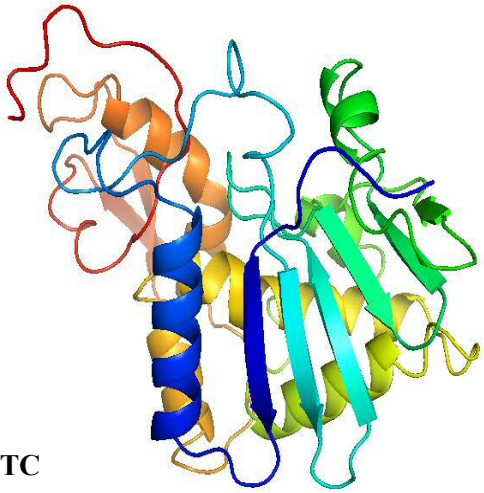
2.1. Types of lectins

Lectins, from the Latin "leggere" (to select), were traditionally defined as "hemagglutinating, multivalent carbohydrate-binding proteins that are not antibodies". Since many lectins function as signaling molecules, their multivalency may promote cross-linking of relevant cell-surface receptors and be required for signaling.

A central question in lectin biology is how the exquisite target specificity of these proteins for certain cellular glycans can be explained. For instance, some lectins have a circumscribed set of immune function, which tether circulating leukocytes to inflamed endothelium to initiate the process of diapedesis. Transmembrane calcium-dependent lectins recognize various types of danger signals from both microbial pathogens and damaged or altered host cells. Soluble lectins like ficolins, collectins, and mannose-binding protein opsonize microbial pathogens. Siglecs regulate leukocyte activation. The first evidence on animal lectins (agglutinin activity) appeared in the studies of Ginsburg in the 1960s, in which blood leukocytes from rats were treated with bacterial glycosidases and then injected back into the circulation (Varki et al 2009; Gabius et al 2011; Thiemann and Baum 2016).

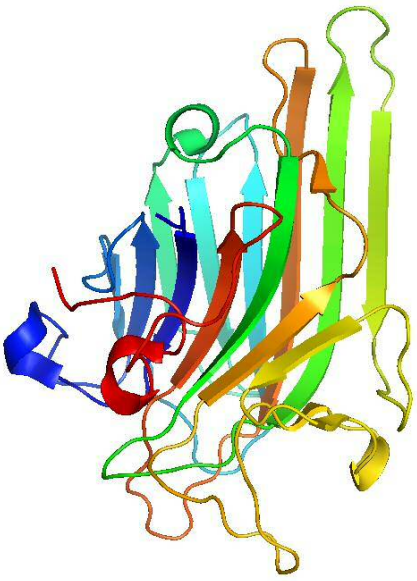
At the present time, there is no single universally accepted classification of lectins, though it is based in sequence homologies and evolutionary relationships. Accordingly, the following classification is widely accepted:

2.1.1. R-type lectins. – Members of this subfamily contain a structurally similar CRD to the one in ricin, the first lectin discovered. R-type lectins are present in bacteria, plants and animals (Varki et al 2009; Bassik et al 2013).



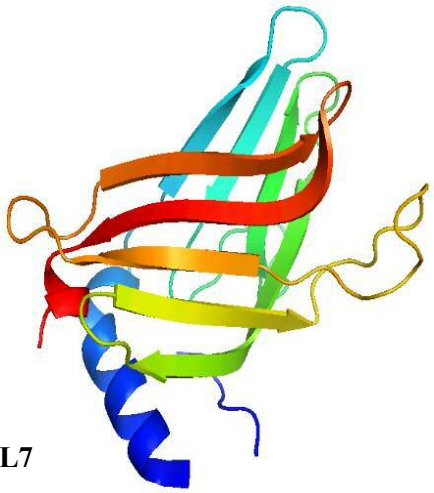
R-type
PDBID: 1RTC

2.1.2. L-type lectins. – This subfamily was first discovered in the seed of leguminous plants. It is now known that they have structural motifs present in other eukaryotic organisms' GBPs (Varki et al 2009; Cooper 2002).



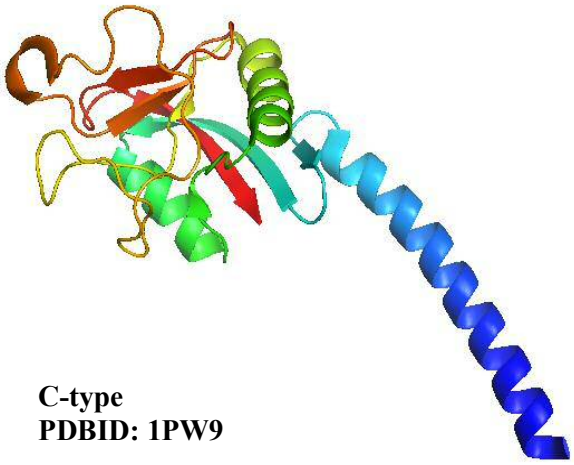
L-type
PDBID: 3NWK

2.1.3. P-type lectins. – These proteins specifically recognize mannose-6-phosphate. After binding to this sugar, they regulate lysosomes trafficking in the cells (Varki et al 2009).



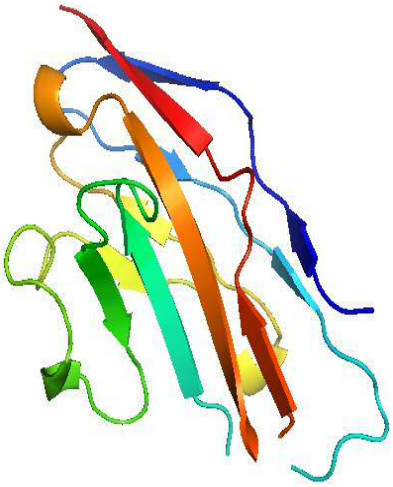
P-type
PDBID: 2RL7

2.1.4. C-type lectins. – These lectins are Ca^{2+}-dependent glycan-binding proteins that share primary and secondary structural homology in their CRDs, and are found in all organisms. This large family includes collectins, selectins, endocytic receptors and proteoglycans. Some of these proteins are secreted and others are transmembranal. Their capability for oligomerization increases their avidity for multivalent ligands. The first reported lectin in animal cells was of this type, the hepatic glycoasialoprotein receptor (Varki et al 2009).



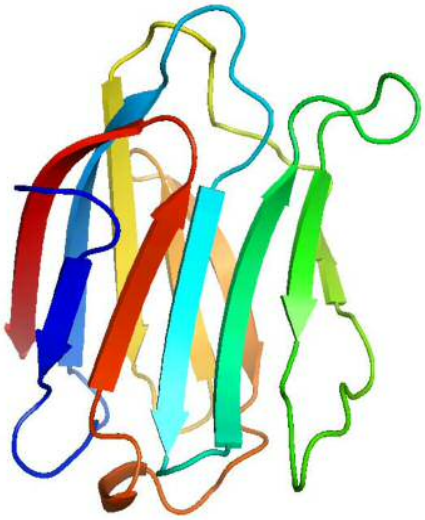
C-type
PDBID: 1PW9

2.1.5. I-type lectins. – These are proteins that belong to the immunoglobulin superfamily (>500 protein-predicted genomes), excluding antibodies and T-cell receptors. The Siglec family of sialic acid-binding lectins is the only well-characterized group of I-type lectins, both structurally and functionally (Varki et al 2009; Cagnoni et al 2016).



I-type
PDBID: 1QFO

2.1.6. Galectins. – Galectins are β-galactoside-binding proteins that share a highly conserved primary structural homology in their CRDs. Composed by 15 members in mammals, this family adds by sequence similarity other galectin-like proteins (Cooper 2002; Leffler et al 2004; Varki et al 2009).



Galectins
PDBID: 1A3K

3. THE GALECTIN FAMILY

The galectin family is evolutionarily ancient with representatives in vertebrates, invertebrates, fungi and even protists. The presence of galectins in so many species, including very ancient ones, implies that they evolved to play fundamental roles in cell biology, while the presence of multiple galectins within single species suggests that they have since diverged to participate in a variety of more specific functions (Cooper 2002).

During studies on the possible presence of galectins in the electric organs of the electric eel in 1975, a protein that required the inclusion of β-mercaptoethanol in isolation buffers to maintain its activity was found, suggesting the presence of one or more free **Cys** residues, thus galectins were originally referred to as S-type lectins. Later on was discovered that the transitions from free sulfhydryl groups to disulfide bridges appear to trigger shape changes. These redox-dependent shape changes, having potential as switches to intersubunit or intrasubunit disulfide bridges upon oxidation, suggest that the presence of **Cys** residues has a functional dimension. However, electrolectin, unlike most galectins, does not contain **Cys** residues but its key **Trp** residue in the carbohydrate-binding site can be oxidized, causing loss of activity (Teichberg et al 1975; Varki et al 2009).

In the early 1980s, a 35 kDa GBP (CBP35, now known as galectin-3), with capacity for β-galactoside binding was identified from mouse fibroblasts. All of the galectins discovered until then demonstrated hemagglutinin activity, but the choice of erythrocytes was crucial. Tripsinized rabbit erythrocytes, which display more terminal galactose residues than human erythrocytes, are readily agglutinated by most galectins, whereas human's require treatment with neuraminidase to enhance their agglutinability. The nomenclature for galectins was systematized in 1994. The first galectin found (electrolectin) was renamed galectin-1 (Gal-1), its nearest homologue was termed Gal-2, and CBP35 (also named εBP, L-29, and L-31) was termed Gal-3. The subsequent galectins discovered were numbered consecutively by order of discovery (Cowles et al 1990; Cooper 2002; Varki et al 2009).

Proceeding from the identification of the *jelly-roll* folding, with capacity to accommodate glycans and a common sequence signature, the systematic search for homologous proteins in a species and in phylogenesis stood as the next step on the way toward network analysis of galectins. Results of data base mining then set the stage to characterize expression and localization profiles as well as functional cooperation of all members of the corresponding family of lectins, which have arisen from an

ancestral gene by divergence through duplications/losses and sequence deviations. Indeed, using **BLAST** and **PSI-BLAST** search algorithms to screen genomic and mRNA databases for sequences similar to known galectins, Altschul et al (1997) identified many potential vertebrate and invertebrate candidates for membership in the galectin family, as well as new human galectin-related proteins. Certain other galectin-like genes appeared to be pseudogenes, for instance because stop codons interrupt their CRDs. A total of 15 galectins have now been found in mammals, but only 12 galectin genes are found in humans, including two for galectin-9 (Altschul et al 1997; Gabius 2000; Cooper 2002; Thijssen et al 2013).

Although much is known about galectins, many of their mechanisms of action remain unknown. Galectins can function intracellularly and can also be secreted to bind to cell surface glycoconjugate counterreceptors, though the secretion pathway of galectins is unidentified as of today. Some galectins are made by immune cells, whereas others are secreted by cells such as endothelial or epithelial cells. Galectin-binding to a single glycan ligand is a low-affinity interaction, but the multivalency of galectins and the glycan ligands present on cell surface glycoproteins result in high-avidity binding that can reversibly scaffold or cluster these glycoproteins. The binding of galectins to other glycoproteins is mostly regulated by several enzymes that mediate carbohydrate ligands expression by a cell. Furtherly, the effect of binding avidity in galectins is a result of ligand clustering or retention by these glycoprotein counterreceptors (Gabius et al 2011; Rabinovich and Thijssen 2014; Thiemann and Baum 2016).

Lastly, the detection of a Galectin-Related Protein (GRP) has its origin in gene expression cataloguing of human $CD34^{+}$ hematopoietic stem/progenitor cells that led, among 300 cDNA clones, to an mRNA termed **HSPC159** (Cooper 2002). Systematic alignments of its predicted amino acid sequence disclosed that it encompasses 51 positions of the 64 amino acids set most shared among galectins. Presence of the gene has been found to be exclusive of vertebrates, and initial comparison revealed an exceptionally high degree of similarity that implies a very strong positive selection. As predicted by sequence homology, GRP preserves the *jelly-roll* folding but it does not bind galectins' canonical ligands. GRP's presence in bone marrow as executer of HSPCs transportation to the blood stream points to the direction of its distinct non-galectin function, although its natural ligands are unknown, even when the human GRP structure has been elucidated (Wälti et al 2008a; Ruiz et al 2014).

3.1. Galectins classification and function

It is remarkable that β-galactosides at the branch ends of glycan antennae of cell surface glycoconjugates are not only accessible for molecular recognition as sensors, but they are also subject to an array of enzymatic substitutions, such as α2,3-sialylation. These structural alterations are very specific, as they modulate the ligand properties of the respective glycans to galectins. Through these interactions, galectins mediate multiple cellular responses, which harbor their ideal properties as high-density storage of biological information (Gabius 2009; Vasta et al 2012).

Most galectins are non-glycosylated soluble proteins, but a few exceptions have transmembrane domains. Although galectins lack a typical secretion signal peptide, they are present not only in the cytosol and the nucleus (where they are engaged in processes such as pre-mRNA splicing and protein regulation), but also in the extracellular space. From the cytosol, galectins may be targeted for secretion by non-classical mechanisms, possibly by direct translocation across the plasma membrane. In the extracellular space, galectins can bind to glycans at the cell surface and/or the **ECM**, and to potential pathogens. Also, galectin-mediated lipid raft assembly may modulate turnover of endocytic receptors and signal transduction pathways (Yang et al 2008; Vasta et al 2012).

The expression of galectins is modulated during the differentiation of individual cells and during the development of organisms and tissues, and is changed under different physiological or pathological conditions. Importantly, in most cases, protein-protein interactions, rather than lectin-carbohydrate interactions, are involved (Liu et al 2002).

Based on their domain organization, galectins are classified in three types (**Fig 2**):

(i) *Prototype* galectins contain one CRD per subunit and can form non-covalently linked homodimers.

(ii) *Tandem-repeat* galectins have two non-identical CRDs joined by a functional linker peptide ranging from 5 to 50+ amino acids in length.

(iii) *Chimera* galectins have a C-terminal CRD and a distinct N-terminal domain rich in proline and glycine (**N-PG**) that contributes to self-aggregation.

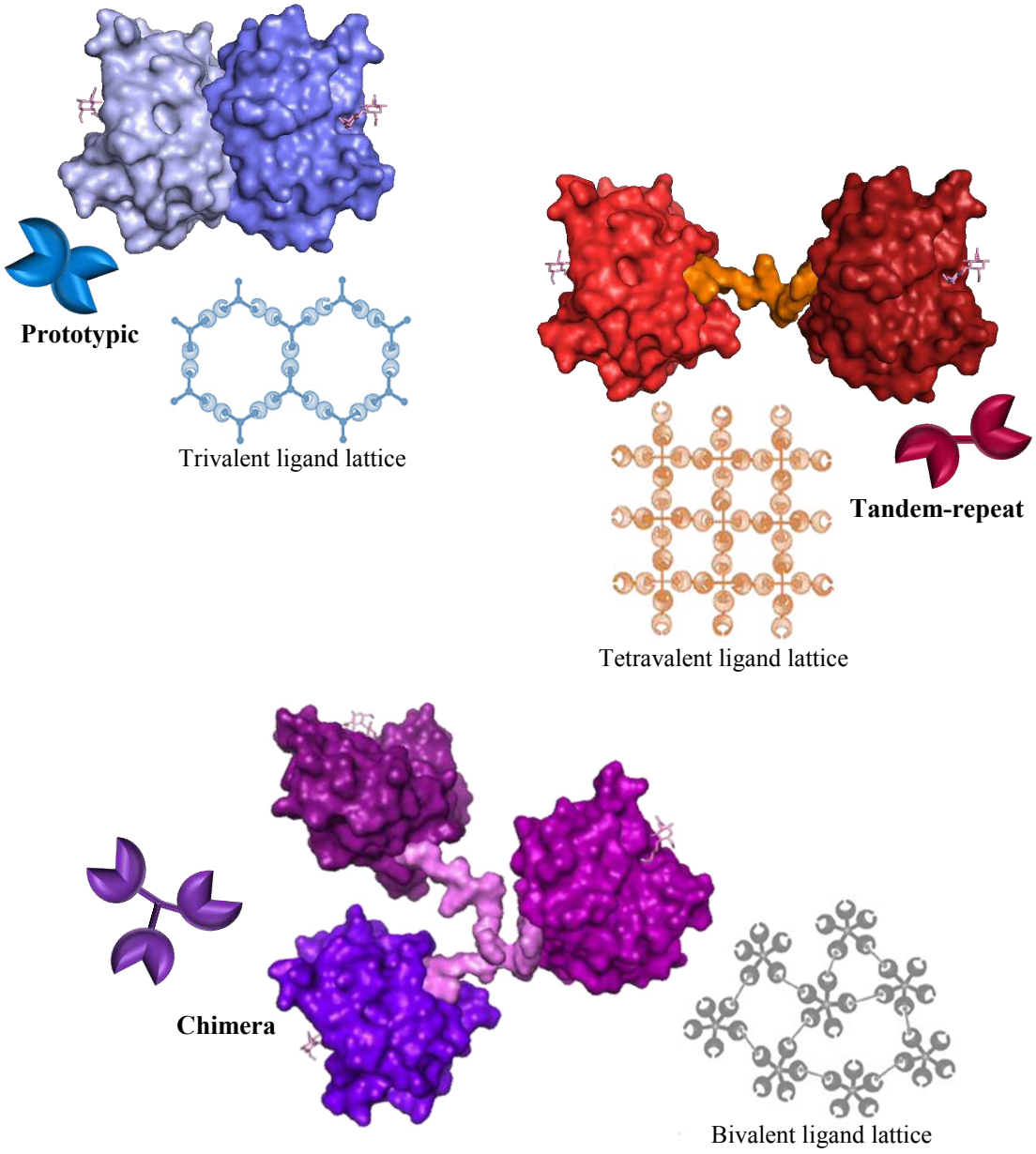


Fig 2. Galectin types and lattices. Proto-type galectins (**blue**) are most commonly non-covalently linked homodimers, tandem-repeat type (**red**) are usually heterodimers joined by a functional linker, and chimera-type (**purple**) have a C-terminal CRD and a chimeric N-terminal for multimerization. Galectins bind to cell surface β-galactoside-containing glycolipids and glycoproteins leading to the formation of lattices that cluster these ligands into lipid raft microdomains required for optimal transmission of signals relevant to cell function. Distinct types of lattice that can be formed hypothetically between multivalent galectins and multivalent glycans, such as trivalent (prototypical), tetravalent (tandem-repeat) and bivalent (Gal-3), are illustrated (Rabinovich and Toscano 2009).

Proto and tandem-repeat types comprise several separate galectin subtypes. Galectin(Gal)-1, -2, -5, -7, -10, -11, -13, -14 and -15 are prototypical, Gal-4, -6, -8, -9 and -12 are tandem-repeat, and Gal-3 is the only chimera type in vertebrates (Gabius et al 2004; Solís et al 2010). The association of prototype galectin monomers as non-covalently bound dimers via a hydrophobic interphase is critical for their function in mediating interactions, *lattice* formation localized at the cell surface, and downstream effector functions. Tandem-repeat galectins can recognize different saccharide ligands with a single polypeptide, although they can also form higher order aggregates that enhance their avidity. For chimera galectins, oligomerization takes place mainly in the presence of multivalent oligosaccharides in solution or at the cell surface, display binding cooperativity (**Fig 2**) (Rabinovich et al 2002a; Vasta et al 2012).

Susceptibility to galectins is controlled by the cell and may be regulated at three levels: (1) synthesis and modification of glycan ligands by glycosyltransferases and glycosidases, (2) presentation of glycan ligands by specific glycoprotein counterreceptors, and (3) intracellular signaling pathways initiated by galectins binding to glycoprotein counterreceptors. However, the translational and clinical applications of altered glycan structures have not yet been completely accomplished, probably due to the complex regulation of the glycosylation process: heterogeneity is inherent to the language of glycans and crucial for their diverse biological roles as information carriers for lectins (Hernández and Baum 2002; Cagnoni et al 2016).

While Gal-1 and -3 are detected ubiquitously in an organism, other galectins are more specifically located, such as Gal-2 and -4, which are preferentially found in the gastrointestinal tract, Gal-7 is highly abundant in skin tissue, Gal-10 in eosinophils, and Gal-12 in adipose tissue (Cagnoni et al 2016). That is, individual galectins can act on multiple cell types and multiple galectins can act on the same cell (**Fig 3**).

Cancer regulation: Different galectins have discrete yet complementary roles in the regulation of tumor progression. There is direct evidence that Gal-1 and -3 expression is necessary for the initiation of the transformed phenotype of tumors. Oncogenic Ras requires Gal-1 mediated anchorage to the plasma membrane, and Gal-1 expression also results in the sustained activation of **RAF1** and **ERK**. Gal-3 promotes the activation of RAF1 and **PI3K**, and contributes to the selective activation of signaling cascades and the regulation of gene expression at the transcriptional level (Liu and Rabinovich 2005; Gao et al 2014).

Apoptosis regulation: To date, only two families of proteins have been described as death-inducing ligands: the **TNF** family and the galectin family. Pro-apoptotic galectins bind to specific saccharide ligands on cell surface glycoproteins and/or glycolipids to initiate cell death (Hernández and Baum 2002). In Gal-3, the **NWGR** sequence motif, found in **Bcl-2** and other apoptosis-regulating proteins, has been implicated in the intracellular Gal-3 anti-apoptotic effects on tumorous cells, therefore contributing to the survival of metastasizing tumor cells. On the other hand, Gal-1 inhibits full T cell activation, induces growth arrest and apoptosis of activated T cells, and suppresses the secretion of pro-inflammatory cytokines (Umemoto et al 2003; Liu and Rabinovich 2005; Lichtenstein and Rabinovich, 2013).

Metastasis: Gal-3 acts as a mediator in both **VEGFR** and hematogenous metastasis, **bFGF**-mediated angiogenic response and cell adhesion to the endothelium. Galectins recognize, in a carbohydrate-dependent manner, one of the most extensively studied families of cell adhesion molecules: the integrins. Gal-3 upregulates integrin expression. The level of Gal-1 expression is positively correlated with the migratory phenotype and biological aggressiveness of human astrocyte tumors. Gal-8 ligates integrins and triggers integrin-mediated signaling cascades, resulting in cytoskeletal changes and cell spreading (Glinsky et al 2003; Markowska et al 2011; Liu and Rabinovich 2005).

Inflammation and infection: Some galectins are amplifiers of the inflammatory response, whereas others activate homeostatic signals to shut off immune effector functions. Gal-1 recognizes a number of cell-surface glycoproteins in T cell lines in a carbohydrate-dependent manner, and can cause a redistribution of some of these proteins into segregated microdomains within the plasma membrane. Gal-3 induces the activation of various inflammatory cell types and can function like a chemokine, attracting monocytes and macrophages. Gal-3 might also reduce the immune response under certain circumstances by downregulation of **IL-5** production and blocking the differentiation of B lymphocytes into plasma cells, with the consequence of reducing the number of circulating antibodies. Gal-2 can induce T cell apoptosis and control the secretion of lymphotoxin-α by macrophages. Gal-4 activates $CD4^+$ T cells. Gal-9 induces T cell apoptosis and function as an eosinophile chemoattractant. In addition, besides mediating endogenous functions, and innate as well as adaptive immune processes, galectins also bind exogenous glycans on the surface of pathogenic microorganisms (Frigeri et al 1994; Liu et al 1995; Rabinovich et al 2002b; Nieminen et al 2005; Vasta et al 2012).

Other functions: Finally, their expression has been found in non-infectious, non-carcinogenic diseases such as arthritis, chronic inflammation, heart failure and diabetes. Galectins are highly expressed at the maternal-fetal interface, and their dysregulated expression is observed in the *great obstetrical syndromes* (Than et al 2012; Lichtenstein and Rabinovich 2013; Thijssen et al 2013; Gao et al 2014).

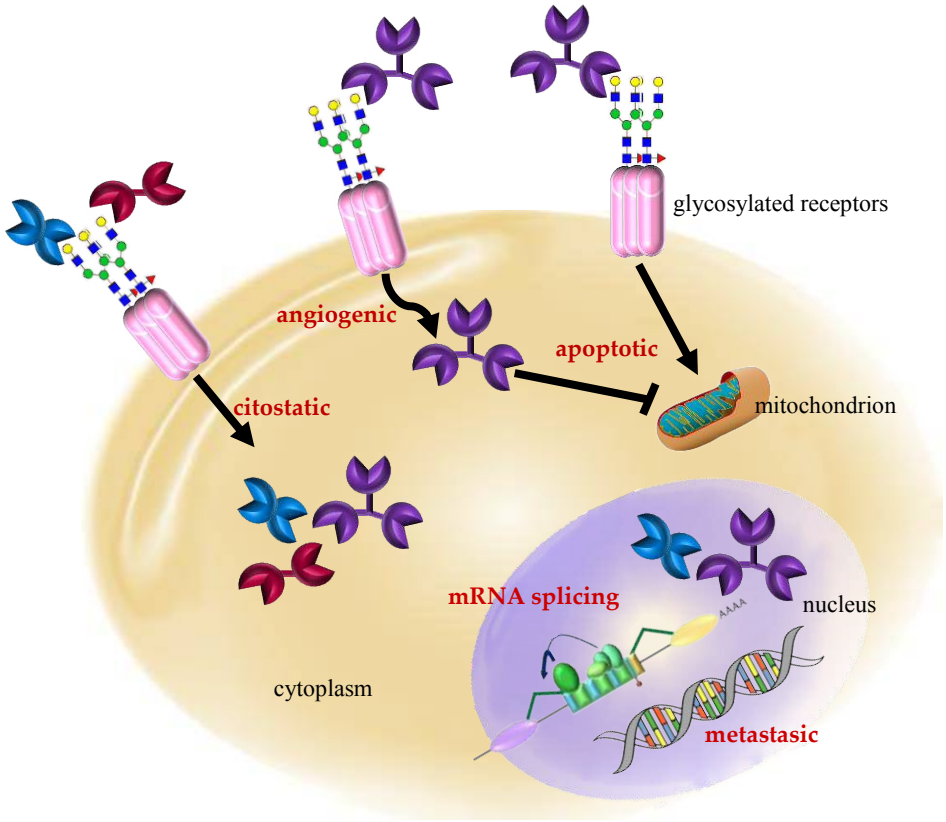


Fig 3. Galectins' biological functions. Both by self-association or association with other galectins, these proteins are able to selectively promote or inhibit cell death, as well as being mRNA splicing factors and cell-cycle regulators (Liu and Rabinovich 2005).

4. Structural characterization of galectins and GRP

The CRD of most galectins is composed of a five-stranded (F1–F5) and a six-stranded (S1–S6) anti-parallel β-sheets connected by a 3_{10} helix, forming a β-sandwich arrangement, a folding called *jelly-roll*. In the canonical dimeric Gal-1 and -2, β-strands F1 and S1 from each monomer extend the antiparallel β-strand interactions across the two-fold symmetry at the dimer interface, whereas the S1 and F1 β-strands of galectin-3 form a solvent-exposed surface. The carbohydrate binding site is formed by three continuous concave strands (4–6) containing all key residues for binding, such as **His45**, **Asn47**, **Arg49**, **Arg70**, **Glu68** and **Trp65** (numbering is for Gal-2), that are involved in direct interactions with β-galactosides (**Fig 4**). Galectin-Related Protein and even some viral galectins share this topology (Seetharaman et al 1998; Wälti et al 2008a; Vasta et al 2012; Thiemann and Baum 2016).

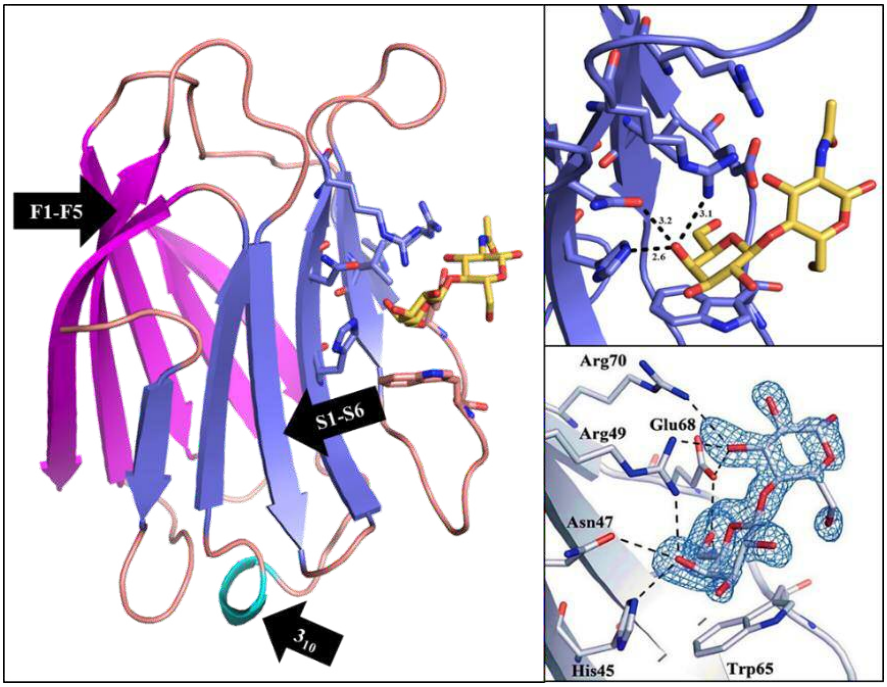


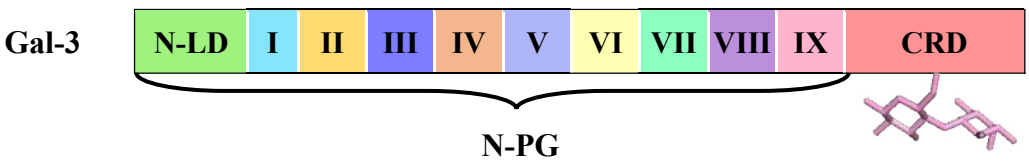
Fig 4. "Jelly-roll" folding and Carbohydrate Recognition Domain. Galectins' CRD is tightly folded in a β-sandwich folding called *jelly roll*, with the CRD site between strands S4-S6 in Gal-3 (Seetharaman et al 1998).

It is generally accepted that galectins require three **OH** groups of **Galβ1-3/4(Glc/Gal)NAc** units, the **4-OH** and **6-OH** groups of galactose, and the **3-OH** group of **(Glc/Gal)NAc**. Some galectins also show a particular preference for **α1-2Fuc-, α1-3Gal-, α1-3GalNAc-**, or **α2-3NeuAc**-modified glycans (Hirabayashi et

al 2002). For instance, Gal-3 co-crystallization with **Galβ1-3/4GlcNAc** disaccharides revealed that the **Asn180**, **Trp181** and **Gly182** amino acids are important for the recognition of galactose moieties (Sorme et al 2005). The amino acids interacting with the docked oligosaccharide in the galectins active sites are important, but so are the surface-exposed amino acids that orient glycans spatially and participate in their movement into the active site (Nahalka 2012).

4.1. Galectin-3

Galectin-3 (Gal-3), a 29-kDa protein, is the only chimera type galectin, with a CRD domain joined to an N-terminal domain with several repeats of a peptide sequence rich in proline and glycine (N-PG) that is similar to collagen repeating domains (Gabius et al 2011; Vasta et al 2012). The C-terminal half, composed of approximately 130 amino acids, that forms a globular structure, accommodates the entire carbohydrate-binding site and is thus responsible for the lectin activity of Gal-3. The N-PG contains 110-130 amino acids and is highly conserved among different species (Dumic et al 2006; Elola et al 2007). Different parts of the N-PG have been subclassified into a short N-terminal leader domain (**N-LD**), which corresponds to the first 12 amino acids, and the collagen-like repeats, which contain I–IX repeats of short amino acid segments **Pro–Gly–Ala–Tyr–Pro–Gly** (Gong et al 1999; Cooper 2002; Nangia-Makker et al 2007).



Self-association of Gal-3 has been reported to be mostly dependent on the N-PG domain (Rabinovich et al 2007; Rapoport et al 2008). Since Gal-3 displays di- and oligomeric conformations that enable ligand cross-linking or high-affinity binding of clustered ligands to multiple CRDs, the involvement of the N-PG tail in these properties have been studied in a number of chemical and physiological scenarios (Liu and Rabinovich 2005; Kaltner et al 2011).

It seems that the N-PG region accounts for Gal-3 oligomerization via tail-to-tail interaction, for instance, in the experiments of Kaltner et al (2011), since the proteolysis of the N-terminal repetitive domain by metalloproteinase-2 and -9 (**MMP-2**, -

9) leaves an intact C-terminal galectin domain with full lectin activity but losing much of the Gal-3 propensity to multimerize (Kaltner et al 2011).

A methodology based on the measurement of residual dipolar couplings from NMR spectra was used to characterize differences in protein structure along the backbone in the presence and absence of ligand, as well as the binding geometry of the ligand itself. The data on β-galactoside ligands are consistent with the ligand binding geometry found in Gal-3 crystal structures of the complexed state. However, a significant rearrangement of backbone loops near the binding site appears to occur in the absence of ligand (Umemoto et al 2003).

Gal-3 may be transported to the early/recycling endosomes and then partitioned into two routes – recycling back to the plasma membrane or targeting to the late endosomes/lysosomes. The carbohydrate-recognition domain is required for binding and endocytosis since the N-PG alone cannot mediate these processes. Although it was largely irrelevant for the protein trafficking to the early endosomes, presence of the N-PG domain was required for targeting Gal-3 to the late endosomes/lysosomes (Gao et al 2012). Moreover, the CRD alone fails to mediate neutrophile adhesion, and fails to induce migration and capillary tubule formation in endothelial cells *in vitro* and angiogenesis *in vivo*, further underscoring the importance of the chimeric extension N-PG (Sato et al 2002; Nangia-Makker et al 2007).

This chimeric region plays an as yet ambiguous role in Gal-3 aggregation and ligand recognition. Results of mutation studies revealed, for instance, that the N-PG has biological importance in binding; intact Gal-3 has a 3.8-fold higher affinity for carbohydrates than the CRD alone (Funasaka et al 2014b). The affinity difference suggests that the N-terminal non-CRD contributes to Gal-3's enhanced affinity for the extended structures of basic recognition units and the ability to recognize other carbohydrates and/or proteins. Of note, *cis*-binding to cell surface counterreceptors and *trans*-binding for cell bridging depend on distinct N-PG properties. However, in biological experiments, Gal-3 with MMP truncation did not impair cell proliferation and hemagglutination, unless left the CRD alone, and sometimes the cleavage even induced angiogenesis in cancer models (Nangia-Makker et al 2010). The application of engineered Gal-3 N-PG variants of different lengths offers a promising perspective to delineate detailed structure-activity profiles (Ochieng et al 1994; Hirabayashi et al 2002; Flores-Ibarra et al 2015).

The first structural report on Gal-3 was given by the CRD-only crystallographic resolution (Seetharaman et al 1998). Although associations involving both the N- and C-terminal domains could easily be achieved by a parallel orientation of monomers, an arrangement seen in the collectins, analysis of the CRD alone from Gal-3 revealed an apolar patch in the face of the 5-stranded β-sheet which may provide a site for monomer-monomer interactions (Drickamer 1993; Seetharaman et al 1998). Thirty Gal-3 CRD-only structures have been solved by X-ray crystallography ever since, but with all this data to this day, nineteen years after the first one reported, there is no solved structure of the protein with the N-PG domain (http://www.rcsb.org/pdb). NMR studies have shown that this domain is highly flexible in solution, which may account for the difficulty to crystallize it (Seetharaman et al 1998; Birdsall et al 2001).

4.2. Galectin-Related Protein

Shared *exon-intron* organization suggests that vertebrate galectins originated from an ancestral mono-CRD galectin by gene duplication, divergence and subfunctionalization. Systematic alignments of the predicted sequence of a Galectin-Related Protein (**GRP**) in human HSPCs disclosed its similarity to galectins. Comprehensive listing of the currently mapped species with positivity for the GRP gene showed that, in all cases, the copy number for the protein is one (**Fig 5**) (Cooper et al 2002).

The region of the CRD following the N-terminal tail in GRP is highly conserved, as it is the second half of the 36-amino acid N-terminal section. Still, the most notable feature of GRP is its lack of β-galactosides binding activity, in spite of its sequence correspondence (51 out of 64 most common amino acids) with canonical galectins. Because GRP is a hematopoietic precursor, it resides in specific niches that control survival, proliferation, self-renewal and differentiation of lymphocytes. In mammals, the continuous trafficking of GRP between the bone marrow and blood compartments contributes to the maintenance of normal *hematopoiesis*. This trafficking is induced in adult animals by treatment with cytokines and/or cytotoxic drugs (Wright et al 2002; Katayama et al 2006), making the basis for innovative ligands to be come upon.

A peculiar feature is seen at the central position in the CRD of galectins: the **Trp** moiety that establishes carbohydrate/π (CH/π) contacts to a ligand's galactose residue. This **Trp** is present in birds, fish and amphibians GRP in the same place as in galectins, whereas mammals consistently present a substitution by **Arg**/**Lys**. *Gallus gallus* was explicity chosen in this work for mapping the structures of all galectins

to track down themes of recognition and divergence, and to analyze the difference between mammal GRP structure (*e.g.* **PDBID: 2JJ6**) and avian GRP structure (Oda and Kasai 1983; Cooper 2002; Wälti et al 2008a; Ruiz et al 2015).

General pre-requisites for the adhesion/growth-regulatory galectin group are fulfilled in the species *G. gallus* with a total of only five canonical galectins and one expressed sequence tag with similarity to a bone marrow human GRP found in chicken bursal lymphocytes (**C-GRP**). C-GRP arises as an important protein with very strong positive selection and because (i) this feature emerges from inter-species considerations that imply special functionality, and (ii) the obvious requirement to bring characterization of the chicken GRP to the same levels as it has been done for the five canonical chicken galetcins (**CG**s). Over 200 compounds, comprising glycans, [lipo]polysaccharides and glycosaminoglycans have been tested by microarray to cover a wide spectrum of glycan structures that might bind to C-GRP, along with CG-1A, -1B and -2 as positive controls. However, in spite of the presence of the active-site **Trp**, no significant reactivity between C-GRP and any glycocompound could be detected. This has an analogy in C-type lectin-like domains that have lost capacity to bind sugars but have gained reactivity to other types of epitopes (Wälti et al 2008b; Ruiz et al 2015; García, Flores-Ibarra, Michalak et al 2016).

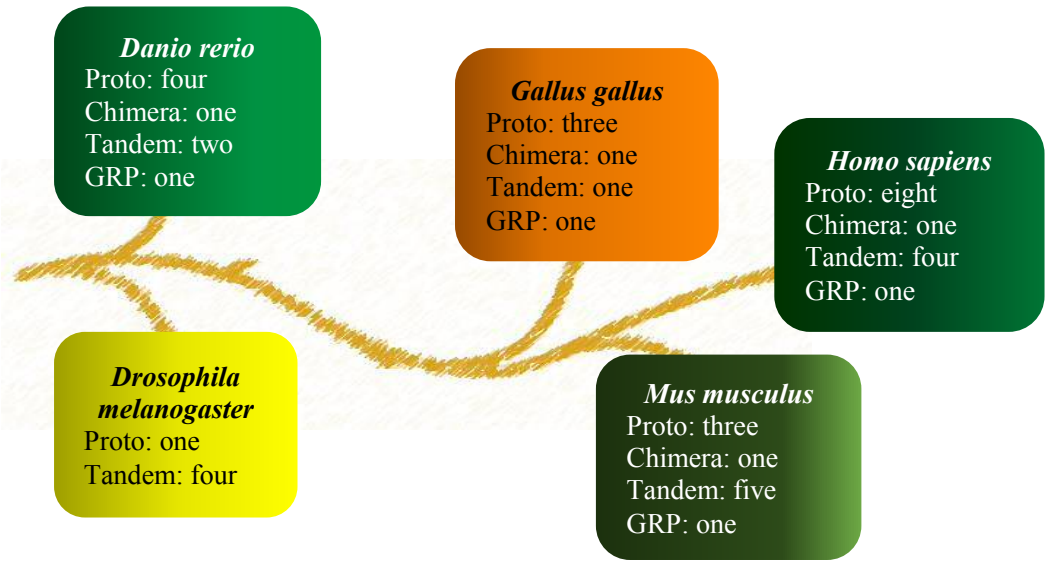


Fig 5. Galectins and GRP phylogeny. Members of the galectin family have arisen from an ancestral gene by divergence, though not all of them have canonical carbohydrate-binding activity. The detection of a human Galectin-Related Protein revealed a protein with sequence similarity to galectins but unknown natural ligands, present in other orders notwithstanding not in other phyla (Cooper et al 2002; García, Flores-Ibarra, Michalak et al 2016).

In conclusion, the current insight into how sugar-encoded messages are translated by interactions between galectins and glycans provides a deep understanding into fundamental mechanisms, from cell activation, differentiation and death/survival to resolution of inflammation, chronic infection, autoimmunity and neurodegeneration (Lichtenstein and Rabinovich 2013; Solís et al 2015). Relevance of lectin-glycan recognition systems is thus undeniable. Drug design and specificity in effector molecules to activate, inhibit or change galectins' functions are based on the structural studies of the protein of interest. Being its CRD solved nineteen years ago, no part of the Gal-3 N-PG domain has been solved thus far (Flores-Ibarra et al 2015). Many of this protein's functions relies in the N-PG domain, such as protein oligomerization and post-translational phosphorylation in the N-LD **Ser6** residue mediating a key change that activates apoptosis, autophagy and nuclei transport into the cytoplasm (Lichtenstein and Rabinovich 2013). Structural knowledge of the N-PG section will complement the biological knowledge on Gal-3, for example, by pinpointing the accommodation of **Ser6** in a Gal-3 oligomer, or the properties for such oligomerization. The high flexibility inherent to the N-PG has made it such a difficult task to crystallize it, thus by choosing a protein engineering approach, this work produces the first structural resolution on a section of the Gal-3 N-PG domain.

On the other hand, a fully comprehensive study of the galectin family means resolving for all the types predicted by DNA database mining present in an organism. Besides galectins, a particular Galectin-Related Protein has been described in all vertebrates (but not in invertebrates) with a sequence homology that places it among the members of the galectin family but that lacks the main trait describing galectins: the ability to bind β-galactosides. This binding is profusely described as to happen through specific H-bonds and CH/π interactions. A notorious difference between mammal GRP and avian GRP is that the former lacks the presence of the **Trp** residue responsible for the CH/π interaction with the galactose moiety, while it is present in the latter (García, Flores-Ibarra, Michalak et al 2016). Still, avian GRP does not bind β-galactosides. This work answers to the necessity of studying the difference between C-GRP and human GRP structures and the difference with canonic galectins that may account for its lack of binding to canonical ligands. The model of *Gallus gallus* is ideal for characterizing all galectin types present in vertebrates in a minimum amount of proteins.

OBJECTIVES

MATERIAL AND METHODS

OBJECTIVES

- To obtain high-purity samples of Gal-3[N VII-IX] for structural experiments.
- To crystallize Gal-3[N VII-IX] and obtain good quality, good sized diffracting crystals.
- To determine the three-dimensional structure of Gal-3[N VII-IX] by X-ray crystallography.
- To biophysically characterize Gal-3[N VII-IX] and Gal-3_{FL} by means of Small-Angle X-ray Scattering.
- To explain the participation of the N-terminal poly-Proline/Glycine domain in the Galectin-3 structure, that may shed light on function.

- To obtain high-purity samples of C-GRP-C for structural experiments.
- To crystallize C-GRP-C and obtain good quality and good sized diffracting crystals.
- To determine the three-dimensional structure of C-GRP-C and compare the structure to that of human GRP-C and canonical CGs.
- To biophysically characterize the protein by means of Small-Angle X-ray Scattering and Analytical Ultracentrifugation.
- To relate the lack of affinity for β-galactosides of the Galectin-Related Protein regarding its structure and cell location.

MATERIAL AND METHODS

1. PROTEIN EXPRESSION AND PURIFICATION

The group of Prof. Hans-Joachim Gabius at Ludwig-Maximilians University (Munich, Germany) provided purified samples of proteins human Galectin-3 and chicken Galectin-Related Protein. More specifically, studies were made with **full length Gal-3 (hGal-3$_{FL}$)** and an engineered form **HUMAN GALECTIN-3 DELETION MUTANT Δ23-75 (GAL-3[N VII-IX])**, **full length chicken Galectin-Related Protein (C-GRP$_{FL}$)** and **CHICKEN GALECTIN-RELATED PROTEIN DELETION MUTANT Δ1-36 (C-GRP-C)**. As proof of concept and for further tests, cellular expression and purification of C-GRP-C was also made in-house.

1.1. *E. coli* expression of C-GRP-C

1.1.1. DNA cloning

GRP synthetic cDNA was provided by ATG: biosynthetics GmbH (Merzhausen, Germany) and oligonucleotides **GRP*NdeI*** and **GRP*XhoI*** by Sigma-Aldrich (Missouri, USA). Both the cDNA and oligonucleotides were stored at 253 K.

First, cDNA was amplified using polymerase chain reaction (**PCR**) with the designed specific primers, and a positive control, using KOD polymerase (Novagen, Merck Biosciences, New Jersey, USA) and dNTPs (Thermoscientific, Massachusetts, USA). The PCR protocol was:

POOL MIXTURE	
Buffer for KOD Hot Start pol 10X	5 µl
dNTPs (2 mM)	5 µl
$MgSO_4$	3 µl
H_2O	32.5 µl
KOD polymerase	1 µl

DNA MIXTURE	
GRP cDNA	0.5 µl
NdeI	1.5 µl
XhoI	1.5 µl
Positive control	2 µl

The melting temperature (T_m) was calculated using the following formula [since the Mastercycler (Eppendorf, Hamburg, Germany) used for these PCRs uses Celsius as temperature unit, calculations were made in Celsius degrees]:

T_m = [(G + C) x 4 ºC] + [(A + T) x 2 ºC]

T_m → Number of guanines and cytosines times four plus number of adenines and thymines times two.

T_m = 64.3 ºC

Reactions were prepared in 50 µl volumes with the followoing reaction conditions:

MASTERCYCLER PROGRAM			
Initial cycle	5 mins	98 ºC	
Denaturalization cycles	30 secs	98 ºC	30 cycles
Extension cycles	60 secs	60 ºC	
KOD Polymerase cycles	2 mins	72 ºC	
Final cycle	10 mins	70 ºC	

PCR products were checked in a 1% agarose gel. Both GRP cDNA bands in the gel were cut out of the agarose gel and purified using a QIAquick PCR Purification Kit (QIAGEN, Hilden, Germany).

1.1.2. DNA ligation

Vectors for bacterial expression used for ligation of the GRP DNA fragment were of the **pET28** series of vectors with a Poly-Histidine tag (**pET28-PP**) for conveying specificity to the protein for posterior purification. DNA insertion vector was digested with *NdeI* and *XhoI* restriction enzymes using manufacturer's protocol (New England Biolabs, USA):

DIGESTION MIXTURE	
DNA vector	25 µl
Buffer 4 10X	6 µl
BSA 100X	0.6 µl
H_2O	23.4 µl
NdeI	2.5 µl
XhoI	2.5 µl

The digestion mix was left 3 hours at 310 K. Afterwards, the ligation of the GRP fragment was made, using the following protocol (New England Biolabs, USA). The ligation was done at room temperature with varying time from 2 hours to overnight. As negative control was used a ligation reaction without DNA insert:

LIGATION MIXTURE	
Buffer 10X T4 DNA ligase	1.5 µl
Vector	5 µl
GRP DNA fragment	5 µl
T4 DNA ligase	1 µl
H_2O	2.5 µl

1.1.3. DNA transformation

The ligation reactions were transformed into *Escherichia coli* DH5α chemically competent cells (Promega, Mannheim, Germany) by heat shock.

Transformation of E. coli DH5α by heat shock [all in a sterile environment]:

1. Mix 60 µl of competent cells with 10 µl of the ligation reaction and incubate for 15 mins.
2. Place cells for heat shock at 315 K for 1 min and 30 secs.
3. Incubate cells at 277 K for 5 mins.
4. Add 1 ml of 2XTY medium and place cells for recovery at 310 K for 1 hour on a shaker.
5. Spin cells for 3 mins at 6708 *g* in a MiniSpin Plus centrifuge (Eppendorf, Hamburg, Germany) and decant or pipet out all but ~50 µl of volume.
6. Resuspend cells in 200 µl 2XTY medium and plate by massive streaking in LB agar containing kanamycin (KAN, 50 µg/mL).

Plated cells were incubated overnight at 310 K to allow colonies to grow. Fifteen colonies were randomly selected and single-streaked into freshly LB-KAN agar plates. The remaining bacteria were placed in 20 µl of MilliQ water (Millipore Co., Massachusetts, USA) and heated to 373 K for 2 mins to screen the colonies by PCR analysis. The colony PCR was performed following the protocol listed below with Taq polymerase (Invitrogen, California, USA) instead of high fidelity KOD polymerase.

POOL MIXTURE	
Buffer for TAQ pol 10X (with $MgSO_4$)	5 µl
dNTPs (2 mM)	5 µl
H_2O	15 µl
TAQ polymerase	1 µl

DNA MIXTURE	
GRP-pET28	1 µl
Nde oligo (FW)	1.5 µl
Xho oligo (BW)	1.5 µl
Positive control	2 µl

MASTERCYCLER PROGRAM			
Initial cycle	5 mins	98 ºC	
Denaturalization cycles	30 secs	98 ºC	30 cycles
Extension cycles	60 secs	50 ºC	
KOD Polymerase cycles	90 secs	72 ºC	
Final cycle	10 mins	72 ºC	

Positive colonies were identified in a 1% agarose gel and sequenced in the in-house service Secugen to confirm gene integrity.

1.1.4. Heterologous protein over-expression

Protein expression can be done in an *E. coli* host strain from a vector with a tag and/or fusion partner. The choice of the host strains depends more on the nature of the heterologous protein (Miroux and Walker, 1996).

Once sequences were confirmed by DNA sequencing, plasmid DNAs were transformed in BL21(DE3) electrocompetent cells with a MicroPulser electroporator (BioRad, California, USA), mixing 1 µl of plasmid DNA in 90 µl of cells and transferring the sample to 0.1 cm electroporation cuvettes.

For recovery after electroporation, cells were added 1 ml of 2XTY medium and incubated 1 hour at 310 K on a shaker, before plating into LB-KAN plates.

Single colonies were inoculated in 10 ml of 2XTY medium containing 50 µg/ml KAN and incubated overnight on a 310 K shaker at 250 rpm. The overnight cultures were transferred into a flask with 950 ml of fresh 2XTY-KAN medium and left to grow at 310 K on a shaker until the optical density at 600 nm (OD_{600}) reached ~0.6-0.8. Protein expression was induced with **IPTG** (isopropyl β-D-1-thiogalactopyranoside, Gold Biotechnology, Missouri, USA) to a final concentration of 1 mM and incubated overnight on a 293 K shaker at 250 rpm (Rosano and Ceccarelli 2014; EMBL Protocols).

Cell cultures were harvested by centrifugation for 15 mins at 6238 *g* in a JLA8.1000 rotor (Beckman Coulter, California, USA) keeping it at 277 K during the entire process. Cell pellets were resuspended in 40 ml of lysis buffer (50 mM Tris-HCl pH 7.5, 500 mM NaCl, 0.5% Tween 20), and sonicated in a Misonix S4000 Sonicator (Hielscher Ultrasonics, Teltow, Germany). The lysate was put to centrifuge for 30 mins at 200K *g* in a 45Ti rotor (Beckman Coulter, California, USA), at 277 K, and the supernatant was collected after centrifugation.

1.2. Protein purification

Purifying a protein to homogeneity is crucial for protein crystallization, as contaminants can hamper crystal growth. Protein purification via affinity chromatography is a powerful technique that separates biomolecules on the basis of a reversible interaction between a biomolecule and a specific ligand coupled to a stationary matrix.

Complementarily, size exclusion chromatography separates molecules based on their size in which molecules with partial access to the pores of the matrix elute from the column in order of decreasing size (www.gelifesciences.com/handbooks). As such, size exclusion was the final polishing step of purification.

Gal-3[N VII-IX]

Suspensions of precipitated material were dialyzed against PBS containing 4 mM β-mercaptoethanol (BME) with a Slide-A-Lyzer Dialysis Cassette (3500 MWCO, Thermo Scientific, Rockford, USA) followed by affinity chromatography on a lactosylated Sepharose 4B column to remove any inactive material. Eluted fractions were then checked by SDS-polyacrilamide gel electrophoresis (SDS-PAGE). Solutions were concentrated and loaded in a HiPrep 16/60 Sephacryl S-100 high-resolution column (GE Healthcare, Little Chalfont, UK) equilibrated with 20 mM Na^+/K^+–PO_4 buffer pH 7.0 with 150 mM NaCl, 4 mM BME and 5 mM lactose using an ÄKTA Prime system (GE Healthcare, Little Chalfont, UK), at flow rate 0.5 $ml \cdot min^{-1}$ and 277 K.

The column was calibrated beforehand with the following molecular weight markers: blue dextran (M_r = 2000 kDa), aldolase (M_r = 158 kDa), albumin (M_r = 67 kDa), ovalbumin (M_r = 44 kDa), chymotrypsinogen (M_r = 25 kDa) and vitamin B_{12} (M_r = 1.35 kDa).

Protein-containing fractions were concentrated to 12 $mg \cdot ml^{-1}$ with a Slide-A-Lyzer Dialysis Cassette (3500 MWCO, Thermoscientific, Massachusetts, USA) and a SDS-PAGE was run, to verify its purity. Protein concentration was determined by measuring the absorbance at 280 nm using the calculated extinction coefficient of 9970 $M^{-1} \cdot cm^{-1}$ (http://web.expasy.org/protparam).

C-GRP-C

A HiTrap (GE Healthcare, Little Chalfont, UK) column binds with specificity to the Poly-Histidine tail in expression vectors. This column was set in an ÄKTA Prime system (GE Healthcare, Little Chalfont, UK), at flow rate 0.4 $ml \cdot min^{-1}$ and 277 K, with 5 ml $NiSO_4$ and equilibrated with the same lysis buffer used in the previous step. The protein with the **pET28-PP** was run through with a buffer with 50 mM Tris-HCl pH 7.5, 200 mM NaCl and a gradient of imidazole ranging from 10 mM to 500 mM for elution of the protein of interest only.

Both elutions were run separately in a Superdex 75 (GE Healthcare, Little Chalfont, UK) with buffer 20 mM Tris-HCl pH 7.5, 100 mM NaCl and 2 mM DTT. The column was calibrated as described in the previous section. The protein with the **pET28-PP** showed a peak at the C-GPR-C M.W. ~16 kDa. The Poly-Histidine tag was removed by incubation and dialysis with protease 3C, further purification was made by affinity chromatography and molecular exclusion chromatography. The final samples were concentrated to 16 $mg \cdot ml^{-1}$ with a Slide-A-Lyzer Dialysis Cassette (3500 MWCO, Thermoscientific, Massachusetts, USA) and an SDS-PAGE was run, to verify its purity. Protein concentration was determined by measuring the absorbance at 280 nm using the calculated extinction coefficient of 8730 $M^{-1} \cdot cm^{-1}$ (http://web.expasy.org/protparam).

2. PROTEIN CRYSTALLIZATION

"Finding the exact conditions to produce good crystals of a specific protein often requires many careful trials, and is perhaps more art than science" Gale Rhodes in Crystallography Made Crystal Clear.

2.1 Nucleation and crystal growth

Protein crystal growth implies at least two major stages: formation of an initial solid object, or nucleus, in the metastable phase, and the nucleus growing. Crystal growth is a first order phase transition in which there must be (**Fig 6**): (i) a crystalline surface, separating the crystal from the original phase, (ii) a metastable zone around the point of thermodynamic equilibrium, and (iii) a nucleation barrier, preventing further nucleation in any other part of the original phase. The crystallization process of a biomacromolecule requires lowering the solubility at a slow and constant rate within an adequate time frame. In the labile region of the curve, crystal nucleation and growth compete, whereas in the metastable region only crystal growth appears. In the unsaturated region the crystals dissolve (Mendoza and Moreno, 2015).

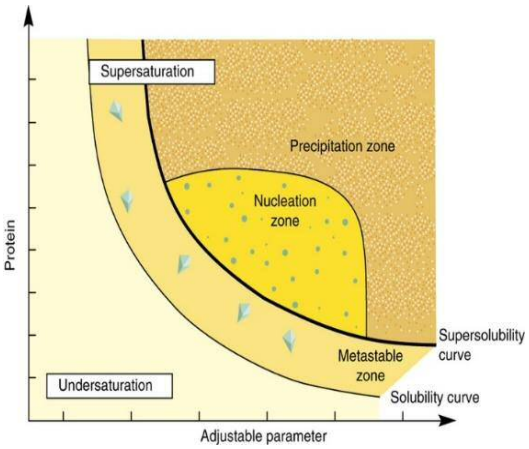


Fig 6. Protein nucleation and crystal growth. Protein solutions need to reach supersaturation to create crystalline nucleus from which crystals will grow in the metastable zone.

Practically, the four main stages of protein crystallization are (**Fig 7**):

- **(a)** phase separation.- marks the co-existence of a dense liquid in a less dense one.
- **(b)** precipitation and nucleation.- the phase separation favors the ordering of the protein molecules inside the dense droplets and the generation of nucleation sites.
- **(c)** crystallization.- around the nuclei formed, the crystals grow in proportion with the protein's density increase and the solution's density gradual reduction.
- **(d)** crystal growth.- if the process is to continue, that is, the density equilibrium is not broken, the protein crystals are susceptible to keep growing in size.

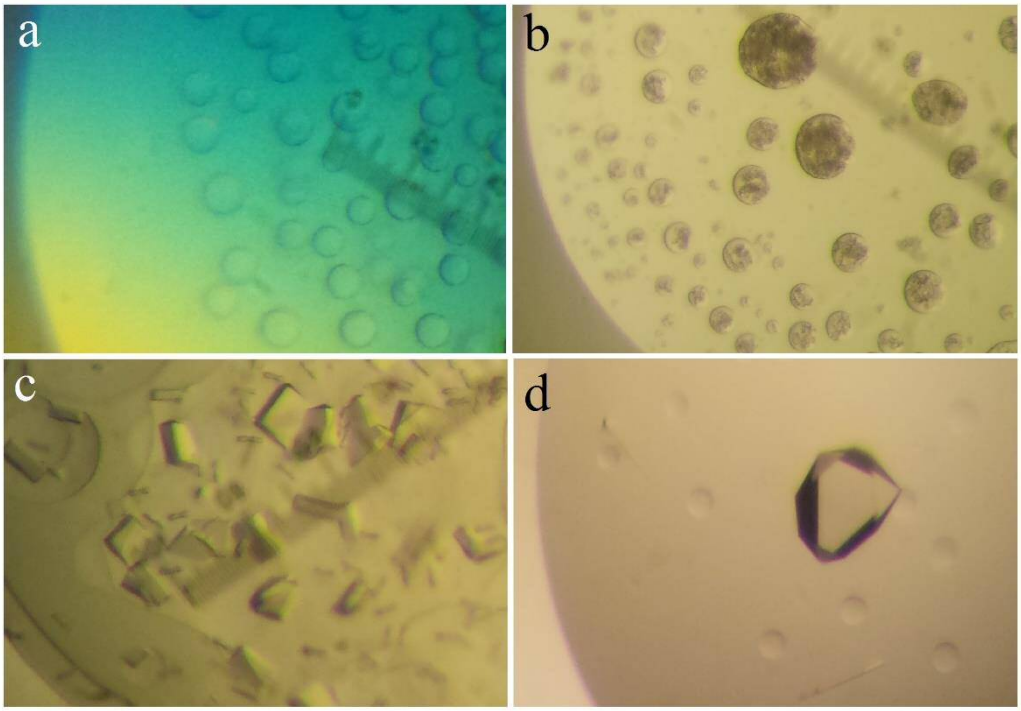


Fig 7. Stages of protein crystallization, as seen through a magnifying glass on the crystallization plates. (**a**) Phase separation, (**b**) precipitation and nucleation, (**c**) crystallization, (**d**) crystal growth.

2.2 Protein solubility

Protein solubility can be influenced by several factors, such as ionic strength, pH and counter ions, temperature and organic solvents. In aqueous solution, each ion is surrounded by an atmosphere of counter ions; this atmosphere influences the interactions of the ion with water molecules and hence the solubility. As for the pH, the more soluble a protein, the larger is its net charge, with the minimum solubility being found at the isoelectric point. The packing in solid state (in the crystal) is favored when electrostatic interactions happen without the accumulation of a net charge in high energy. Protein solubility is temperature-dependent for in the solution energy equation, $\Delta G^0 = \Delta H - T\Delta S$, the entropy term has an increasing influence with increasing temperature (Messerschmidt 2007).

GAL-3[N VII-IX]

Screening for crystals was conducted in 96-well sitting-drop plates (Swissci MRC, Molecular Dimensions, Suffolk, England) at 295 K using a Cartesian Honeybee System (Genomic Solutions, Irvine, USA) and the JBScreen Classic (Jena Bioscience, Jena, Germany), Wizard Classics I–III (Emerald Bio, Bainbridge Island, USA) and

Index (Hampton Research, Aliso Viejo, USA) commercial kits. The drops were 0.4 μl in volume, consisting of 0.2 μl protein solution and 0.2 μl precipitant, equilibrated against 50 μl reservoir solution.

Further attempts to optimize the initial crystallization conditions were performed using optimization screens prepared with a Freedom EVO liquid-handling robot (TECAN, Männedorf, Switzerland). The selected conditions were:

(i) 18%(w/v) PEG 8000, 100 mM Tris–HCl pH 8.5, 200 mM Li_2SO_4; and

(ii) 22%(w/v) PEG 8000, 100 mM MES pH 6.5, 200 mM $(NH_4)_2SO_4$.

Different grids varying PEG molecular weight (PEG 6000, PEG 8000 and PEG 10000), PEG concentration (15–30%) and pH from 6.0 (100 mM MES) to 9.0 (100 mM Tris–HCl) were prepared, with Li_2SO_4 as additive. Crystals were cryo-protected with the reservoir solution supplemented with 12.5%(w/v) PEG400, mounted on nylon loops and flash-cooled in liquid nitrogen for X-ray data collection.

C-GRP-C

Systematic screening for conditions to crystallize C-GRP-C was performed in 96-well sitting drop plates (Swissci MRC, Suffolk, England) at 295 K using a Cartesian Honeybee system (Genomic Solutions, Irvine, USA) and commercially available 96-well kits: JBScreen Classic (Jena Bioscience™), Wizard Classic Screen I–III (Emerald Bio™) and Index Crystal Screen (Hampton Research™). The drops were 0.4 μL in volume, a mixture of 0.2 μL of the protein solution and 0.2 μL of precipitant, equilibrated against 50 μL of the reservoir solution. Diffracting crystals grew after a couple of weeks in the presence of 2.0 M $(NH_4)_2SO_4$, 100 mM MES (pH 6.5) and 5% (w/v) PEG 400. Crystals were cryo-protected with the reservoir solution supplemented with a 1.0 M sodium malonate solution, mounted on nylon loops and flash-cooled in liquid nitrogen for X-ray data collection.

3. PROTEIN CRYSTALLOGRAPHY

3.1 Principles of crystalline symmetry

"The characteristic symmetry of a physicochemical phenomenon is the highest symmetry compatible with the phenomenon existence" Pierre Curie.

In the experiment performed by Friedrich and Knipping in 1912, using $CuSO_4$, X-rays produced with a tungstene anti-cathode and a photographic plaque as detector, an interference pattern was observed that demonstrated properties like (Friedrich et al 1912):

- crystalline solids have a periodic arrangement;
- X-rays wavelengths have similar dimensions to the interatomic distances.

Spatial relations between the constituents of a crystalline solid occur at short and long ranges, hence influenced by both the physical and chemical properties of the crystal. The basic repeating unit from which the crystal is constructed is the *unit cell* (Cowtan 2001; Mendoza and Moreno 2015). For each crystal, a *reciprocal lattice* of the X-ray scattering can be constructed, which is useful when interpreting the atoms positions (see Ewald's sphere below).

3.2. Solvent content in crystals

Crystals solvent content was broadly discussed by Matthews et al (1968), who established for the first time how to quantify water content inside a protein crystal. The Matthews volume (V_M) is defined by the crystal unit cell volume radius over the molecular weight of the protein. Thus, V_M represents the crystal volume per mass unit of molecular protein, which is practically independent from the *asymmetric unit volume* and mainly dependent on the solvent content. Values for V_M generally range from 1.6 $Å^3 \cdot Da^{-1}$ to 3.5 $Å^3 \cdot Da^{-1}$ (Messerschmidt 2007).

3.3. Synchrotron radiation

As electrically charged particles such as electrons or positrons of high energy are kept under the influence of magnetic fields and travel in a pseudocircular trajectory, *synchrotron radiation* is emitted. In the practice, these particles are injected into a storage ring directly from a linear accelerator or through a booster ring, and circulate in a high vacuum for several hours at a relative energy. In order to keep the bunched particles traveling in a near-circular path, an arrangement of bending magnets is set up around the storage ring. As the particle beam traverses each magnet, the path of the beam is altered, and synchrotron radiation is emitted. The loss of energy of the

particle beam is compensated by an oscillating radiofrequency electric field at each cycle. Synchrotron radiation is highly polarized, and can be channeled through different beamlines for use in research (Messerschmidt 2007).

3.4. Data collection

The primary beam in a synchrotron radiation facility leaves the X-ray source and passes the X-ray optics collimator. The crystal is mounted on a goniometer head capable of perform spatial movements around the center of the beam source, either in a quartz capillary or in a flash-cooled cryo-loop. The X-ray detector, which registers the diffracted intensities, is mounted on a device which allows the translation and rotation of the detector. A piece of lead is placed in the path of the primary beam just behind the crystal to prevent damage to the detector and superfluous gas scattering. The X-rays are scattered from every point in the crystal, with a strength in proportion to the concentration of electrons at that point. An X-ray reflection is generated when a point of the reciprocal lattice lies on a sphere of radius 1/λ, whose origin is 1/λ away from the origin of the reciprocal lattice in the direction of the primary beam. The direction of such a diffracted beam is along the center's connection of the so-called Ewald sphere (radius 1/λ), and the intersection of the reciprocal lattice point on this sphere (**Fig 8**). The apparent reciprocal lattice extends only to a given radius which defines the resolution sphere. Each individual exposure is processed and the data stored electronically in a computer. These raw data images are evaluated subsequently to provide the intensities and geometric reference values (*indices*) for each collected intensity (Cowtan 2001; Messerschmidt 2007).

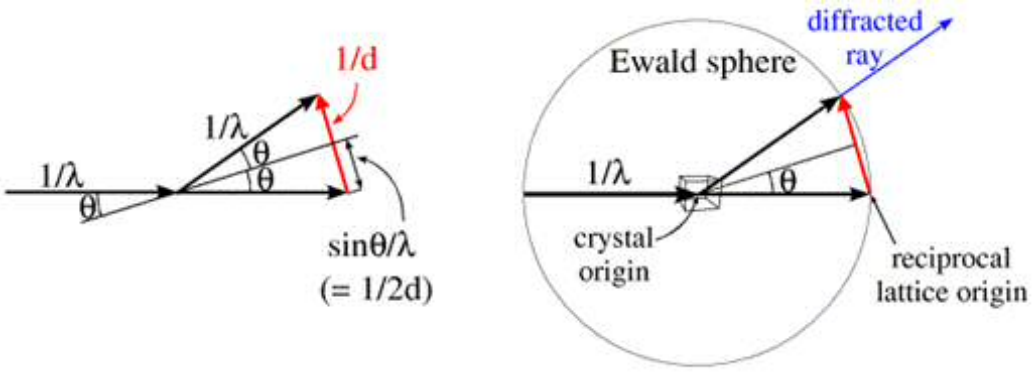


Fig 8. The Ewald sphere for reciprocal lattices. The angle 2θ = 0–180º. The diffracted ray can go in any direction in three dimensions, the vector representing it can have its tip anywhere at the surface of a sphere with a radius of 1/λ. The diffracted ray has its base at the center of the sphere, considered the origin of the crystal. But the origin of the reciprocal space has to be at the point where the direct beam exits the sphere (Drenth 1994).

3.5. Rational mathematical representation of X-ray diffraction

The simplest periodic process in time (oscillation) is given mathematically by the sine or cosine function. If the process is recurring v times per second, it is written as:

or **(3.5.1)**

where A is the amplitude and ν is the frequency. The actual value of $2\pi\nu t$ is called the phase angle, and this determines the momentary state, that is the *phase*.

Computation becomes much easier if the imaginary exponential function is used instead of the cumbersome trigonometric functions.The trigonometric functions are related to the exponential function by Euler's formula:

(3.5.2)

This representation leads to:

(3.5.3)

where z is a complex number, the representation point of which rotates on a circle of radius A with an angular velocity of $2\pi\nu$. The projections onto the real and imaginary axis are:

; **(3.5.4)**

An equation between complex numbers means that both the real and the imaginary part of each side fulfill the equation. Thus, the real and the imaginary parts of the equation can be taken as the physical meaning of the equation.

The square of the amplitude A^2 can be obtained in the complex computation by multiplying the oscillation magnitude by its complex conjugated value :

(3.5.5)

3.6. X-ray scattering by a crystal

As X-rays are electromagnetic waves, Maxwell's equations are valid in the physical process of X-ray scattering by atoms. X-ray propagate in a vacuum with the velocity of light; they are transversal waves with the electric field component **E** and the magnetic field component **H**, respectively, oscillating perpendiculary to the direction of

propagation and perpendicularly to each other. For a three-dimensional crystal, the unit cell is spanned by the unit cell vectors **a**, **b** and **c**, and is repeated periodically by the corresponding vector shifts , being integers, in the respective spatial direction. The scattering of the X-ray waves by a crystal can then be written:

(3.6.1)

The three sum terms are geometric series which can be calculated. The intensity can be obtained by squaring . The squares of the complex magnitudes are established by multiplying it with its complex conjugate. Carrying out these multiplications reveales concise expressions and the intensity :

(3.6.2)

The last three factors are known as the interference function . The function has maxima of when the three subsequent conditions are fulfilled (*h*, *k* and *l* being the corresponding diffraction intensities at the reciprocal lattice points):

(3.6.3)

These conditions are known as Laue equations. If the constant magnitudes in Eq **(3.6.2)** are neglected, *the equation for the total scattered wave for a three-dimensional crystal with a unit cell containing N atoms* is:

(3.6.4)

with from Laue's equation, and:

(3.6.5)

with being the amplitude and being the phase angle. The intensity of the scattered wave is obtained as the structure factor multiplied by its complex conjugate value, according to this equation:

(3.6.6)

3.7. The Patterson function and Friedel's law

The measured X-ray intensities are proportional to the square of the absolute value of the structure factor according to Eq **(3.6.6)**. It is possible to use the intensities directly to calculate a function that contains structural information. By calculating a convolution of the electron density with itself, Patterson (1934) showed that this is just the Fourier transform (FT) of the intensities:

(3.7.1)

In the case of real atomic scattering factors *f*, the diffraction intensities are centro-symmetric according to Friedel's law:

(3.7.2)

From Eqs **(3.6.4)**, **(3.7.1)** and **(3.7.2)**, *i.e.* Friedel's law, it is possible to obtain the following equation:

(3.7.3)

that shows the equation for the structure factor is the FT of the electron density (Messerschmidt 2007).

3.8 Software for crystallographic resolution

On this mathematical basis, several specialized softwares have been developed for resolving protein structures by X-ray crystallography. The programmers' web pages

contain thourough information on both the developing and the usage of their crystallography software packages (**CCP4**: Evans and Murshudov 2013; **Phenix**: Adams et al 2010; **XDS**: Kabsch 2010; **Coot**: Emsley et al 2010; **PyMol**: DeLano 2012).

Image processing with XDS can be done in Unix or Mac OS platforms, and provides important data such as the total number of unique reflections, the resolution range, the unit cell dimensions (*a, b, c; α, β, γ*) and the symmetry of the crystal lattice. This software's output is the first step to decide on the quality of the data (Kabsch 2010). The reflections' scaling is done with the program Aimless (Karplus and Diederichs 2012; Evans 2011), from the CCP4 software package (Evans and Murshudov 2013), which scales the diffraction data merging intensities, and alternatively with scaled but unmerged observations, giving out results in terms of the Fourier transform coefficients.

From the X-ray diffraction data, the amplitudes are derived but not the phases (Cowtan 2001). If both parameters were known, the protein's electronic density could be obtained by simply applying the inverse Fourier transform to the experimental structure factors. Thus, it is necessary to use an alternative method that allows the estimation of the structure phases factors in order to calculate the experimental electronic density: this is known as *the phase problem*. For the calculation of the phases, different approaches have been used. In this work, the method used was that of *molecular replacement*, with Phenix (Adams et al 2010).

Molecular replacement

When the molecule under study is reasonably similar to another molecule whose structure is already known, the molecular replacement method allows phases to be obtained from the known structure. This is done by calculation of the rotation and translation functions: the known molecule is first rotated in three dimensions, and for each orientation, structure factors are calculated from the model. The agreement between the calculated structure factors and the observed values from the diffraction experiment is used to identify the orientation of the known molecule that most closely matches that of the unknown molecule in the crystal. Next, the oriented model is placed at every possible position in the unit cell, and again the agreement of the structure factors is used to identify the correct translation (Adams et al 2010).

Crystallographic refinement

Once phase estimates (or phase probability distributions) are available, an initial electron density map may be calculated. At this point it is possible to apply chemical

knowledge to improve this map and the phases. The electron density in regions of disordered solvent does not show identifiable features, so the electron density map can be improved by flattening the solvent and, furthermore, sharpening the features inside the protein region. Once the map has been modified, it is used to calculate a new set of structure factors and phases. Averaging the density between these copies also reduces the noise level in the map. All these calculation may be repeated over several cycles (Cowtan 2001).

Phenix software suite is a highly automated system for macromolecular structure determination that can rapidly arrive at an initial partial model of a structure, through the general formula **(3.7.3)**, which numerically resolved, is the discrete Fourier transform of the electronic density in a sphere of radius r~2 Å around each atom. This achievement has been made possible by the development of new algorithms for structure determination, such as maximum-likelihood molecular replacement (Phaser), automated macromolecular refinement (phenix.refine), and iterative model-building. Phenix builds upon Python, the Boost.Python Library and C++ to provide an environment for automation and scientific computing. Final evaluation of the refinement iterations quality is made by calculating the *R*-factor, which establishes the relative difference between the *observed* structure factors (F_{obs}) and the *calculated* structure factors for a theoretical model (F_{calc}), and it is determined with the following equation (Adams et al 2010):

(3.8.1)

Visualization and manual correction of the generated map around the amino acid structure is done with the software Coot (Emsley et al 2010), a graphical interface that also allows the calculation of electronic density, atypical values in stereochemical parameteres, distances and bond angles, and correspondence with Ramachandran values for the dihedral angles of the protein residues (Ramachandran and Sasisekharan 1968).

3.9. Principles of the self-rotation function

Proteins frequently crystallize with more than one copy of the molecule in one unit cell, or asymmetric unit in the case of crystals with internal symmetry. The self-rotation function can be used to determine the orientation and order of local symmetry, an information that it is also needed for refinement. Multi-subunit proteins often exhibit *non-crystallographic symmetry* relating the subunits. The Patterson

function contains a set of peaks representing intramolecular vectors (self-vectors) for each molecular orientation found in the crystallographic cell. These peaks are distributed around the Patterson origin within a volume of double the molecular diameter, differentiating them from the intermolecular vectors (cross-vectors) which, depending on the molecular packing, can occur anywhere in the Patterson function. The self-rotation function *R(C)* is the auto-correlation function of the native Patterson, $P(\mathbf{x})$, with a rotated version of itself:

(3.9.1)

where $\mathbf{C}(\mathbf{x})$ is the three-dimensional rotation matrix that relates the rotated coordinate system to the stationary one, and:

— **(3.9.2)**

The *R(C)* function will have a maximum of orientations where two molecules in the unit cell superimpose.

Visualization is easier with the spherical polar angle convention that accomplishes any rotation by an appropriate spin κ about a properly chosen axis, specified by two angles ϕ and ω. A molecular *n*-fold rotation will be found in this system in a $\kappa = 360^{\circ}/n$ section. Thus, self-rotation function peaks corresponding to a two-fold axis occur in the $\kappa = 180^{\circ}$, a three-fold in the $\kappa = 120^{\circ}$ section, and so forth (Rossmann and Blow 1962; Cowtan 2001).

GAL-3[N VII-IX]

X-ray diffraction data were collected using a PILATUS 6M detector (Dectris) on the BL13 XALOC beamline at the ALBA synchrotron. Crystals were rotated with an oscillation angle of 0.5º and an exposure time of 0.5 s per image, and had a resolution of 3.3 Å. The data were processed using XDS (Kabsch 2010) and integrated and scaled using Aimless (Evans et al 2011; Evans & Murshudov 2013).

New crystals were obtained in conditions of 18%(w/v) PEG 8000, 100 mM Tris–HCl pH 8.5, 200 mM Li_2SO_4. Data collection showed 2.2 Å of resolution. The structure was solved by molecular replacement using Phaser (Adams et al 2010), and the Gal-3 CRD-only structure (**PDBID: 1A3K**; Seetharaman et al 1998) and the 3.3 Å

resolution structure (Flores-Ibarra et al 2015) as search models, expecting an occupancy of eight to twelve molecules in one unit cell. The initial model was first refined with phenix.refine (Adams et al 2010) and alternating manual building with Coot (Emsley et al 2010). The final model was obtained by repetitive cycles of refinement; solvent molecules, lactose and sulfate molecules were added automatically and inspected visually for chemically plausible positions. The MolProbity software was used to check the protein was in allowed regions of the Ramachandran plot (http://molprobity.biochem.duke.edu/). Data collection statistics are shown in Table 1 (see Results). Structural figures were drawn with PyMOL (DeLano 2012).

C-GRP-C

X-ray diffraction data were collected on a single crystal using a Pilatus 6M detector at the BL13 XALOC beamline at the ALBA synchrotron. A total of 960 rotation images were collected with an oscillation angle of 0.25°. The resulting data set was processed using XDS (Diederichs and Karplus 1997; Kabsch 2010) and scaled using Aimless (Evans 2011; Evans and Murshudov 2013). The structure was solved by molecular replacement using Phaser (Adams et al 2010) and the human GRP structure (**PDBID: 3B9C**; Zhou et al 2008) as a search model. The initial model was first refined with Refmac5 (Evans and Murshudov 2013), alternating manual building with Coot (Emsley et al 2010). The final model was obtained by repetitive cycles of refinement using Phenix (Adams et al 2010). Solvent molecules were added automatically and inspected visually for chemically plausible positions. The placement of PEG, ethylene glycol and sulfate molecules was done using Coot (Emsley et al 2010). The MolProbity software was used to check the protein was in allowed regions of the Ramachandran plot (http://molprobity.biochem.duke.edu/). Data collection statistics are shown in Table 2 (see Results). Structural figures were drawn with PyMOL (De Lano 2012).

4. PROTEIN BIOPHYSICAL CHARACTERIZATION

4.1. Small-Angle X-ray Scattering

A complementary technique for X-ray diffraction is the so-called Small-Angle X-ray Scattering (**SAXS**), which provides structural information about the size, shape and flexibility of a molecule in solution (Putnam et al 2007; Grant et al 2015). Though SAXS data resolution is semi-automatized through specialized software, the major challenge data acquisition presents is radiation damage. Ionizing radiation can cause biomacromolecules to form high-molecular-weight oligomers through cross-linking reactions, disulfide bond formation and/or other hydrophobic/electrostatic interactions, all of which produce changes in the protein's tertiary structure (Franke et al 2012). These effects are reflected in parameters changes such as the Guinier plot (Guinier and Foumet 1955), the radius of gyration (***R_g***) and the maximum particle dimension (***D_{max}***) (Le Maire et al 1990).

The final Guinier plot and the calculated R_g for the global size of a molecule are determined through a somewhat subjective interpretation of what it is an acceptable linear region. The Guinier plot linearity is an important pre-requisite to ensure the sample's monodispersion (Jacques and Trewhella 2010). The assessment of these values is automatized in the software *AutoRg* (Petoukhov et al 2012), that determines the linearity and calculates R_g with the formula:

(4.1.1)

where D_{max} is the maximum particle dimension, ***r*** is the interatomic distance, and ***P(r)*** is the pair distribution function (Putnam et al 2007; Petoukhov et al 2012). D_{max} calculation can then be done from the distribution function ***P(r)*** taking the intensity at angle zero (*I(0)*). In the practice, data cannot be collected to the infinite so the final calculation of D_{max} is made using an indirect method of the Fourier transform, with the program GNOM, finally averaging the calculations with DAMAVER, of the software package ATSAS (Svergun 1992).

GAL-3[N VII-IX]

Using the same procedure for sample purification in Section **1.2.2** of Material and Methods, new Gal-3[N VII-IX] samples were purified and prepared at a series of concentrations of 3, 4, 6, 8 and 10 $mg \cdot ml^{-1}$, and Gal-3 full-length ($Gal\text{-}3_{FL}$) samples

were purified and prepared at a series of concentrations of 2, 4, 6 and 8 mg·ml^{-1}. SAXS data on Gal-3_{FL} and Gal-3[N VII-IX] were collected on the BM29 beamline at the European Synchrotron Radiation Facility (ESRF, Grenoble, France) using the BioSAXS robot and a Pilatus 1M detector (Dectris AG, Baden-Daettwil, Switzerland) with synchrotron radiation at a wavelength λ=0.1 nm.

For each measurement, ten frames were obtained at 1 s exposures of an 80 µl sample flowing continuously through a 1 mm diameter capillary during X-ray exposure, and protein-free buffer processed as control. The scattering images were spherically averaged and buffer's scattering intensities subtracted using software from the ESRF (Grenoble, France). The R_g was calculated with GNOM, which also gives the distance distribution function *P(r)* and the D_{max}. Bead models were obtained using the ATSAS software package. Each model was produced from 20 runs of DAMMIN that were combined and filtered to produce an averaged model using DAMAVER (Svergun 1992).

C-GRP-C

C-GRP-C samples precipitated in $(NH_4)_2SO_4$ were set for overnight dialysis in PBS 1X and BME. Purification was made using a molecular exclusion chromatography column Superdex 75 with and elution buffer containing 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA and 1 mM DTT. Purified C-GRP-C samples were prepared at a series of concentrations of 2, 4, 6 and 8 mg·ml^{-1}. SAXS data on C-GRP-C were collected on the BM29 beamline at the European Synchrotron Radiation Facility (ESRF, Grenoble, France) using the BioSAXS robot and a Pilatus 1M detector (Dectris AG, Baden-Daettwil, Switzerland) with synchrotron radiation at a wavelength λ = 0.1 nm.

For each measurement, ten frames were obtained at 1 s exposures of a 100 µL sample flowing continuously through a 1 mm diameter capillary during X-ray exposure, and protein-free buffer processed as control. The scattering images were spherically averaged and buffer scattering intensities subtracted. The R_g, *P(r)* and D_{max} were calculated with GNOM, Bead models were obtained using the ATSAS software package. Each model was produced from 20 runs of DAMMIN that were combined and filtered to produce an averaged model using DAMAVER (Svergun 1992).

4.2. Analytical ultracentrifugation

Analytical ultracentrifugation (**AUC**) is a versatile and powerful method for the quantitave analysis of macromolecules in solution. AUC has broad applications for

the study of biomacromolecules in a wide range of solvents and over a wide range of solute concentrations. Using specialized sample cells and modern analysis software, researchers can use sedimentation velocity to determine the homogeneity of a sample and define whether it undergoes concentration-dependent association reactions, determining the nature of the species present in solution and their interactions.

When the centrifugal force is sufficiently small, an equilibrium concentration distribution of macromolecules is obtained throughout a cell where the flux due to sedimentation is exactly balanced by the flux due to diffusion. The shape of this concentration gradient can be derived using a variety of approaches. For an ideal single non-interacting species, the equilibrium radial concentration gradient, ***c(r)***, is given by:

(4.2.1)

where c_0 is the concentration at an arbitrary reference distance s_0. The term $M_b\omega^2/RT$ is often referred to as the reduced molecular weight, σ. Sedimentation equilibrium experiments provide a very accurate way to determine M and consequently the oligomeric state of biomolecules in solution. Deviations from the simple exponential behavior described by Eq **(4.2.1)** can result from the presence of either multiple non-interacting macromolecular species, or thermodynamic non-ideality (Cole et al 2008).

C-GRP-C

C-GRP-C solutions of 0.2, 1 and 2 $mg \cdot ml^{-1}$ were prepared and cleared by a centrifugation step for 10 mins at 16,000 *g*. Sedimentation velocity AUC experiments were set at 277 K in an Optima XL-I instrument (Beckman Coulter, Krefeld, Germany) equipped with an AN50-Ti rotor at 180K *g*. Differential sedimentation coefficients were calculated by least-squares boundary modeling of the experimental data using the *c(s)* method implemented in the program SedFit version 14.7 (Schuck et al 2015).

RESULTS

RESULTS

GAL-3[N VII-IX]

Protein purification

Galectins display high-affinity binding to lactose, a property that can be used to purify them by affinity chromatography on β-galactoside supports. A suspension of human Gal-3[N VII-IX] in ammonium sulfate was set for dialysis overnight, and afterwards bound to a lactosylated Sepharose 4B column (as described in Material and Methods). After loading, the column was washed with approximately 5 column volumes of washing buffer (20 mM PBS, pH 7.2, 150 mM NaCl and 4 mM BME) until a stable baseline was reached. The protein was then eluted with buffer containing 200 mM lactose and 4 mM BME.

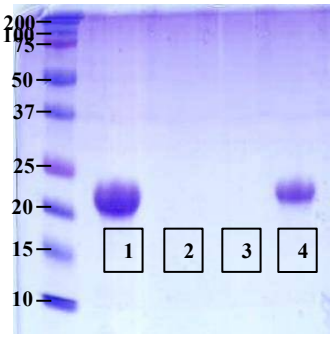


Fig 9. SDS-PAGE gel of hGal-3[N VII-IX]. Molecular weight marker (far left), protein after dialysis **[1]**, flow-through **[2]**, wash fraction [**3]** and elution fraction **[4]**.

Controls for purity were performed by standard SDS-PAGE (**Fig 9**). The lack of protein in the flow-through and wash fractions underscores the optimal binding of the protein to the lactose matrix.

In order to remove any other contaminant, a step of further purification of the sample was carried out through size exclusion chromatography (HiPrep 16/60 Sephacryl S-100) on an ÄKTA Prime equipment (**Fig 10a**). The last elution peak of the chromatogram corresponding with the appropriate size (~21 kDa) of the protein (lanes 17 to 21), showed a pure protein as could be seen by SDS-PAGE (**Fig 10b**). These fractions were pooled together, concentrated with a Slide-A-Lyzer Dialysis Cassette (3500 MWCO, Thermo Scientific, Rockford, USA) to 16 $mg \cdot ml^{-1}$ for crystallization

experiments, and to concentrations of 3, 4, 6, 8 and 10 mg·ml^{-1} stored at 193 K for SAXS experiments.

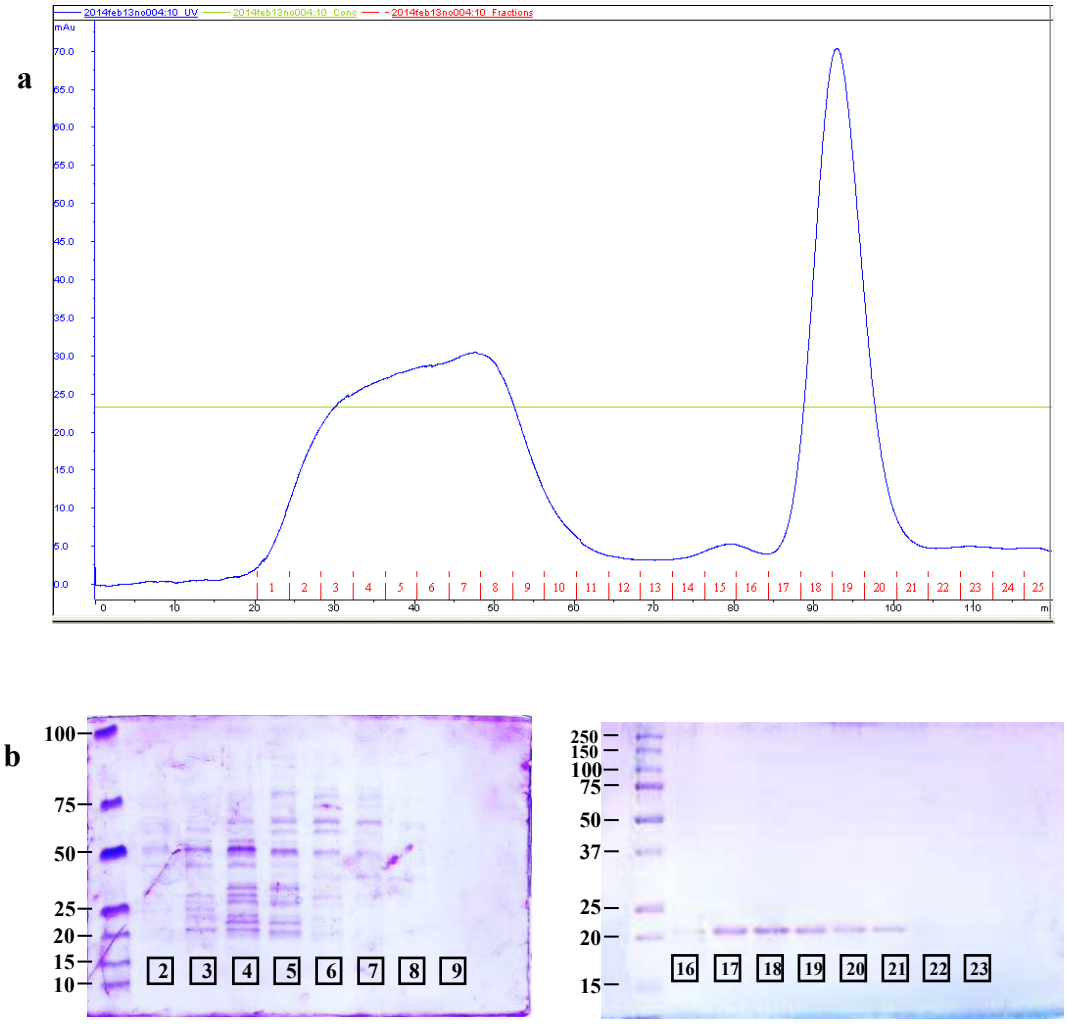


Fig 10. Size exclusion chromatography, Gal-3[N VII-IX] a) UV chromatogram, b)SDS-PAGE. Two marked peaks were found in the size exclusion chromatography, all fractions were collected and run through an SDS-PAGE to test their purity. The second peak fractions showed bands with high purity in the molecular weight of Gal-3[N VII-IX] ~21 kDa.

The purification protocol for full-length Gal-3 was similar to the previous one for the Gal-3[N VII-IX] variant.

X-ray diffraction data resolution

Fig 11. Crystals of Gal-3[N VII-IX]. Image of grown crystals seen with polarized light.

Crystals of Gal-3[N VII-IX] grew to a size of 0.05 x 0.05 x 0.03 mm plates (**Fig 11**) in the conditions: *(i)* 18%(w/v) PEG 8000, 100 mM Tris–HCl pH 8.5, 200 mM Li_2SO_4; and *(ii)* 22%(w/v) PEG 8000, 100 mM MES pH 6.5, 200 mM $(NH_4)_2SO_4$. The crystals were cryoprotected, stored in a dry-shipping Dewar and taken to the light source line XALOC (ALBA synchrotron, Cerdanyola del Vallès, Spain) for X-ray diffraction data collection. This batch of crystals diffracted to a resolution of 3.3 Å (**Fig 12**), from which some information was obtained. For instance, data analysis revealed that crystals belonged to the orthorhombic space group $P2_12_12_1$, as indicated by systematic absences. Data processing showed signs of radiation damage after exposure ($\varphi = 200º$), thus limiting the resolution value to ensure better quality. However, at 3.3 Å resolution, a noticeable signal was observed as indicated by the value $I/\sigma(I) \sim 10.1$ (overall) and 2.8 (last resolution shell), with 99.8% completeness and an R_{merge} of 0.169.

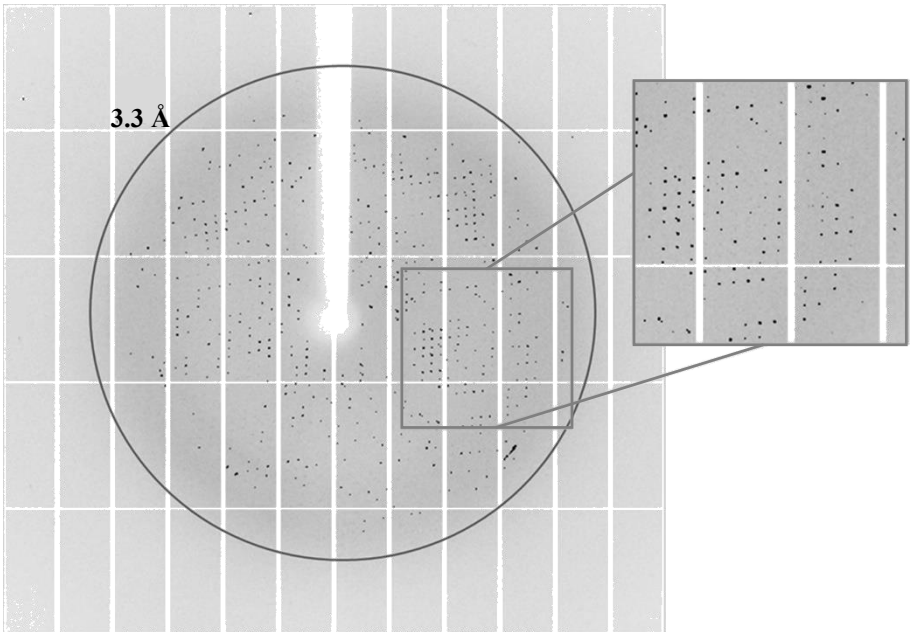


Fig 12. X-ray diffraction of Gal-3[N VII-IX]. The X-ray diffraction image is at $\varphi = 60º$ with a resolution ring at 3.3 Å.

Also, a new feature appeared for this structure that had not been reported for Gal-3 CRD crystal structures: the protein is forming oligomers. A method to quantitatively determine the presence of more than one protein molecule in a unit cell is through the solvent content determination as described in the previous section. The Matthews coefficient for Gal-3[N VII-IX] was indicative of the presence of between six (V_M = 4.28 $Å^3Da^{-1}$, solvent content 71.04%) and twelve (VM = 2.14 $Å^3Da^{-1}$, solvent content 42.08%) molecules in the asymmetric unit, being difficult to narrow down the actual number of molecules. To facilitate this task, Patterson self-rotation function analysis was performed using MOLREP (Vagin and Teplyakov 2010). This analysis showed peaks corresponding to a 4-fold axis (with κ = 90º), parallel to the *z* axis, and two 2-fold axes (with κ = 180º) in the *xy* plane (**Fig 13**). These results are summarized in **Table 1** and constituted the first published information on Gal-3 with a section of the N-PG and its oligomerization in crystal form (Flores-Ibarra et al 2015, see Publications).

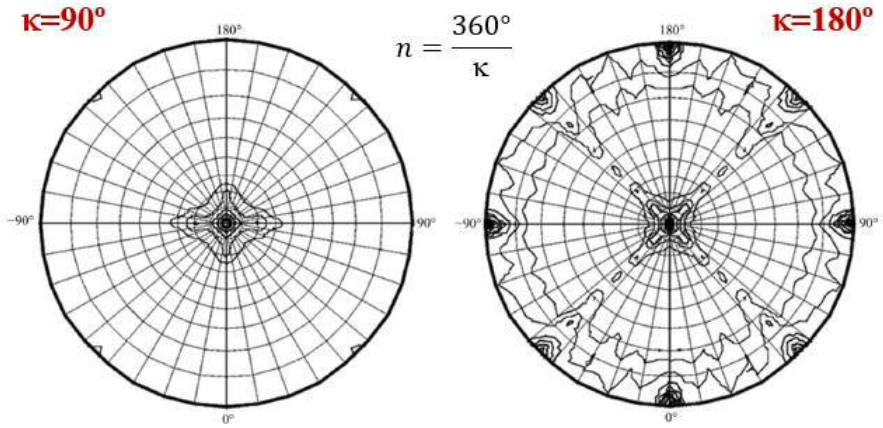


Fig 13. Self-rotation function of Gal-3[N VII-IX] crystals. These results had an integration radius of 30 Å and data was between 15 and 4 Å resolution. The crystallographic *a* axis is along $\varphi = 0º$, $\psi = 90º$ and the c* axis is perpendicular to the plane of the figure.

The structure was then solved by molecular replacement using Phaser (Adams et al 2010) and the coordinates of the human Gal-3 CRD (**PDBID: 1A3K**, Seetharaman et al 1998) as search probe. The initial rotational and translational searches identified eight monomers in the asymmetric unit with values of TFZ = 41.3 and LLG = 6485. This partial solution was then refined using phenix.refine (Adams et al 2010) along with manual building in Coot (Emsley et al 2010). In **Fig 14** is depicted the density

map (blue density) for the first round of molecular replacement for Gal-3[N VII-IX], where eight molecules accommodate (ribbon representation, each monomer in a different color), as well as strong positive electronic density areas (green density, white arrows), indicating the presence of more subunits. Moreover, in agreement with the Matthew's coefficient and the self-rotation function, it became clear that the structure had a non-crystallographic symmetry in which the basic unit in the crystal was a tetramer, which makes it possible then to obtain either an octamer or a dodecamer in the unit cell. With this information it was decided to run a new search and refinement for the data with a different model than **1A3K**.

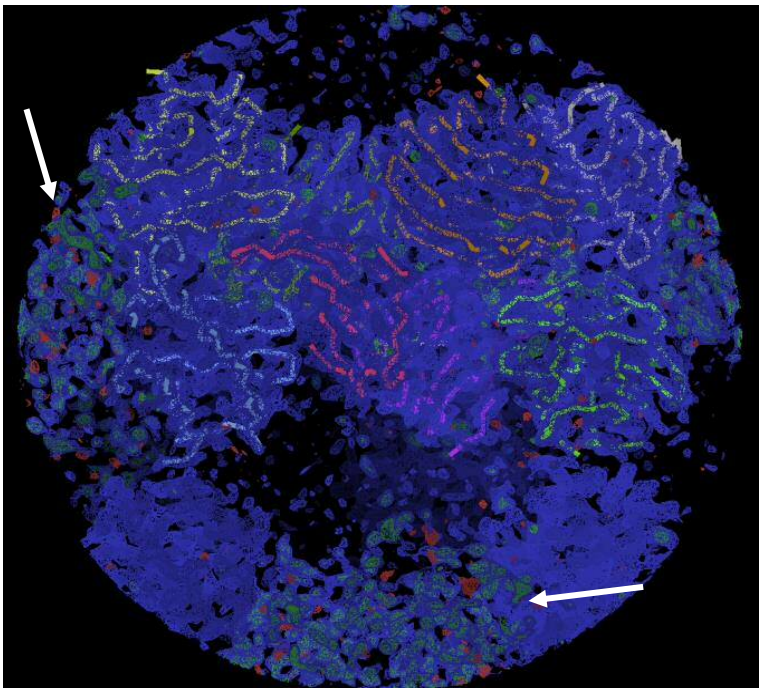


Fig 14. $2F_o$-F_c electron density map of Gal-3[N VII-IX] contoured at 1σ level. The first round of data refinement showed eight monomers in one unit cell and additional positive electron density (white arrows) suggesting the presence of more molecules.

Searches for the remaining units were performed using Phaser (Adams et al 2010) and the model of a single tetramer from the Gal-3[N VII-IX] data set as probe in a new query. The search resulted in the discovery of a third tetramer giving a total content of 12 molecules arranged as three tetrameric units (TFZ = 92.5, LLG = 16089) (**Fig 15**).

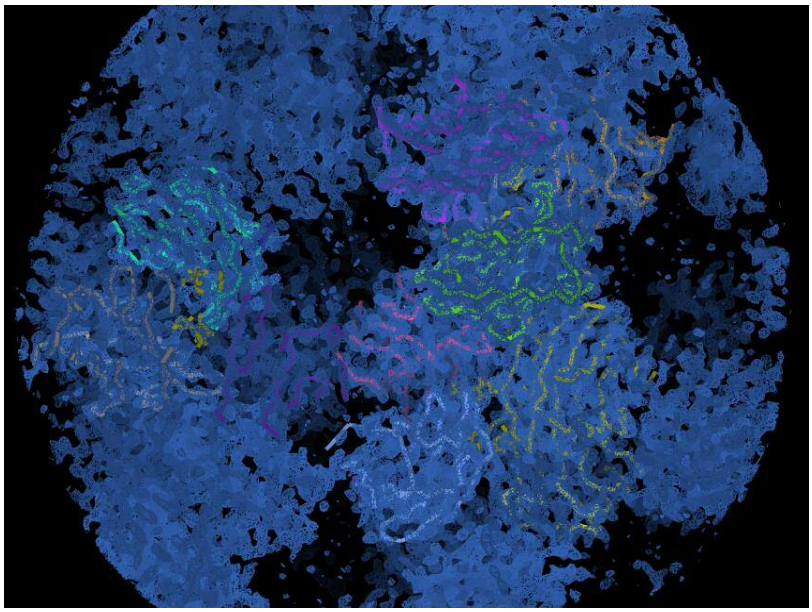


Fig 15. $2F_o$-F_c electron density map of Gal-3[N VII-IX] contoured at 1σ level. New rounds of manual building and refinement asserted the presence of twelve molecules in the unit cell.

Although at this resolution side chains are poorly defined, the characteristic *jelly-roll* folding of galectins was clearly observed (**Fig 16a**). Furthermore, the active site was

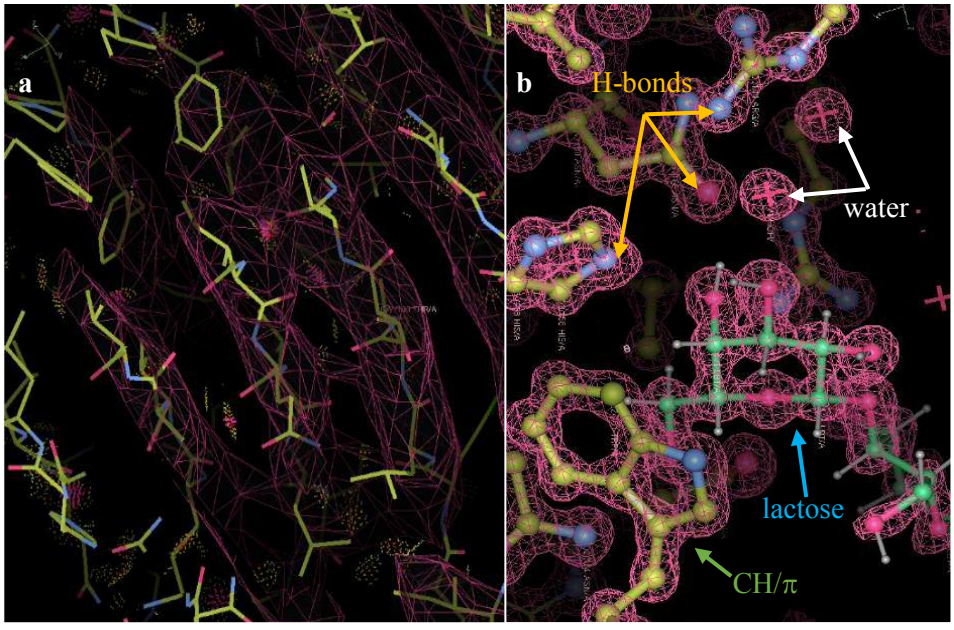


Fig 16. Crystal structure of Gal-3[N VII-IX]. The protein preserves (**a**) the *jelly-roll* folding characteristic of galectins, and (**b**) all the key contacts for β-galactoside binding.

clear enough to show the hydrogen bond contacts of lactose with **Arg162**, **His158** and **Asn160**, and the carbohydrate-π interactions with **Trp181**, as well as the interactions with waters (**Fig 16b**).

Following these preliminary results, optimization screening conditions were prepared to improve crystals and data quality. These crystals were also taken in a dry-shipping Dewar to the light source line XALOC (ALBA synchrotron, Cerdanyola del Vallès, Spain), and they diffracted to 2.2 Å resolution (**Fig 17**). Data were integrated with XDS (Kabsch 2010) and a local scaling was applied to reduce systematic errors with the CCP4 suite (see Materials and Methods). Details on the X-ray data collection are in **Table 1**.

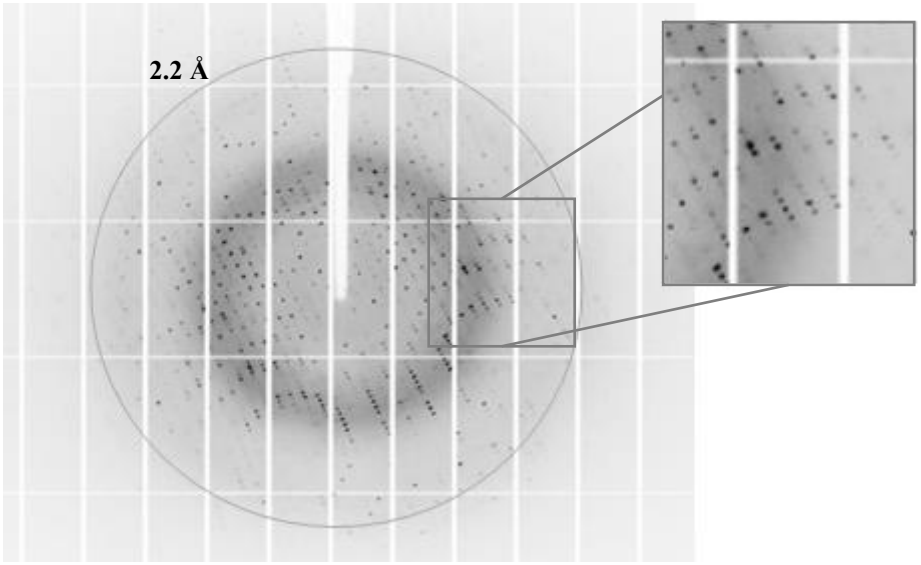


Fig 17. Diffraction pattern for the high resolution crystals. The X-ray diffraction image is at φ = 76º with a resolution ring at 2.2 Å.

The structure of this higher-resolution crystal was solved applying differential Fourier techniques in Phenix, taking the dodecamer model of Gal-3[N VII-IX] at low resolution as an initial model for refinement. With this model and after a first round of refinement with phenix.refine (Adams et al 2010), the R_{work} and R_{free} factors lowered to reasonable values of 0.263 and 0.314, respectively. A first look shows that every component of the dodecamer conserve the *jelly-roll* topology of two β-sheets made of antiparallel strands (**S1-S6**; **F1-F5**).

TABLE 1. X-RAY DATA COLLECTION FOR GAL-3[N VII-IX]

Beamline	BL13-XALOC (ALBA)
Wavelength (Å)	0.9794
Space group	$P2_12_12_1$
Unit cell parameters (Å)	a = 93.85, b = 98.19, c = 237.81
Resolution range (Å)*	49.10 – 2.20 (2.32 – 2.20)
No. of observations	1012467 (147079)
No. of unique reflections	112248 (16179)
Multiplicity	9.0 (9.1)
Completeness (%)	100 (100)
Mean $I/\sigma(I)$	11.7 (1.8)
Molecules per asymmetric unit [V_M ($Å^3Da^{-1}$)]	12 (2.15)
R_{merge}[a]	0.134 (1.233)
R_{meas}[b]	0.142 (1.308)
$CC_{1/2}$[c] (%)	99.7 (55.8)
Wilson B-factor	43.25

[a]$R_{merge} = S_{hkl}S_i\,|I_i(hkl) - \langle I(hkl)\rangle| \,/\, S_{hkl}S_iI_i(hkl)$, [b]$R_{meas} = \Sigma_{hkl}\,(N-1)^{-1/2}\,\Sigma_i|I_i(hkl) - \langle I(hkl)\rangle| \,/\, \Sigma_{hkl}\,\Sigma_i\,I_i(hkl)$, where $I_i(hkl)$ is the intensity measured for the *i*th reflection and $\langle I(hkl)\rangle$ is the average intensity of all reflections with indices *hkl*. [c]$CC_{1/2}$ is the correlation coefficient between two random half datasets (Karplus & Diederichs, 2012). *Values in parentheses are for the highest resolution shell.

Strong electron density corresponding to ordered parts of the N-terminal tail could only be detected in three of the twelve monomers of Gal-3[N VII-IX], most likely

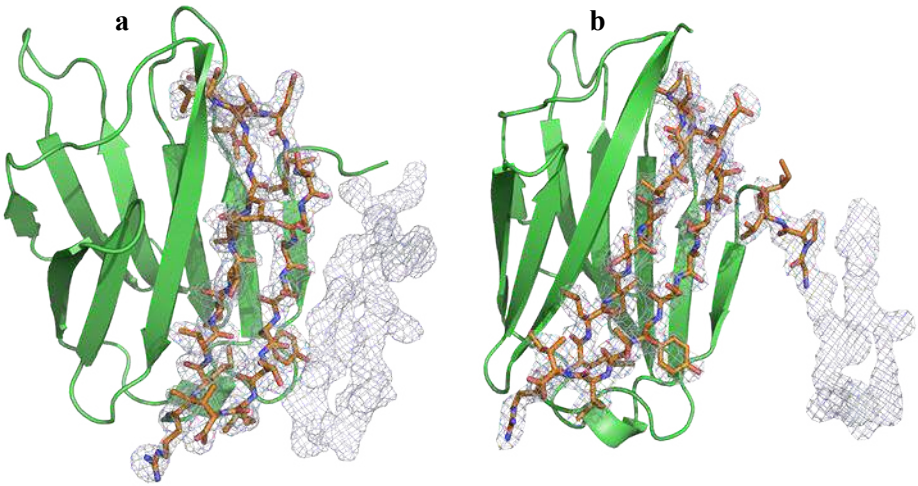


Fig 18. 2Fo-Fc electron density maps for two ordered segments of the N-terminal tail contoured at 1σ. (a) Specific characteristics of amino acids such as **His** and **Phe** allowed for their unequivocal identification in the electron density maps when building the N-LD section. **(b)** A different region was identified as part of the N-PG unit IX and residues preceding the CRD.

due to its flexible nature. For these three monomers, the electron density clearly indicated the presence of two different regions of the N-PG rather than two alternative conformations of the same segment (**Fig 18a/b**). To build a model in both electron density regions, the sequence available for the Gal-3[N VII-IX] construct (**Fig 19**) was used to accurately assign the sequence for identifying and fitting the side-chain amino acids.

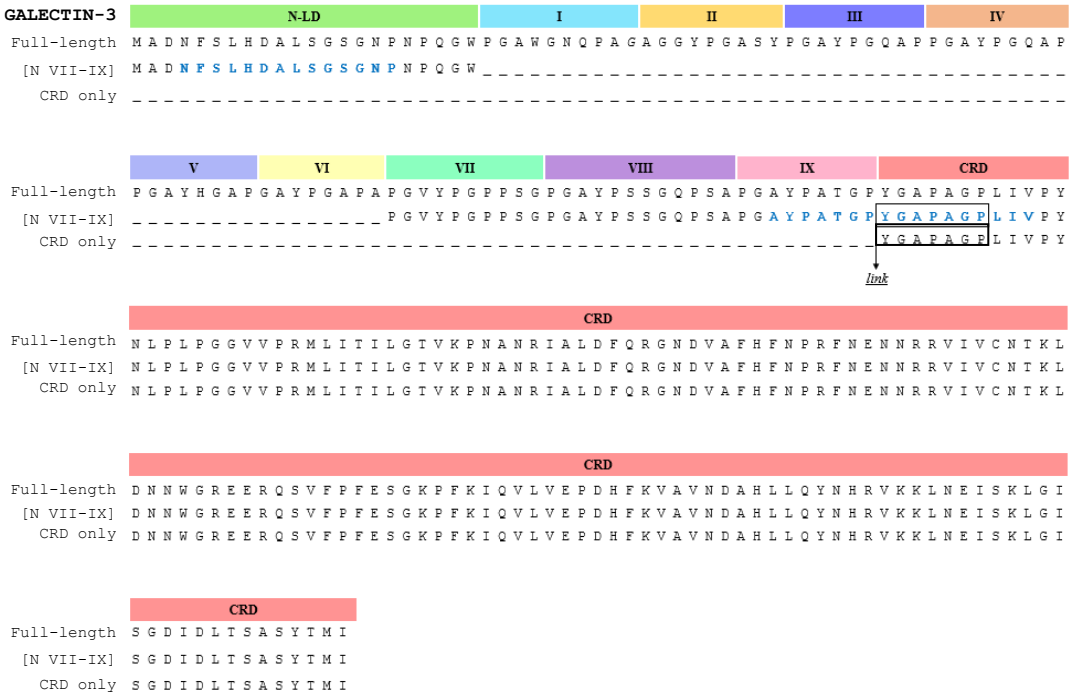


Fig 19. Sequences of Gal-3 full length (Gal-3$_{FL}$), Gal-3[N VII-IX] and Gal-3 CRD only. The electronic densities obtained from the X-ray diffraction data were compared with the available sequences on the proteins to figure out amino acid accommodation. The letters in light blue mark the N-PG amino acids built, **Asn4-Pro117** of the N-LD and **Ala100-Val116** of section IX, and the link between the N- and C-terminal (black square).

Once the dodecamer of Gal-3[N VII-IX] was solved, it was evident that, while many biological reports state that Gal-3 oligomerizes through the N-PG, the tetramers in this structure are facing each other's CRDs. The amino acid sequences were compared to the electronic densities obtained with the X-diffraction data, and several rounds of manual building in Coot (Emsley et al 2010) followed by refinement with phenix.refine (Adams et al 2010) allowed to unequivocally assign the two following distinct fragments of the N-terminal tail:

(A) in two monomers, residues **Asn4** to **Pro17** corresponding to the N-LD section can be modelled with high confidence due to specific sequence landmarks (**Fig 20**), like the imidazol ring of histidine in position 8 (**His8**) and the phenyl ring of phenylalanine in position 5 (**Phe5**). Notably, this region at the N-terminal forms a double-stranded antiparallel β-sheet stabilized by a hydrogen bonding network (**Table A**). The presence of the additional β-strands, called **F0** and **F-1**, extend the β-sheet on the Gal-3 β-sandwich face **F1-F5**, reducing the flexibility of the N-LD amino acids, which then may facilitate their availability for phosphorylation in key positions such as **Ser6** and **Ser12**.

Table A. Hydrogen bond interactions stabilizing the N-LD

Atom 1	*Atom 2*	*Distance (Å)*
O^{γ} Ser12	O^{γ} Thr246	2.74
O^{γ} Ser12	N Asp9	2.73
N Gly13	O Tyr247	3.05
O Gly13	N Tyr247	3.04
N Ser14	O Leu7	2.71
O Ser14	N Leu7	2.76
N Gly15	O Ala245	2.76
N Asn16	O Phe5	2.53
O^{γ} Ser6	$N^{\delta 2}$ His8	2.52

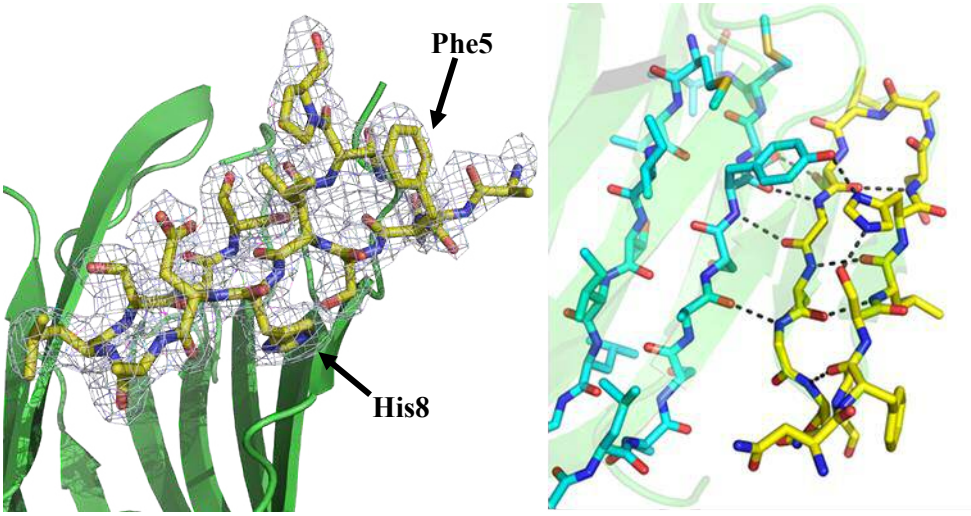


Fig 20. Building the N-LD region. The rings formed by **His8** and **Phe5** are indicated with black arrows (left). A specific set of contacts between the C-terminal region of the CRD and the N-LD residues reduces the flexibility of the newly constructed section (right).

Fig 21 shows a stereoscopic view of the newly constructed amino acids in the N-LD with residues of the C-terminal CRD stabilizing them.

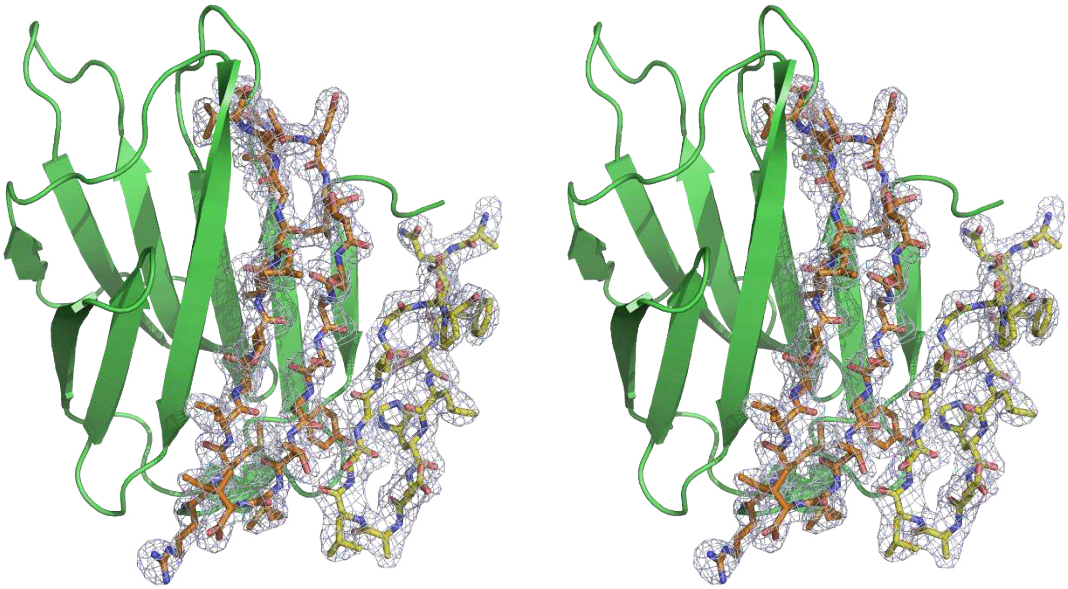


Fig 21. Stereoscopic view of the constructed amino acids Asn4-Pro17. With sticks representation, in orange: the amino acids belonging to the previously solved Gal-3 CRD; in yellow: the missing N-terminal **Asn4-Pro17** residues, forming a double-stranded antiparallel β-sheet on the F-face (namely **F0** and **F-1)** .

(B) on the other hand, a continuous electron density was observed near the CRD of one monomer (**Fig 18b**), which allowed to trace part of the N-PG unit IX and residues preceding the CRD (**Ala100** – **Val 116**) (**Fig 22**). As can be seen, this region displays a flexible random-coil-like conformation stabilized by hydrogen bonding between backbone atoms of sequencial residues, and additionally by the interaction of **Tyr107** with **His217** of a symmetry related neighbour (**Table B**).

Table B. Hydrogen bond interactions stabilizing the N-PG unit IX		
Amino acids in contact		*Distance (Å)*
N Ala111	O Ala103	3.0
O Ala111	N Ala103	2.76
O Tyr107	$N^{\delta 1}$ His217*	2.89

*symmetry related residue

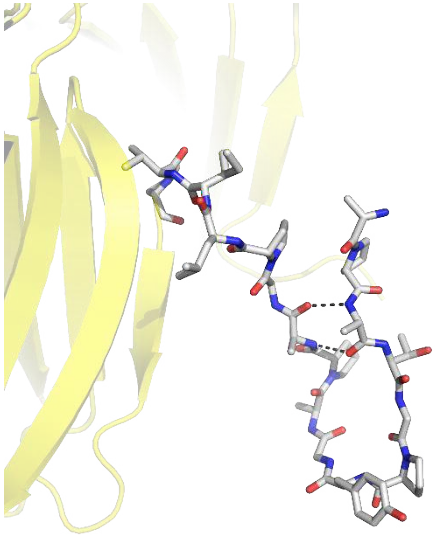


Fig 22. Close-up view around the N-PG region (Ala100-Val116). Two hydrogen bonds between **Ala103** and **Ala111** stabilize this region.

Fig 23 shows a stereoscopic view of the newly constructed amino acids from section IX and the link between the protein's N-LD and CRD.

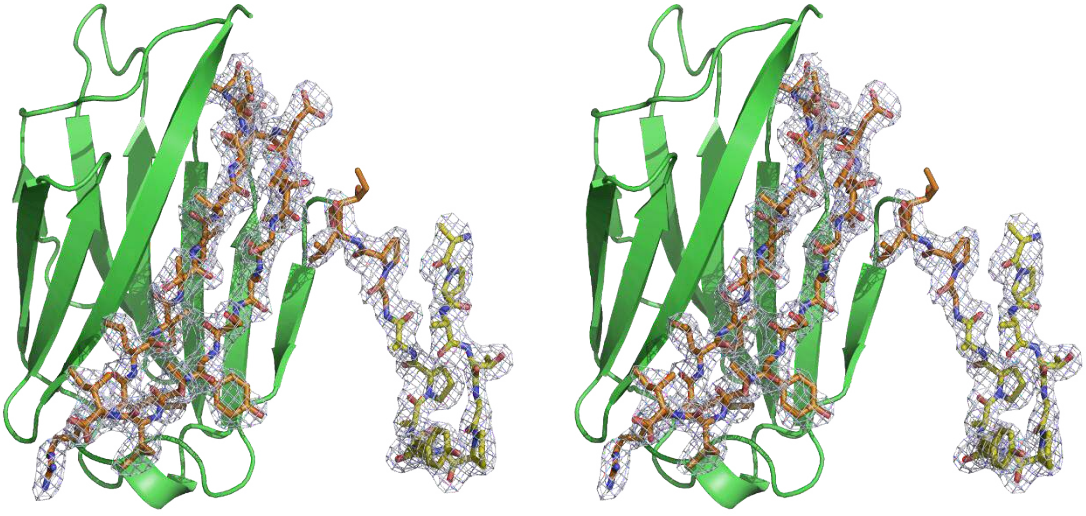


Fig 23. Stereoscopic view of the electron density map around the N-PG region. With sticks representation, in orange: the amino acids belonging to the previously solved Gal-3 CRD; in yellow: region corresponding to the residues joining the N-terminal CRD and those of section IX of the N-PG.

Finally, water and lactose molecules were placed using Coot (Emsley et al 2010) and refinement of the final model had values of R_{work} = 0.243 and R_{free} = 0.284. A summary of the final refinement parameters is given in **Table 2**. Validation of the final model was done with the Molprobity software, 96.4% of residues were located in favored regions of the Ramachandran plot (**Fig 24**).

TABLE 2. REFINEMENT STATISTICS FOR GAL-3[N VII-IX]

Refinement	
Resolution range (Å)	49.50 – 2.20 (2.26 – 2.20)
Working reflections	106537 (7797)
Testing reflections	5607 (413)
Completeness (%)	99.97
R_{work}	0.213 (0.342)
R_{free}	0.264 (0.350)
No. of non-H atoms	
Protein	13566
Lactose	276
Sulfate ions	120
Water	128
Average B factors (Å^2)	
Protein	47.15
Lactose	31.14
Sulfate ions	56.26
Water	34.91
R.m.s. deviations	
Bond lengths (Å)	0.014
Bond angles (°)	1.83
Ramachandran plot statistics	
Favoured (%)	96.4
Allowed (%)	3.4
Outliers (%)	0.2

Values in parentheses are for the highest resolution shell.

With all this information, significant progress has been made in the structural resolution of human Gal-3 showing images for two different regions of the N-PG tail. However, reconstruction of the full-length Gal-3 and the truncated Gal-3[N VII-IX] need other biophysical techniques, like SAXS, to provide a more complete picture.

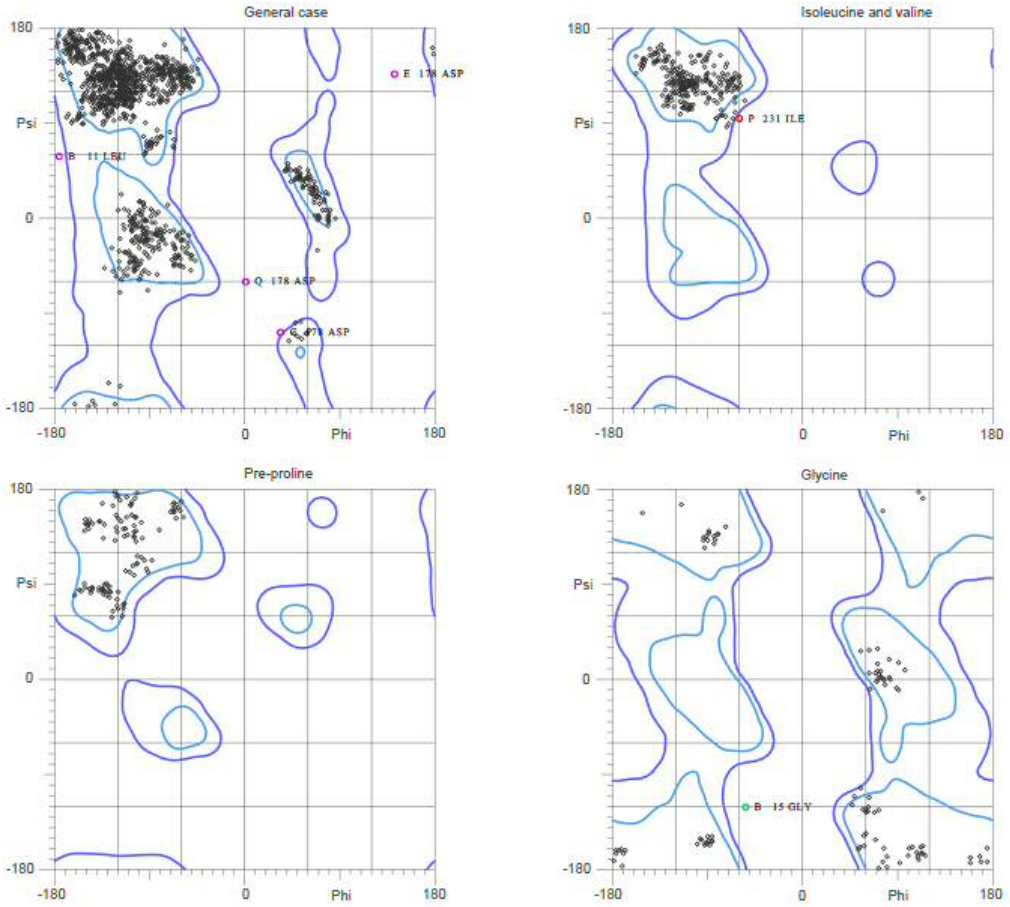


Fig 24. Ramachandran plot for Gal-3[N VII-IX]
Asp 178 located in a flexible loop is in a disallowed region. All the other residues are in the favored and allowed regions (>98%).

Small-Angle X-ray Scattering

Small-angle X-ray Scattering of Gal-3[N VII-IX] and Gal-3_{FL} were measured using synchrotron radiation at the ESRF facilities (Grenoble, France). Solutions of purified galectins were prepared at concentrations between 3 and 10 $mg.ml^{-1}$ for Gal-3[N VII-IX] and from 2 to 8 $mg.ml^{-1}$ for the full-length. To prevent freezing damage, the samples were transported at 277 K to the ESRF. Still, the samples' freezing process had a negative effect on protein stability leading eventually to protein precipitation, particularly at higher concentrations. However, the data collection possessed an appropriate quality and indicated that Gal-3[N VII-IX] and Gal-3_{FL} are monomeric over the concentration range tested.

Curve fitting of the experimental data for Gal-3[N VII-IX] (**Fig 25a/b**) and for Gal-3_{FL} (**Fig 26a/b**) was used to build the corresponding bead models (**Fig 25c** and **Fig 26c** for Gal-3[N VII-IX] and Gal-3_{FL}, respectively) using the ATSAS software package (Svergun 1992). The calculated molecular weights were 20.98 kDa for Gal-3[N VII-IX] and 31.54 kDa for Gal-3_{FL}, in good agreement with the theoretical values. R_g and D_{max} provide secure information on the nature of the interactions between domains and of the total protein size (Putnam et al 2007). Thus, the three-dimensional shapes obtained for Gal-3[N VII-IX] and Gal-3_{FL} corresponded to elongated particles, confirming the pair-distance distribution functions (**Fig 25b** and **Fig 26b**).

The final structures modelled for Gal-3[N VII-IX] and Gal-3_{FL} were compared with one another and with the crystal structure of the solved Gal-3 CRD with PyMol (DeLano 2012) in order to observe the distinctive characteristics of all the structures of the same protein.

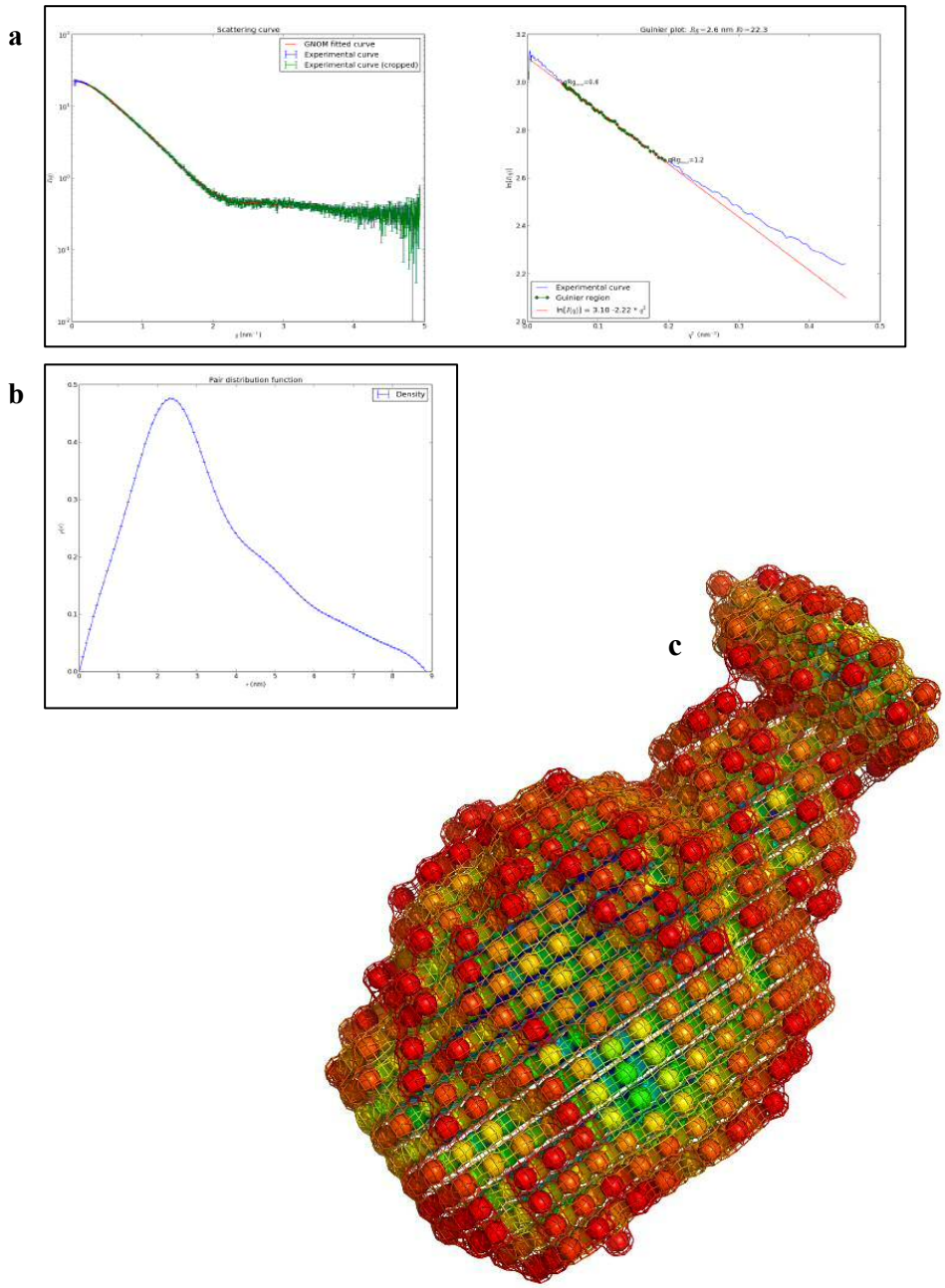


Fig 25. Small-Angle X-ray scattering (**a**) GNOM curve fitting for a 6 mg.ml^{-1} Gal-3[N VII-IX] solution, (**b**) Pair-distance distribution function for Gal-3[N VII-IX], (**c**) bead model representation calculated using the ATSAS software package.

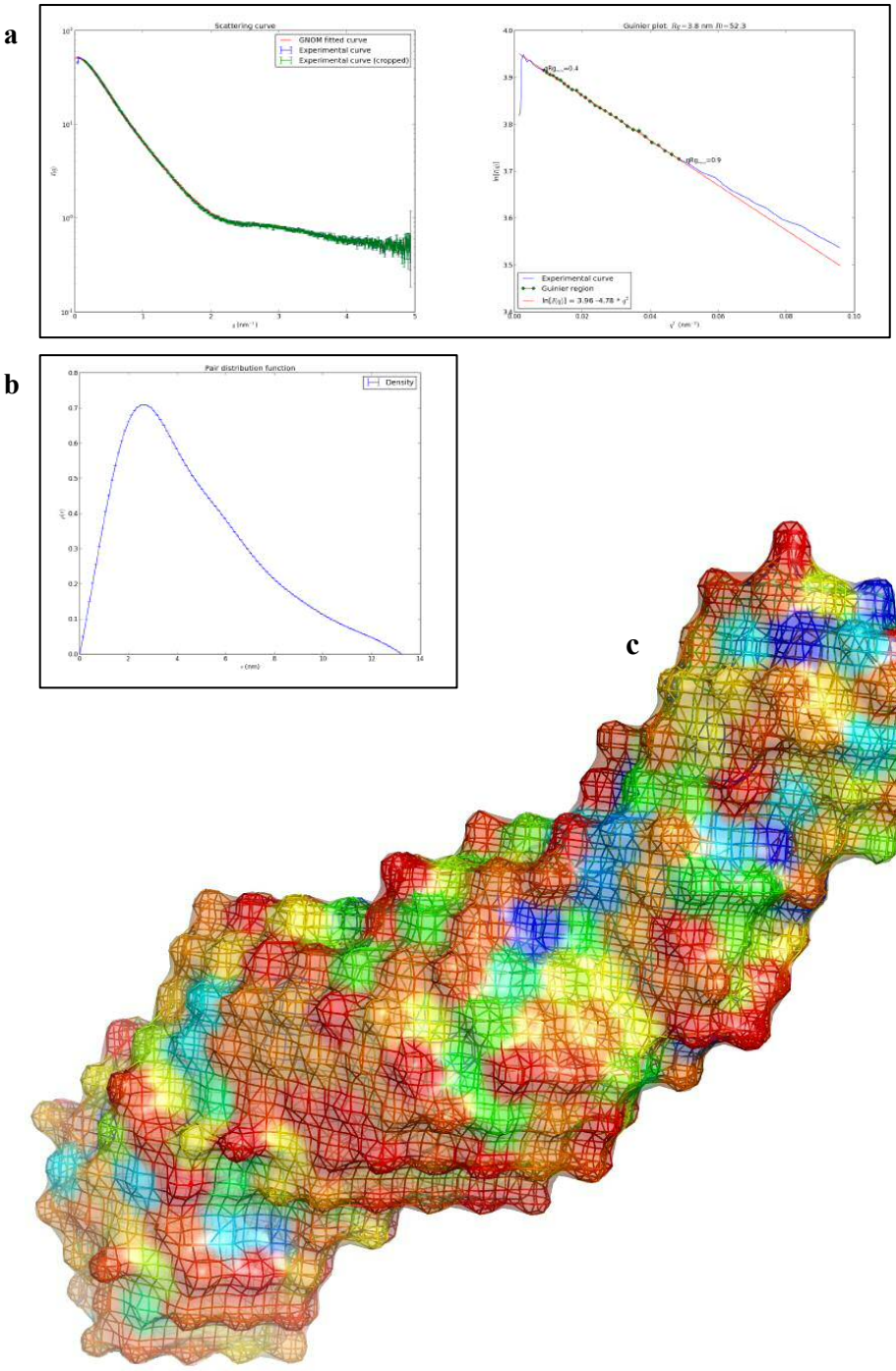


Fig 26. Small-Angle X-ray scattering (**a**) GNOM curve fitting for a 2 mg.ml^{-1} Gal-3$_{FL}$ solution, (**b**) Pair-distance distribution function for Gal-3$_{FL}$, (**c**) bead model representation calculated using the ATSAS software package.

***E. coli* expression and protein purification**

C-GRP-C synthetic cDNA (ATG: biosynthetics, Merzhausen, Germany) with primers **GRP*NdeI*** and **GRP*XhoI*** designed from the nucleotide sequence (Sigma-Aldrich, Missouri, USA) was amplified by PCR. As shown in **Fig 27**, a PCR product of the expected size (426 bp) appeared at the Mg^{2+} recommended concentration provided by the manufacturer (**Fig 27a**). One step PCR subcloning allowed the insertion of the C-GRP-C gene into the bacterial plasmid expression His-tagged vector pET28-PP between the *NdeI* and *XhoI* restriction sites (see Materials and Methods).

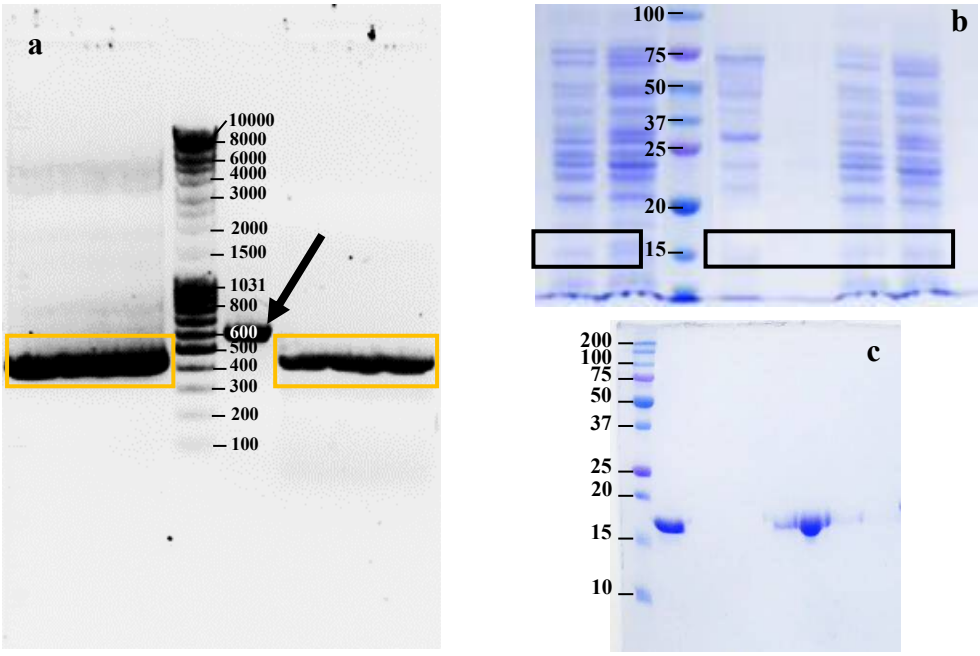


Fig 27. Gene and protein expression of C-GRP-C. (**a**) PCR amplification of C-GRP-C was tested in an agarose gel (yellow boxes), the PCR positive control is marked with a black arrow. (**b**) Protein presence in the *E. coli* lysates was checked by SDS-PAGE. A protein of ~16 kDa M.W. was spotted (black boxes). **Protein purification of C-GRP-C.** (**c**) After purification, the presence of the protein was checked by SDS-PAGE.

E. coli BL21(DE-3) strain was used for large-scale expression of the protein. Cell lysates were run on a 12% SDS-PAGE, which showed that C-GRP-C (~16 kDa M.W., **Fig 27b** black boxes) was among the other bacterial proteins. The soluble

fraction was loaded onto a nickel charged HiTrap chelating column followed by incubation and dialysis with 3C-protease to remove the His tag. The concentrated fractions' purity is shown in **Fig 27c**.

Further purification was carried out through a size exclusion chromatography step. The chromatogram (**Fig 28**) presents two differentiated peaks, which could be an indication of the presence of C-GRP-C oligomers in solution. Both peaks were verifyied for purity through an SDS-PAGE (**Fig 28, inset**). Each fraction was tested individually, lanes **[9]** to **[11]** and lanes **[17]** to **[20]**.

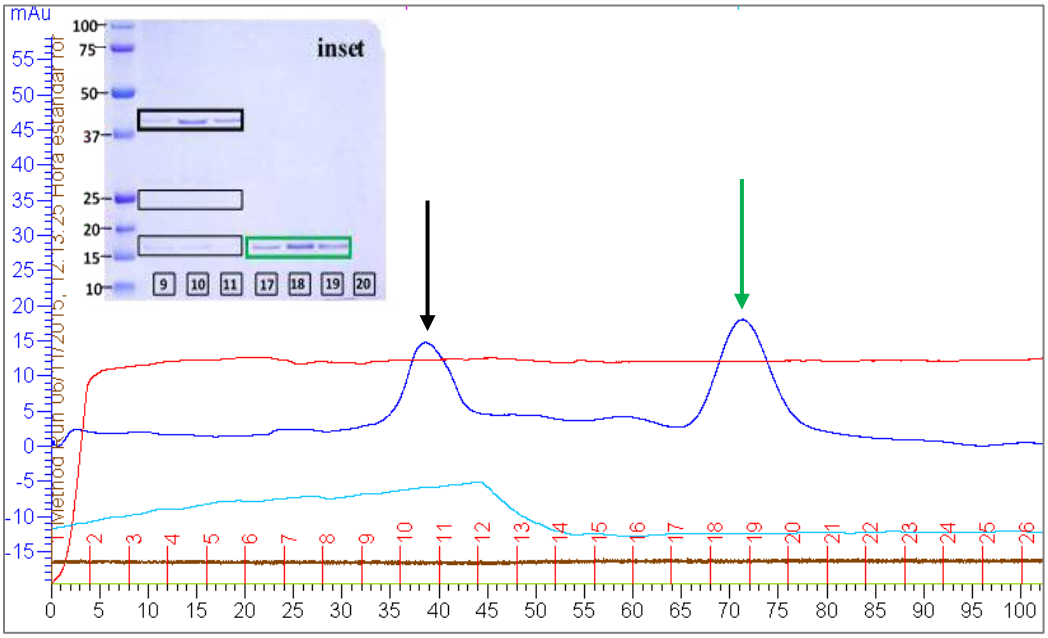


Fig 28. Purification of C-GRP-C by size exclusion chromatography. The elution profile showed two differentiated peaks. The inset shows the analysis by SDS-PAGE of the corresponding fractions for both peaks.

In **Fig 28**, the first elution peak (black arrow) shows a mixture of three different molecular species (black boxes). On the other hand, the second elution peak (green arrow) shows a very pure C-GRP-C with the expected molecular mass (green box). These fractions were treated separately and used for crystallization, SAXS and analytical ultracentrifugation experiments.

X-ray diffraction resolution

Fig 29. C-GRP-C crystal. In this image, a single crystal of the protein is located in the lower left corner.

For C-GRP, initial attempts to crystallize it were not successful. It was clear that a well ordered structure of the N-terminal could not be attained to enable crystal formation. Therefore, as previously performed in His-tagged human GRP for its crystallization (Zhou et al 2008), a 36 amino acid sequence portion was deleted to produce a shortened form. Diffracting C-GRP-C crystals grew after two weeks (**Fig 29**). A high-resolution data set was collected in the XALOC line at the ALBA synchrotron. The crystals diffracted to a resolution of 1.55 Å (**Fig 30**), and the crystals belonged to the body-centered orthorhombic space group $I2_12_12_1$ with unit cell parameters a = 38.6, b = 106.9 and c = 114.1 Å. Analysis of the crystal solvent content indicated one monomer in the asymmetric unit, with a Matthews coefficient of 3.57 $Å^3Da^{-1}$ (corresponding to a solvent content of 65.52%). The crystal structure was determined by molecular replacement using the coordinates of the human homologue (**PDBID: 3BC9**; Zhou et al 2008). The final model

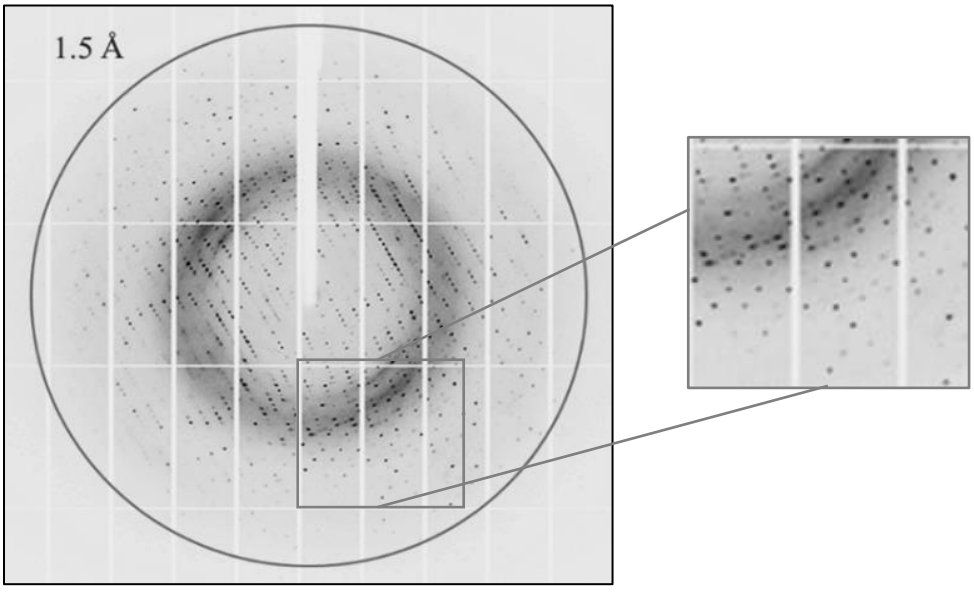


Fig 30. X-ray diffraction data of C-GRP-C. The X-ray diffraction image is at φ = 176º with a resolution ring at 1.5 Å.

includes 134 amino acid residues, 2 sulphate ions, 3 polyethylene glycol, 3 ethylene glycol and 104 water molecules in the asymmetric unit (**Fig 31**). The C-GRP-C monomer consists of a single β-sandwich domain composed of two antiparallel five-stranded (**F1-F5**) and six-stranded (**S1-S6**) β-sheets which is characteristic of the *jelly-roll* topology shared by all known galectins (Gabius et al 2011; Solís et al 2015).

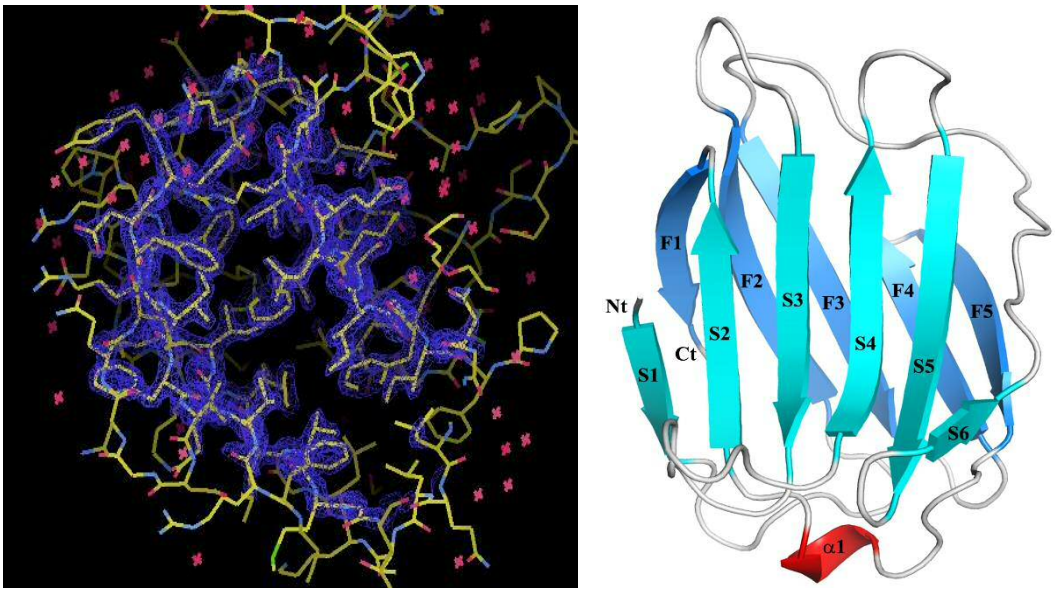


Fig 31. Crystal structure of C-GRP-C. The protein preserves the *jelly-roll* folding as expected by its sequence, but does not possess the ability to bind β-galactosides. The strands of the two β-sheets are labelled as well as a short 3_{10} helix placed between strands F5 and S2.

In addition to the sequence similarities between galectins and GRP, significant dissimilarities can be observed at the residues that form the architecture of the carbohydrate-binding site in galectins. The stereochemical quality of the model, with unambiguous electron density, allowed a clear tracing of this region (**Fig 32**). This concave surface of the β-sheet, comprising residues from β-strands S4-S6, characterizes the recognition site for β-galactosides in galectins. Looking at the sequence signature for operative binding in canonical *G. gallus* galectins (**His45**, **Asn47**, **Arg49**, **Asn58, Trp65**, **Glu68**, and **Arg70** for CG-2) is turned into **Glu86**, **Lys88**, **Val90**, **Asn99, Trp106**, **Glu109** and **Ser111** in C-GRP-C. This shows that only three of seven positions of the highly conserved binding residues are maintained: **Asn99**, **Trp106** and **Glu109**. In particular, the **Trp106** is present in avian GRP but not in human GRP, and is the responsible for the carbohydrate-π interactions in galectins' CRD.

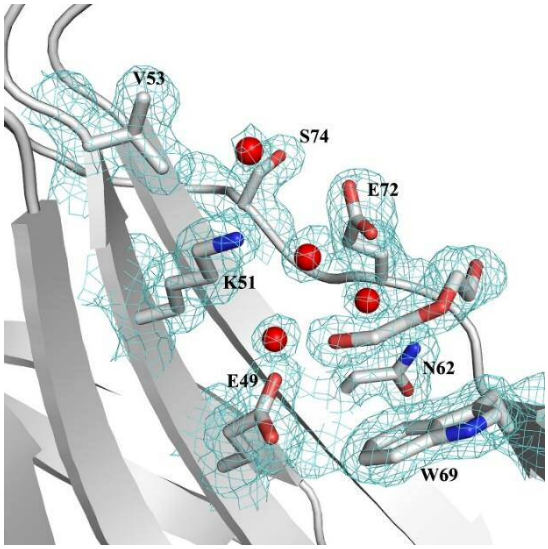


Fig 32. Close-up view of the putative contact site for a β-galactoside in C-GRP-C. Respective residues in C-GRP-C as well as an ethylene glycol molecule are shown in ball-and-stick mode. Water molecules are depicted as red spheres. The observed electron density map 2Fo-Fc is contoured at 1.0 σ.

The final model had a good geometry with 99.2 % of the residues located in the most favored regions of the Ramachandran plot (**Fig 33**). Summary statistics for the C-GRP-C geometry obtained for the final model are shown in **Table 3**.

TABLE 3. DATA COLLECTION AND REFINEMENT STATISTICS FOR C-GRP-C

Data collection	
Beamline	BL13-XALOC (ALBA)
Wavelength (Å)	0.9794
Space group	$I2_12_12_1$
Unit cell parameters (Å)	a = 38.6, b = 106.9, c = 114.1
Resolution range (Å)	57.04 - 1.55 (1.63 - 1.55)
No. of observations	300868 (40181)
No. of unique reflections	34844 (4882)
Multiplicity	8.6 (8.2)
Completeness (%)	99.7 (97.6)
Mean I/s(I)	12.6 (2.2)
Molecules per asymmetric unit [V_M ($Å^3Da^{-1}$)]	1 molecule (V_M= 3.91)
R_{merge}[a]	0.069 (0.977)
R_{meas}[b]	0.074 (1.043)
$CC_{1/2}$[c]	99.9 %
Mosaicity	0.20
Wilson B-factor	21.62
Refinement	
R_{work}/R_{free}	0.168 (0.267) / 0.189 (0.30)
Working reflections	33160 (2285)
Testing reflections	1593 (101)
Protein atoms (non H)	1064
PEG and ethylene glycol	36
Sulfate ions	10
Water molecules	102
Mean B factors (Å²)	
Protein	27.23
PEG and ethylene glycol	57.45
Sulfate ions	58.01
Water molecules	42.42
rmsd bond lengths (Å)	0.008
rmsd angles (°)	1.36
Ramachandran plot statistics	
Favored (%)	99.2
Outliers (%)	0
PDB accession ID	5IT6

[a]$R_{merge} = S_{hkl}S_i\ |I_i(hkl) - <I(hkl)>| / S_{hkl}S_iI_i(hkl)$, [b]$R_{meas} = \Sigma_{hkl}\ (N-1)^{-1/2}\ \Sigma_i|I_i(hkl) - <I(hkl)>| / \Sigma_{hkl}\ \Sigma_i\ I_i(hkl)$, where $I_i(hkl)$ is the intensity measured for the *i*th reflection and $(I(hkl))$ is the average intensity of all reflections with indices *hkl*. [c]$CC_{1/2}$ is the correlation coefficient between two random half datasets (Karplus & Diederichs, 2012). Values in parentheses are for the highest resolution shell

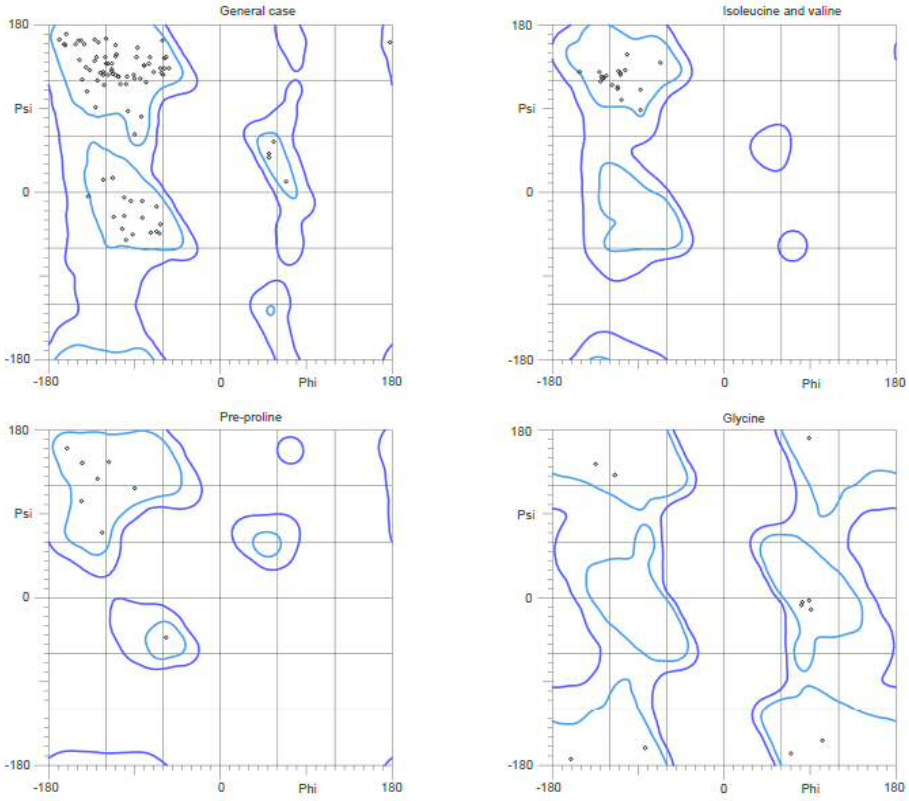


Fig 33. Ramachandran plot for C-GRP-C
All residues are located in the favored and allowed regions.

Small-Angle X-ray Scattering

Small-angle X-ray Scattering of C-GRP-C was measured using synchrotron radiation at the ESRF facilities (Grenoble, France). Solutions of purified C-GRP-C used in the SAXS measurements were prepared at concentrations in the range between 2 and 8 mg·ml^{-1}. To prevent freezing damage, the samples were transported at 277 K to the ESRF. The estimated radius of gyration using the Guinier approximation was R_g = 20 Å. There was no noticeable dependence of R_g on protein concentration. Curve fitting of the experimental data (**Fig 34a**) and final bead model (**Fig 34b**) was made with the ATSAS software package (Petoukhov et al 2012). The calculated molecular weight of C-GRP-C was 16.2 kDa, so this speaks of the behavior of C-GRP-C as a monomer in solution even at 8.0 mg·ml^{-1}. The value of R_g obtained from the GNOM fit was 18.3 ± 0.7 Å, in good agreement with the value obtained by the Guinier approximation, and a D_{max} = 54 Å. An additional experiment, to prove the protein is monomeric in solution, was provided by AUC.

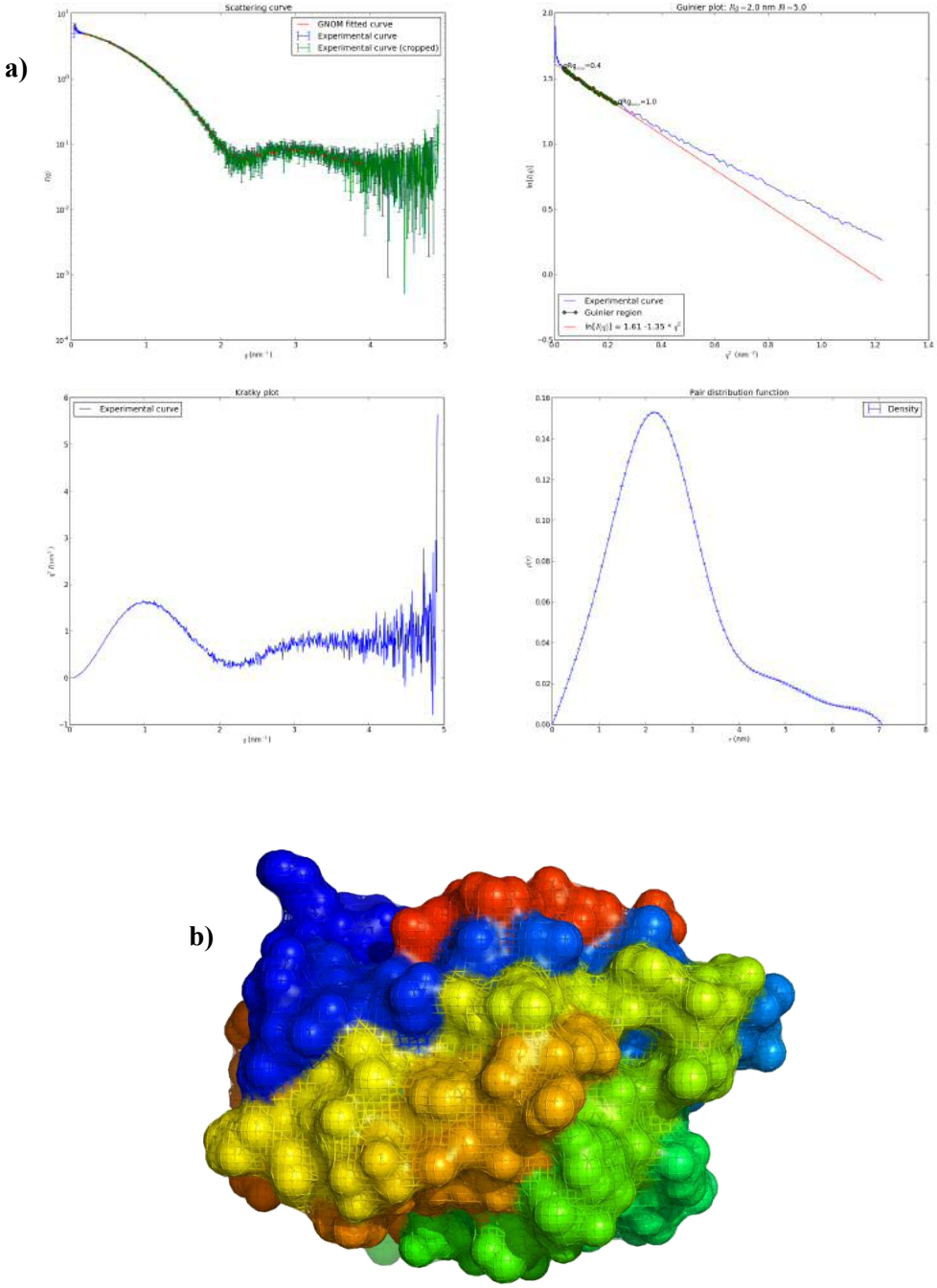


Fig 34. Small-Angle X-ray Scattering of C-GRP-C. The SAXS values for this protein account for the presence of a monomer in solution. (**a**) Experimental data (**b**) final bead model obtained from the data using the ATSAS software package.

Analytical ultracentrifugation

Analytical ultracentrifugation was run using C-GRP-C samples at concentrations of 0.2, 1 and 2 mg·ml^{-1}. Differential sedimentation coefficients were calculated by least squares boundary modeling of the experimental data using the *c(s)* method. C-GRP-C was monomeric over the full concentration range tested. The calculated sedimentation coefficient was 1.161 ± 0.006 S with a frictional ratio of 1.30 (**Fig 35a**). Afterwards, it was decided to test the full-length protein (C-GRP) in order to assess whether the non-CRD (36 amino acids deletion) portion could participate in the protein's oligomerization in solution, as it happens with Gal-3. As for the shortened version, the calculated sedimentation coefficient was similar, with a value of 1.137 ± 0.008 S with a frictional ratio of 1.52 (**Fig 35b**).

Thus, the differential coefficient distribution *c(s)* curves suggest that both C-GRP-C and C-GRP are monomeric in solution, with apparent molecular weights of 16.5 kDa and 19 kDa, respectively. No oligomer formation was seen under all conditions tested.

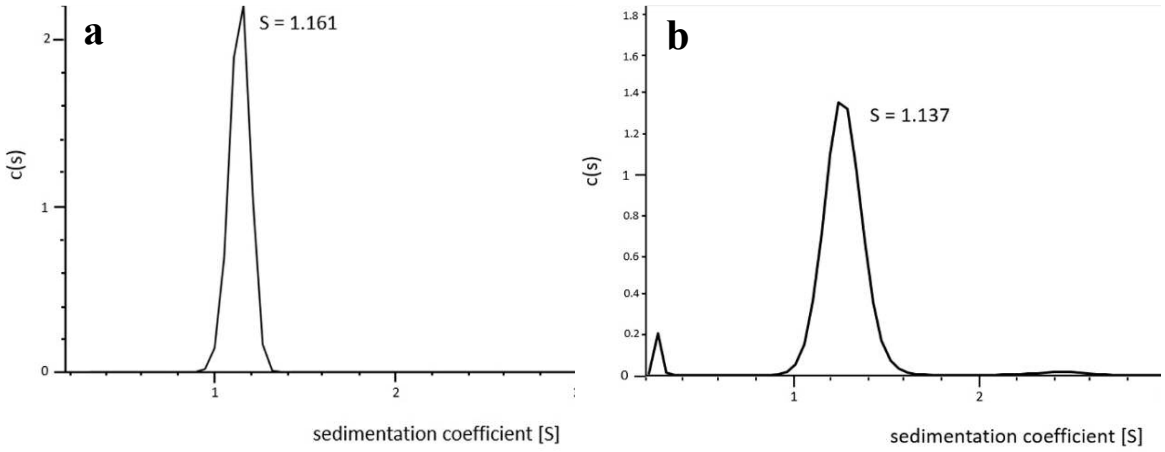


Fig 35. Analytical ultracentrifugation C-GRP-C and C-GRP. In the experiments of sedimentation velocity at 2.0 mg.ml-1, both (**a**) C-GRP-C and (**b**) the full-length protein (C-GRP), are monomeric in solution, with no sign of oligomer formation throughout the entire experiment.

DISCUSSION

DISCUSSION

Galectins are β-galactoside-binding proteins involved in the regulation of several cellular processes, both physiological and pathological. Galectins have been identified as important targets in cancer, angiogenesis and tumor metastasis, both as markers and as inhibitable molecules in the development and progression of cancer (Thijssen et al 2013). Because they promote at the same time tumor cells survival and neutrophile apoptosis in a location-dependent manner, galectin regulators, as a therapy strategy, will importantly contribute to other treatments in cancer such as VEGFR inhibition (Rabinovich et al 2014; Compagno et al 2014). Moreover, besides their participation in cancer, galectins mediate, in a carbohydrate-dependent manner, bacterial adhesion to host cells and inflammatory processes, *e.g.* rheumatoid arthritis, preeclampsia and other obstetric syndromes, diabetes, etc. (Vasta et al 2012; Than et al 2012; Thiemann and Baum 2016).

GAL-3[N VII-IX] AND GAL-3FL

A characteristic of some galectin types is that they are tissue-specific, like Gal-7 and -10, while others are ubiquitous like Gal-1, -3 and -9 (Cooper et al 2002). Therefore, the latter are the most studied among the galectin family members. Both full structures of Gal-1 and -9 have been solved (López-Lucendo et al 2004; Solís et al 2010), as well as a significant number of structures of the Gal-3 CRD domain (Seetharaman et al 1998).

However, the lack of a structure for Gal-3 poses a limitation for the development of pharmaceutical effectors for clinical aplication against the aforementioned diseases. Indeed, being the only chimera type galectin, Gal-3 has eluded full structural resolution since the first time the CRD domain of this protein was described nineteen years ago. While Gal-3 has been widely reported in cancer biology studies to execute its functions by multimerization (Rabinovich et al 2002a; Thiemann and Baum 2016), the only available crystal structures of the Gal-3 CRD do not show how this protein interacts to form such multimers. Moreover, while the CRD alone has an overall globular shape, it is unknown what shape does Gal-3 adopt with the elongated N-terminal flexible domain.

In this work, the structural resolution of Gal-3 was undertaken from a protein engineering approach. Several constructs of distinct Gal-3 deletion variants were obtained by recombinant techniques with different degrees of tail truncation, mimick-

ing the naturally occurring products by cleavage of the N-terminal poly-proline/glycine (**N-PG**) tail with matrix metalloproteinases (**MMPs**) (Kopitz et al 2014). The relationship between Gal-3 and MMPs is important on both sides, as well as for many other extracellular proteins involved in ECM remodeling, angiogenesis, invasion and metastasis. For instance, shortening the N-PG with MMP-2 and -9 renders Gal-3 unable to activate neutrophiles and reduces its angiogenic potency (Funasaka et al 2014a; Machado et al 2014). On the other hand, MMP-9 deficiency leads to Gal-3 accumulation in hypertrophic chondrocytes, impeding normal bone formation. In return, extracellular Gal-3 regulates MMP-1 and -9 expression at the transcriptional level, and induces MMP-9 secretion in a metastatic melanoma cell line (Gao, Liu et al 2017).

A particular construct, consisting of three of the nine PG repeats (seven to nine) plus the N-terminal lead peptide (**N-LD**) and the CRD, namely Gal-3[N VII-IX], crystallized with high reproducibility, and diffracted to 2.2 Å (see Results). The [N VII-IX] elongation in this Gal-3 construct changed the way the protein molecules build up, forming tetramers in the crystal lattice.

The Gal-3[N VII-IX] crystal structure showed new structural features, giving a complete picture of the multimerization mode. According to Gao, Liu et al (2017), protease-mediated Gal-3 cleavage generates both the intact CRD and with N-terminal peptides of varying lengths that retain lectin binding activity but lose multivalence in every instance that the N-PG is incomplete. In apparent contradiction to this, Gal-3[N VII-IX] crystal presented twelve molecules in the *unit cell*, specifically in the form of three tetramers (**Fig 36**). It stands out that oligomerization is mediated by interactions among subunits through the C-terminal end, being the three tetramers facing each other's CRD. This type of oligomerization in Gal-3 *via* self-association of the CRD is called a *C-type self-association* (Gao, Liu et al 2017).

As it is depicted in **Fig 36** (tetramers marked using different colors), for each tetramer the lactose ligands bound to the carbohydrate-binding sites are in close proximity to each other, while the N-terminals (drawn in red) are on the opposite side. In two monomers, the modelled N-terminal corresponding to residues **Asn4-Pro17** has a β-sheet structure (**Fig 36**, red in yellow) conforming two β-strands, **F0** and **F-1**, stabilized in its place by a specific set of contacts between amino acids of the N-LD and the **F1** strand of the CRD. In a third monomer, a few residues of the Pro/Gly-rich repeating unit IX and residues preceding the CRD, **Ala100-Val116**, were built

with a flexible random-coil-like conformation (**Fig 36**, red in blue) stabilized by a hydrogen bond network (see Results).

C-type self-association of Gal-3 has been shown in FRET, using Gal-3 labeled at the C-terminal (Nieminem et al 2007), in NMR experiments (Ippel et al 2016), and when Gal-3 is incubated with asialofetuin as well as with multivalent glycoconjugates located in the same cell membrane (Lepur et al 2012; Vasta et al 2012; Gao, Liu et al 2017).

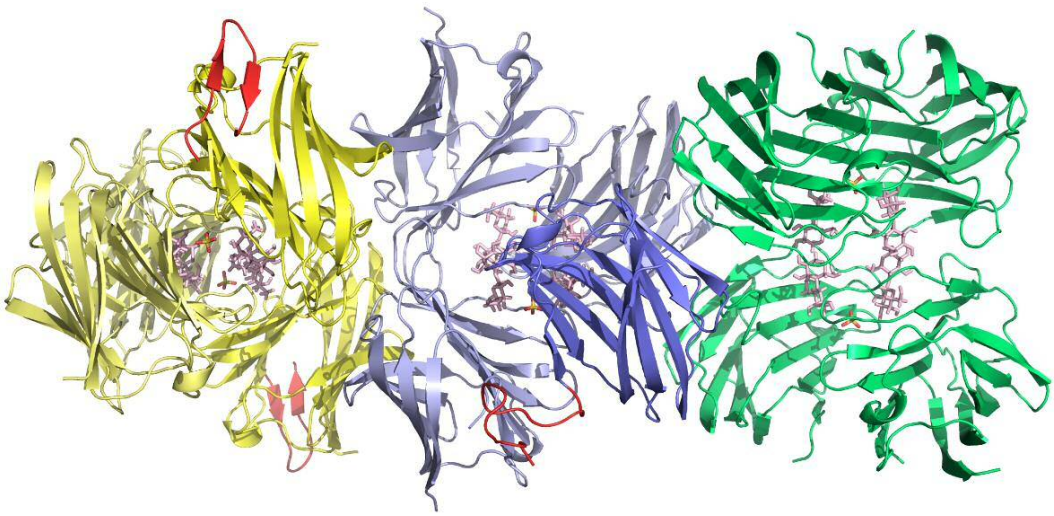


Fig 36. Crystal packing of Gal-3[N VII-IX] within the asymmetric unit. The ribbon representation shows three tetramers with the characteristic *jelly-roll* topology, with the CRDs facing each other. Bound lactose molecules at the CRD are depicted in stick representation (pink). The constructed sections of the [N VII-IX] are drawn in red.

Considering this, it is very likely that the C-type self-association observed in the Gal-3[N VII-IX] crystal structure may be at least partially mediated by the presence of close contacts between the lactose ligands (**Fig 37**, purple arrow), that imitate a cluster in which Gal-3 oligomerizes in presence of its natural ligands. However, in every instance, the N-PG has to be present for oligomerization to occur. This may also explain the differing results with other experiments (Gao, Liu et al 2017) in which truncated forms of Gal-3 with the N-PG do not oligomerize in absence of ligands. In **Fig 37** (black arrow) can also be seen two sulfate molecules, apparently stabilizing the tetramer by replacing water molecules (Seetharaman et al 1998), a feature that is most likely a packing artifact since computational studies show how accommodation of the sulfate group will be disfavored by steric hindrance.

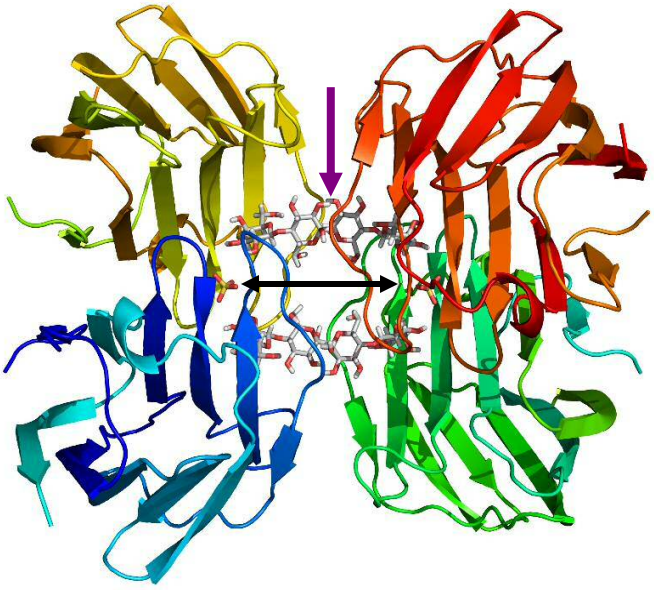


Fig 37. Features in the crystal structure of Gal-3[N VII-IX]. Lactose ligands at the CRDs are arranged in a network with very close contacts, stabilized by the presence of sulfate ions.

A comparison of the monomers, and particularly of the protein surfaces at the N-PG extension (**Fig 38**), allows to clearly identify the start and end points of the amino acid sequence for the constructed N-PG with, unexpectedly, a well-defined structure. The last Pro/Gly-rich repeating unit is located on the opposite side of the CRD. Interestingly, the double-stranded antiparallel β-sheet formed by residues **Asn4-Pro17** (**Fig 38, left**) extends the antiparallel β-sheet on the F-face, thus reducing flexibility of the N-terminal.

This means that residues implied in Gal-3 nucleus to cytoplasm translocation (*i.e.* **Ser6** and **Ser12**) are located in a defined region, which may facilitate the phosphorylation process, and are far away from the carbohydrate-binding pocket that mediates glycan-dependent functions (see below). The other constructed N-PG portion, **Ala100-Val116 (Fig 38, right)**, comprises residues that overlap with the CRD domain and contains a cleavage site for MMPs at the **Pro112/Leu113** bond, leaving out a zone of six amino acids that conform the link between the C- and N-terminal domains. The rest of this region, comprising residues **Ala100-Pro106** from the IX repeat extends the structure away from the **S1** β-sheet of the C-terminal CRD.

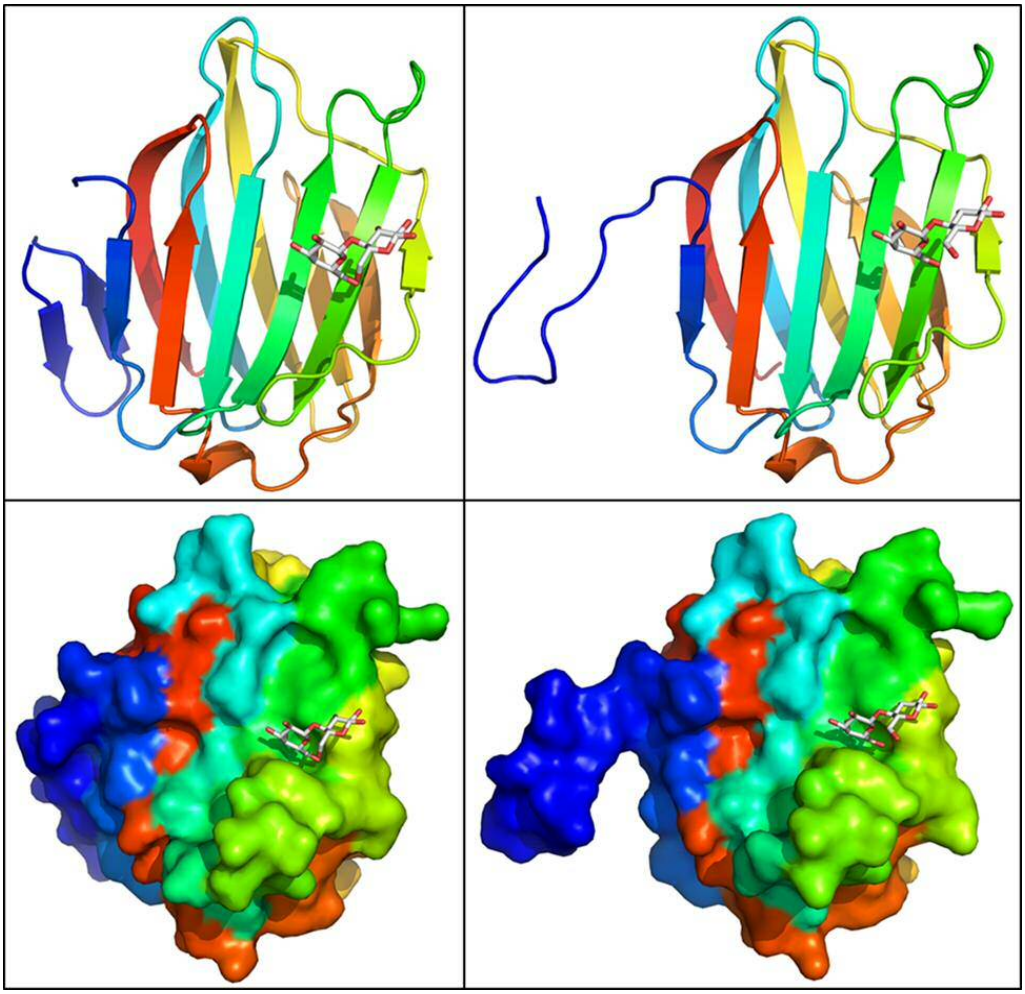


Fig 38. Isolated monomers of Gal-3[N VII-IX] with two different N-PG sections. The amino acids newly constructed in the N-LD section (left) seem to 'come back' from the elongated structure of section IX of the N-PG (right).

Superposition of both monomers (**Fig 39**) shows a well-fitting of the CRD domains (r.m.s.d. 0.8 Å for Cα-atoms), and suggests the way to connect both ends, *i.e.* the entire 22-residue segment of the repeating units VII and VIII and the connection with the N-terminal, since lack of electron density prevents the modelling of this part of the protein. To do a more detailed analysis of the structure and to provide a complete picture of Gal-3, experimental SAXS data was used to determine the global architecture of the protein, with samples from both Gal-3[N VII-IX] and Gal-3_{FL}.

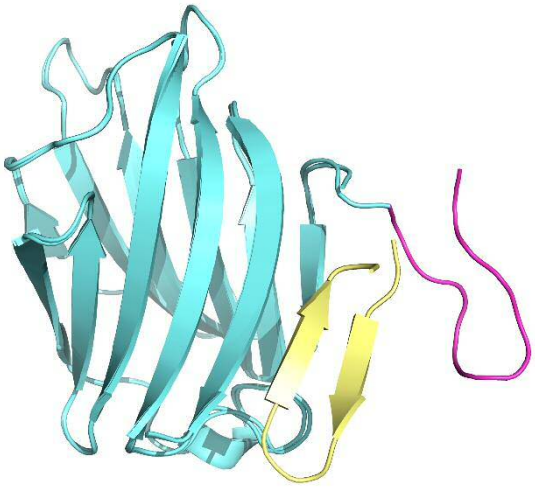


Fig 39. The N-PG structure. Overlapped representation of all the constructed amino acids from the X-ray diffraction data.

As it is known that Gal-3 multimerization is initiated at a certain threshold concentration (nucleation) (Lepur et al 2012), is ligand-dependent (Funasaka et al 2014b), and this self-association can cause aggregation and/or precipitation, size exclusion chromatography with buffer exchange was successfully applied to obtain clear solutions of both Gal-3[N VII-IX] and Gal-3_{FL} for testing in a SAXS beamline (BM29 line, ESRF).

SAXS data indicated that both Gal-3[N VII-IX] and Gal-3_{FL} are monomeric at concentrations below 8 mg·ml^{-1} based on the maximum particle dimension (D_{max}) and Guinier plots. At concentrations above 8 mg·ml^{-1}, Gal-3[N VII-IX] started to aggregate and data could not be analyzed. The full-length protein showed even higher propensity to aggregation. Molecular envelopes were generated *ab initio* (Svergun 1999), selected, combined and filtered to produce an averaged model in which both shapes corresponded to elongated particles (**Figs 25b/26b** from Results).

Based on the surface envelopes calculated from the SAXS data and with the available Gal-3[N VII-IX] crystal structure, it can be modelled, inside the empty surface area, the extended segment corresponding to the VII and VIII repeats missing in the crystal structure. First, for this connecting segment the models generated were analyzed by secondary structure prediction servers, RaptorX and I-TASSER. The predicted structure was an almost linear and long polypeptide chain with the N-terminal pointing in the opposite direction of the CRD. Coot was used to build the model, taking into

account the Gal-3 biochemical properties, *i.e.* the positions of two serine residues as acceptors for phosphorylation (**Ser6** and **Ser12**) and the seven cleavage sites for MMPs and PSA (Kopitz et al 2014). The first requirement is fulfilled since **Ser6** and **12** are correctly positioned in the double-stranded antiparallel β-sheet (**F0** and **F-1**) in a conformation that favours phosphorylation (see below). Secondly, sites for protease cleavage must be solvent exposed (**Gly31/Ala32, Ala62/Tyr62, Pro75/Gly76, Pro89/Ser90, Pro101/Ala102, Tyr106/Gly107** and **Pro112/Leu113**). With this in mind, the corresponding residues were accommodated inside the surface in a closed conformation with an elongated shape and the molecule was submitted to molecular dynamics simulations. The final models are shown in **Fig 40**.

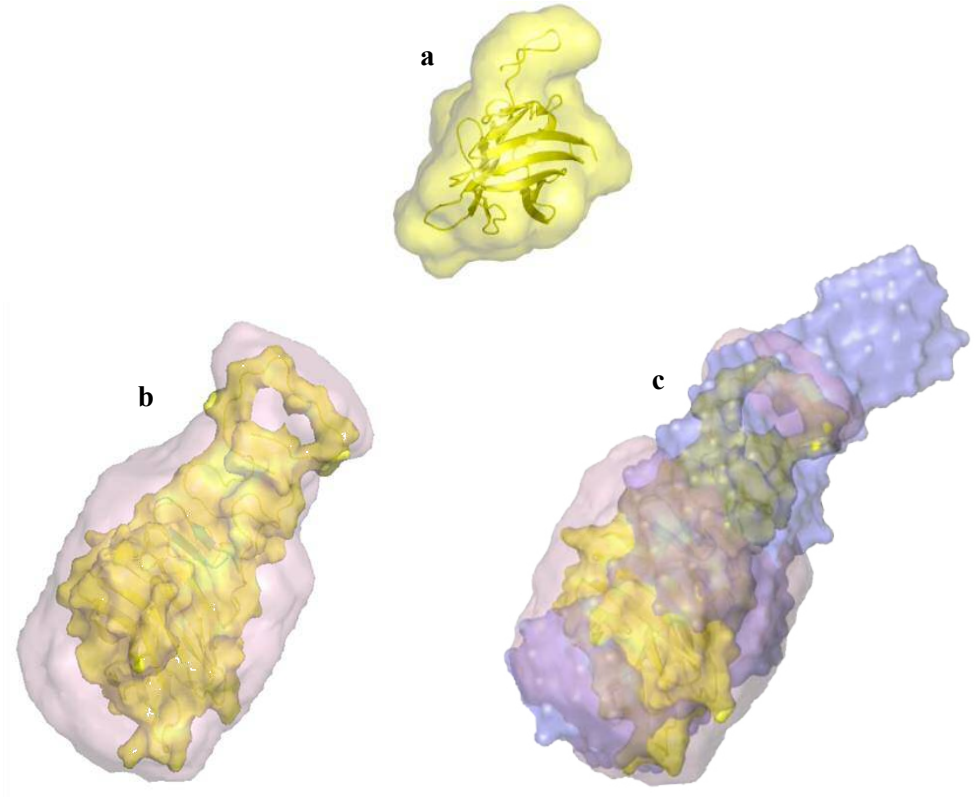


Fig 40. Electrostatic surface representation of crystalline Gal-3 CRD, and SAXS-resolved Gal-3[N VII-IX] and Gal-3$_{FL}$. (**a**) The crystal structure for Gal-3 CRD (yellow) presents a compact shape. (**b**) SAXS data for Gal-3[N VII-IX] (magenta) shows a structure where the [N VII-IX] was built (yellow inside magenta) and that supports the idea of the N-PG not adopting a random-coil conformation. (**c**) Gal-3$_{FL}$ (blue) grows further away from the globular shape, where amino acids from sections I to VI would accommodate.

One of the most important Gal-3 post-translational modifications, that has the potential to alter important biological functions, is the phosphorylation of **Ser6** (**Fig 41**). **Ser12** and several **Tyr** residues in the protein are susceptible to be phosphorylated as well, however, it is **Ser6** phosphorylation/dephosphorylation what has a key role in regulating, for instance, Gal-3 TRAIL-induced apoptosis, cell distribution, and binding ability (Gao, Liu et al 2017). In the Gal-3[N VII-IX] structure modelled with both X-ray crystallography and SAXS data, this **Ser6** residue is readily available in the surface of a pocket where kinases may bind and activate Gal-3. Even for the full-length Gal-3, the position of the **Ser6** is likely to remain as it is.

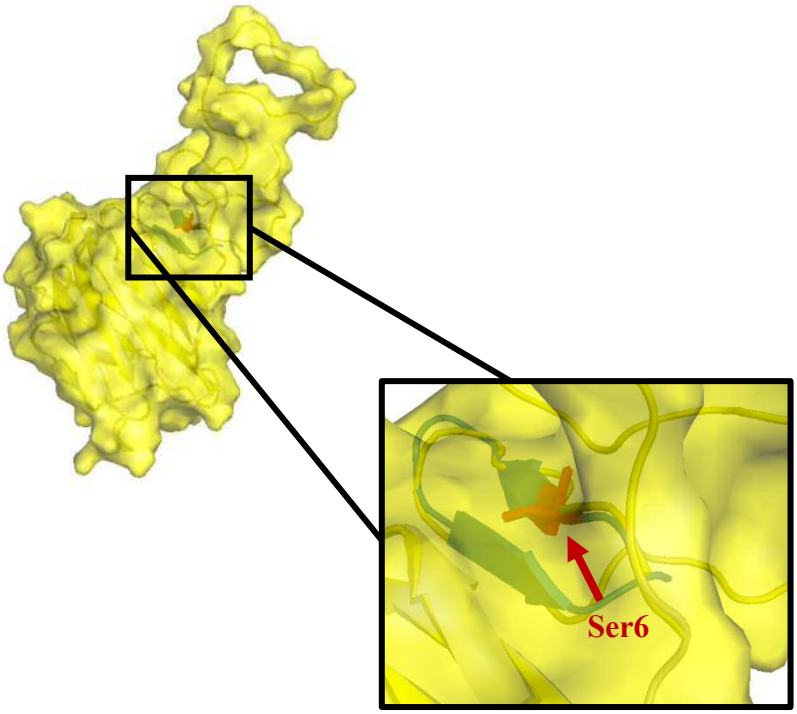


Fig 41. Ser6 phosphorylation site. Modelling of Gal-3[N VII-IX] with X-ray diffraction and SAXS data allows to discover that Galectin-3's major phosphorylation site is readily available in the molecular surface.

More specifically, Gal-3 **Ser6** phosphorylation by Protein Kinase I (formerly known as Casein Kinase I, **CKI**) regulates Gal-3 function producing the only nuclear fraction that can be exported from the nucleus to the cytoplasm and protect cells from mitochondrial membrane permeabilization-mediated apoptosis. Unphosphorylated

mutant Gal-3 accumulates in the nucleus and doesn't exhibit its anti-apoptotic function (Nakahara et al 2006; Gao, Liu et al 2017). A closer look to the model (**Fig 42**) shows that the pocket with the **Ser6** in Gal-3[N VII-IX] forms a groove perfectly complementary in shape with the active site of the CKI (**Arg128, PDBID: 5MQV**; Halekotte et al 2017) that allows this kinase an easy access to the binding site pocket. This means that even MMP-processed Gal-3 would be capable of exiting the nucleus to aid tumorous cells escape the cell death program.

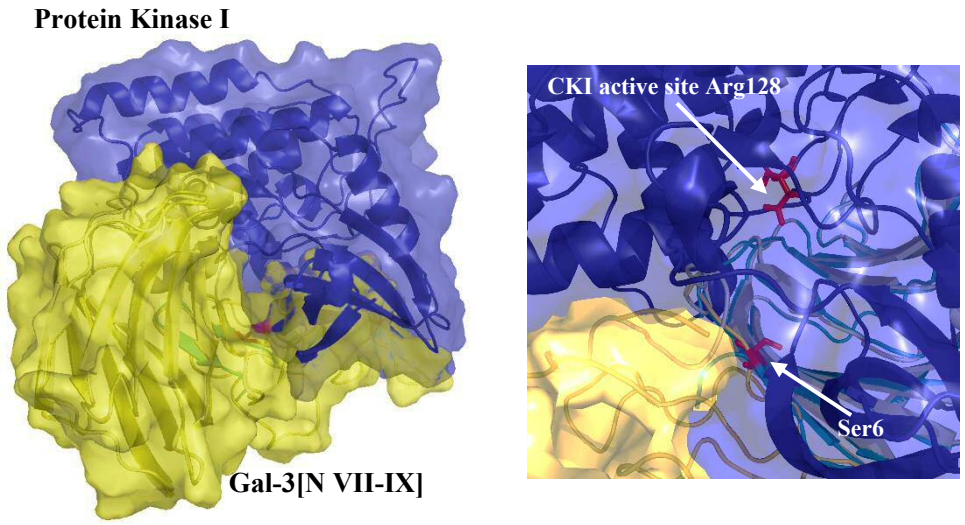


Fig 42. Protein Kinase I binding to Gal-3. The tertiary structure of Gal-3[N VII-IX] forms a pocket that leaves a major phosphorylation site available for protein-protein interaction with Gal-3's natural kinase in processes of cytoplasm to nucleus translocation, CKI.

The anti-apoptotic function of exported **Ser6**-phosphorylated Gal-3 works in favor of cancerous cells, which overproduce Gal-3 and this in turn helps them escape cell death. Prevention of phosphorylation will prevent its nuclear export, but no specific inhibitors for this have been designed. Gal-3 structure deserves further investigation in order to understand its functions, mechanisms of action, and to develop a library of structures that may serve as agonists or antagonists for protein regulation in pathological processes. This new structure with the N-terminal section that contains the phosphorylation site sets the path for development of further research in aid of the existing therapies. In all, the success for the first time on the structural resolution of a truncated variant that contains the chimeric module of Gal-3, although shortened, marks the start point for Gal-3_{FL} structural elucidation.

On to C-GRP-C, the gene for the Galectin-Related Protein is present exclusively in vertebrates, with high-level sequence conservation and similar chromosomal positioning when, in contrast, occurrence of canonical galectins is known to also happen in organisms such as fungi, nematodes and insects. Therefore, the GRP gene ought not to be a recent acquisition by duplication of an ancestral gene within a distinct species but an integral component of the vertebrate genome. This sequence conservation among various species implies a very strong positive selection, as generally seen for genes encoding proteins with multiple aspects involved in critical interactions, say with itself, other proteins and/or specific ligands (Cooper 2002).

Biochemical and structural characterization of GRP, a product from a gene under strong positive selection, is a pre-requisite for functional characterization and network monitoring of the whole galectin family (García, Flores-Ibarra, Michalak et al 2016). Because this protein does not bind β-galactosides, it is not possible to classify it under any galectin type for its CRD, *i.e.* proto, tandem or chimera. Prototype and tandem-repeat galectins normally form homo- and heterodimers, and in human GRP-C crystals, aggregation into dimers and tetramers has been observed (Zhou et al 2008). In contrast, several experiments were performed on C-GRP-C and the protein was always a monomer.

The crystal structure of C-GRP-C, solved at 1.55 Å resolution, revealed its monomeric nature. As expected by similarities in the amino acid sequence with galectins, C-GRP-C conserves the *jelly-roll* fold (**Fig 43a**), formed by a β-sandwich of antiparallel strands, **F1-F5** in one face, **S1-S6** in the other face, and a $\mathbf{3_{10}}$ helix connecting both sides. However, the sequence signature at the canonical binding site, located in strands **S4-S6** in galectins, drastically changes in C-GRP-C, affecting its ability to bind β-galactosides. A peculiar feature of C-GRP-C at the central position of the CRD is the presence of a **Trp106** residue, which is the responsible for the carbohydrate-π interactions with the B-face of galactose in galectins (**Fig 32** from Results). Birds, fish and amphibians have a **Trp** in this position, whereas mammals consistently present a substitution by **Arg/Lys**. However, this CH-π interaction, as fundamental as it is for binding galactose moieties in galectins, still fails to bind these type of carbohydrate in C-GRP-C.

Fig 43b shows a comparative view of the C-GRP-C structure with a canonical chicken galectin, CG-2. The sequence signature for binding in CG-2 (**His45**, **Asn47**,

Arg49, **Asn58**, **Trp65**, **Glu68** and **Arg70**) is turned out into **Glu86**, **Lys88**, **Val90**, **Asn99, Trp106**, **Glu109** and **Ser111** in C-GRP-C: only three out of seven positions are thus maintained. Furthermore, **Val90**, **Glu86** and, especially, **Lys88**, are impairing the set of contacts for the 4′, 6′-hydroxyls of galactose moieties. In particular, it is the **Lys88** (in yellow, **Fig 43b**) that protrudes to an extent that disfavors a comfortable fit for lactose, so that not even the presence of **Trp106** can compensate for this distorsion (García, Flores-Ibarra, Michalak et al 2016).

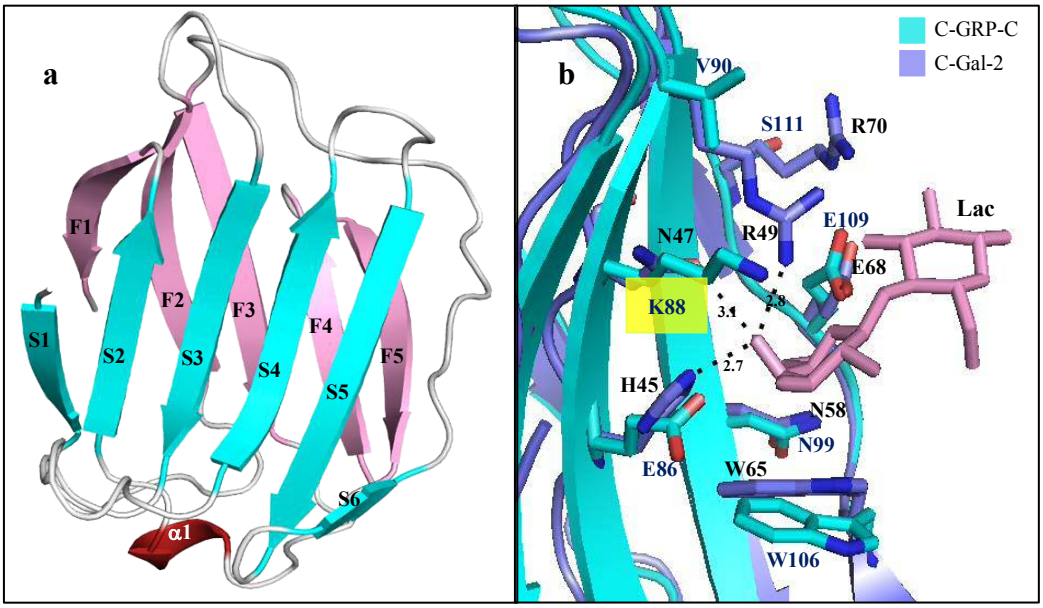


Fig 43. Overall crystal structure of C-GRP-C. (**a**) The protein maintains the *jelly-roll* folding typical of the galectin family. **Superposition of a canonical chicken galectin (CG-2) and C-GRP-C.** (**b**) Sugar binding residues of CG-2 (purple) and those of equivalent positions in C-GRP-C (cyan) are shown in stick mode.

Still, the low root-mean-square deviations (rmsd) for Cα atoms, with values ranging from 1.07 Å to 1.54 Å, between members of the chicken galectin family: CG-1A (1.42 Å), CG-1B (1.54 Å), CG-2 (1.33 Å) and CG-8N (1.07 Å) underscore the close relationship between galectins and GRP. The superposition of the C-GRP-C structure with that of the listed canonical CGs, however, disclosed notable differences in several regions involving loops between adjacent β-strands at the concave face of the CRD pocket. In detail, the S3-S4 loop, which connects antiparallel β-strands F4-F5, is five residues longer in C-GRP-C and CG-8N than in prototype CGs, for whom (CG-1A and -1B) the S4-S5 loop is extended (**Fig 44**).

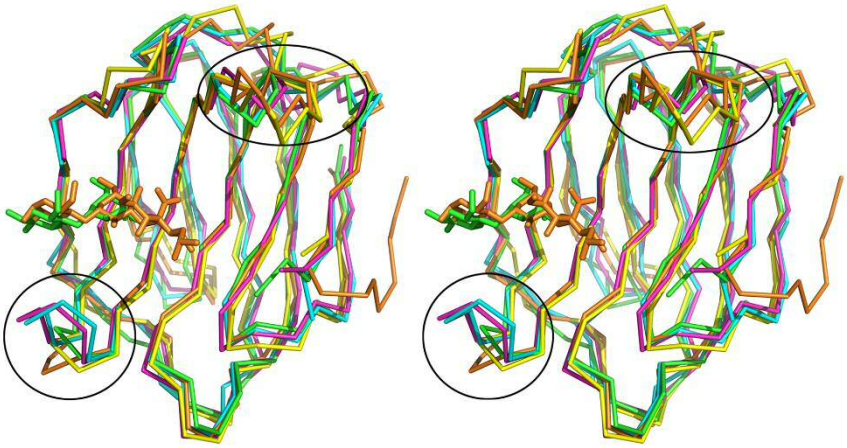


Fig 44. Structural superposition of C-GRP-C (yellow) with canonical CGs. CG-1A (cyan; 1QMJ), CG-1B (purple; 3DUI), CG-2 (green; 2YMZ) and CG-8N (orange; 4WVW). Black circles highlight the major structural differences in two regions involving connecting loops at the concave face of the CRD pocket.

Then it may be that deviations in the C-GRP-C amino acid sequence account for the lack of β-galactoside-binding by this protein, although, it is of note that deviations in position of the signature glycan-binding sequence do not necessarily mean that a member of the galectin family will loose its glycan binding activity. Actually, congerin P, a galectin from peritoneal cells of the conger eel (Watanabe et al 2012), and the *Coprinus cinerea* galectin CGL3 (Wälti et al 2008b) have operative binding despite similar alterations to C-GRP-C.

Consequently, the absence of key amino acids at the CRD binding site could not be the only reason for C-GRP-C distinct behavior. Instead, changes in amino acids that are far from the carbohydrate-binding site have been seen in Gal-8 to impair glycan-binding ability (Ruiz et al 2014). A look at the electrostatic surface of both C-GRP-C and CG-2 (**Fig 45**) shows the prominent **Lys88**, along with a highly acidic surface and the absence of a tunnel-like cavity, present in galectins, that could reduce C-GRP-C likelihood to host galactose moieties.

Additionally, the specific location of GRP in bone marrow cells might be indicative for the different biological functions of C-GRP in contrast to galectins (Cooper 2002). Altogether, the present results on the C-GRP-C structure highlight all the changes present in a Galectin-Related Protein in comparison to the rest of the proteins from the galectin family.

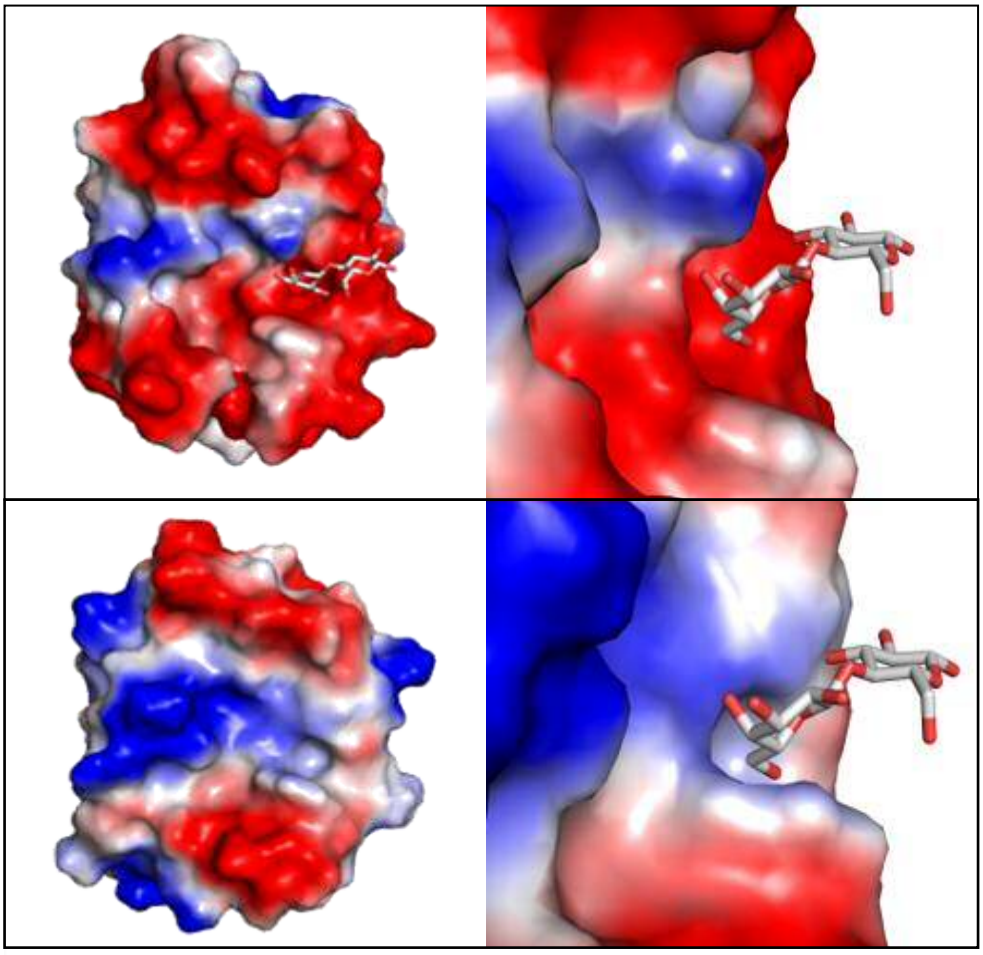


Fig 45. Electrostatic surface potential maps comparing C-GRP-C (top) with CG-2 (bottom) contoured from -10 kT/e (red) to +10 kT/e (blue). The surface of C-GRP-C is highly negative owing to the surface distribution of acidic residues, and the absence of a tunnel-like cavity present in canonical galectins, which accomodates the galactose moieties, is really evident.

Finally, as C-GRP-C, in analogy to human GRP, was engineered to produce a shortened and crystallizable form, a series of biophysical experiments were also conducted to determine if the structure of the engineered variant resembles the entire full-length protein. The monomer presence was independently confirmed by SAXS and AUC for C-GRP-C. Sedimentation velocity experiments demonstrated that C-GRP and the shortened C-GRP-C maintained the same monomeric status (see Results).

SAXS modelling was used to obtain a three-dimensional bead model of a monomer with a maximum dimension of 48.2 ± 2.3 Å and a molecular mass of 16.2 ± 1.6 kDa

that is in good concordance with the C-GRP-C structure (**Fig 46**). Furthermore, cell-binding capacity experiments showed very similar reactivity of both the physiological and the shortened form of C-GRP (García, Flores-Ibarra, Michalak et al 2016).

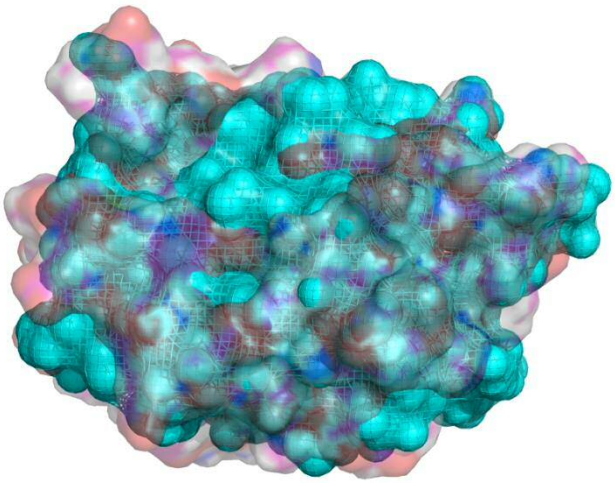


Fig 46. Averaged bead model of SAXS-resolved C-GRP-C in comparison with C-GRP-C crystal structure electrostatic surface. The statistical model obtained by SAXS and the electrostatic surface of the protein's crystal have a good agreement both in shape and size.

Sequence conservation at a high level, as seen for GRP genes in vertebrates, is a sign for the development of a distinct function at the canonical lectin activity's expense (García, Flores-Ibarra, Michalak et al 2016). This structure takes another step towards the entire galectin family fingerprinting, and its crystal structure results fundamental for characterization of GRP potential physiological ligands.

CONCLUSIONS

CONCLUSIONS

1. Gal-3[N VII-IX] crystallographic structure, the first one solved for this protein with a section of the N-PG domain, provides a key starting point for the completion of Gal-3_{FL} structure elucidation.

2. Gal-3[N VII-IX] crystallographic structure establishes the molecular basis for the study of the distinct behavior of Gal-3 with and without its N-terminal domain.

3. Gal-3[N VII-IX] structure shows that a key residue of Gal-3 post-translational modifications, *i.e.* **Ser6** phosphorylation, accommodates in a readily accessible position for binding other effector molecules, such as Gal-3's natural kinase, CKI.

4. Gal-3[N VII-IX] and Gal-3_{FL} SAXS structure clearly shows the deviation of Gal-3 CRD globular structure into an elongated form, where the crystallographic and biochemical structure fits well.

5. C-GRP-C crystallographic structure shows that avian GRP conserves the jelly-roll folding adopted by all galectins and, as previously shown by DNA sequence, presents a **Trp** residue (not present in human GRP) in the same position as the **Trp** in galectins that interacts with the galactose ring.

6. C-GRP-C and C-GRP present themselves in the form of monomers, both in crystal form and in solution, corroborated by SAXS and AUC.

7. C-GRP-C, as human GRP, does not bind β-galactosides, and the crystallographic structure shows the absence of other key residues for galectins' glycan binding, and the presence of a protruding **Lys88**, accounting for the lack of C-GRP-C binding canonical galectin ligands.

8. The highly acidic surface of GRP, and its localization in bone marrow cells, points to a different natural ligand and/or binding site for this protein.

CONCLUSIONES

- La estructura cristalográfica de Gal-3[N VII-IX], la primera resuelta para esta proteína con una sección del dominio N-PG, proporciona un punto de inicio para la resolución estructural de la proteína completa Gal-3_{FL}.

- La estructura cristalográfica de Gal-3[N VII-IX] establece la base molecular para el estudio del comportamiento distintivo de Gal-3 con y sin el dominio N-terminal.

- La estructura de Gal-3[N VII-IX] muestra que un residuo clave en las modificaciones postraduccionales de Gal-3, esto es, la fosforilación de Ser6, se posiciona en una región accesible para la unión de moléculas efectoras, tal como la quinasa natural de Gal-3, CKI.

- Las estructuras resueltas por SAXS de Gal-3[N VII-IX] y Gal-3_{FL} muestran claramente la desviación de la estructura globular del CRD de Gal-3, con una forma alargada donde la estructura cristalográfica se acomoda fácilmente.

- La estructura cristalográfica de la C-GRP-C muestra que GRP en aves conserva el plegamiento *jelly-roll* adoptada por todas las galectinas y que, como se conocía por la secuencia de DNA, presenta un residuo de **Trp** (ausente en GRP humana) en la misma posición que el **Trp** en galectinas que interactúa con el anillo galactósido.

- C-GRP-C y C-GRP se presentan como monómeros, tanto en cristal como en solución, como fue corroborado por SAXS y AUC.

- La C-GRP-C, como la GRP humana, no une β-galactósidos, y la estructura cristalográfica muestra la ausencia de residuos clave para la unión galectina-glicano y la presencia de una Lys88, lo que es determinante en la ausencia de afinidad de C-GRP-C para unir ligandos canónicos de galectinas.

- La alta presencia de residuos ácidos en la superficie de la GRP, así como su localización en células de la médula ósea, apunta a un ligando natural y/o a un sitio de unión diferentes para esta proteína.

REFERENCES

REFERENCES

1. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC and Zwart PH. (**2010**) *PHENIX: a comprehensive Python-based system for macromolecular structure solution*. Acta Crystallogr D, 66(2):213-21. **doi: 10.1107/S0907444909052925**.

2. Altschul SF, Madden TL, Schäffer AA, Zhang ZJ, Miller W and Lipman DJ. (**1997**) *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nuc Acid Res 25(17):3389-3402. **PMID: 9254694**.

3. Bassik MC, Kampmann M, Lebbink RJ, Wang S, Hein MY, Poser I, Weibezahn J, Horlbeck MA, Chen S, Mann M, Hyman AA, LeProust EM, McManus MT and Weissman JS. (**2013**) *A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility*. Cell, 152:909-922. **doi: 10.1016/j.cell.2013.01.030**.

4. Birdsall B, Feeney J, Burdett ID, Bawumia S, Barboni EA and Hughes RC. (**2001**) *NMR solution studies of hamster galectin-3 and electron microscopic visualization of surface-adsorbed complexes: evidence for interactions between the N- and C-terminal domains*. Biochemistry, 40(15):4859-66. **PMID: 11294654**.

5. Cagnoni AJ, Pérez Sáez JM, Rabinovich GA and Mariño KV. (**2016**) *Turning-off signaling by siglecs, selectins, and galectins: chemical inhibition of glycan-dependent interactions in cancer*. Front Oncol 6:109. **doi: 10.3389/fonc.2016.00109**.

6. Cole JL, Lary JW, Moody T and Laue TM. (**2008**) *Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium*. Methods Cell Biol, 84:143-179. **doi: 10.1016/S0091-679X(07)84006-4**.

7. Compagno D, Gentilini LD, Jaworski FM, González I, Contrufo G and Laderach D. (**2014**). *Glycans and Galectins in prostate cancer biology, angiogenesis and metastasis*. Glycobiology, 24(10):899–906. **doi: 10.1093/glycob/cwu055**.

8. Cooper DN. (**2002**) *Galectinomics: finding themes in complexity*. Biochim Biophys Acta, 1572(2–3):209–231. **PMID:12223271**.

9. Cowles EA, Agrwal N, Anderson RL and Wang JL. (**1990**) *Carbohydrate-binding protein 35. Isoelectric points of the polypeptide and a phosphorylated derivative.* J Biol Chem, 265(29):17706-17712. **PMID: 2170392**.

10. Cowtan K. (**2001**) *Phase problem in X-ray crystallography, and its solution*. Encyclopedia of Life Sciences, MacMillan Publishers Ltd, Nature Publishing Group.

11. DeLano WL. (**2012**) http://www.pymol.org**.**

12. Diederichs K and Karplus PA. (**1997**) *Improved R-factors for diffraction data analysis in macromolecular crystallography*. Nature Struct Biol 4:269–275. **doi: 10.1038/nsb0497-269**.

13. Drenth J. (**1994**) *Principles of Protein X-ray Crystallography*. Springer-Verlag: New York.

14. Drickamer K. (**1993**) *Ca^{2+}-dependent carbohydrate-recognition domains in animal proteins*. Curr Opin Struct Biol 3:393400.

15. Dumic J, Dabelic S and Flogel M. (**2006**) *Galectin-3: an open-ended story*. Biochim Biophys Acta – Gen Subj, 1760:616–635. **doi: 10.1016/j.bbagen.2005.12.020**.

16. Elola MT, Wolfenstein-Todel C, Troncoso MF, Vasta GR and Rabinovich GA. (**2007**) *Galectins: matricellular glycan-binding proteins linking cell adhesion, migration, and survival*. Cell Mol Life Sci, 64:1679–1700. **doi: 10.1007/s00018-007-7044-8**.

17. EMBL Protocols, (https://www.embl.de/pepcore/pepcore_services/protein_expression/ecoli/index.html).

18. Emsley P, Lohkamp B, Scott WG and Cowtan K. (**2010**) *Features and development of Coot*. Acta Cryst 66:486–501. **doi: 10.1107/S0907444910007493**.

19. Evans PR. (**2011**). *An introduction to data reduction: space-group determination, scaling and intensity statistics*. Acta Cryst D 67:282–292. **doi: 10.1107/S090744491003982X**.

20. Evans PR and Murshudov GN. (**2013**) *How good are my data and what is the resolution*. Acta Cryst D, 69:1204-14. **doi: 10.1107/S0907444913000061**.

21. Flores-Ibarra A, Ruiz FM, Vértesy S, André S, Gabius H-J and Romero A. (**2015**) *Preliminary X-ray crystallographic analysis of an engineered variant of human chimera-type galectin-3 with a shortened N-terminal domain*. Acta Crystallog F, 71(2):184–188. **doi: 10.1107/S2053230X15000023**.

22. Franke D, Kikhney AG and Svergun DI. (**2012**) *Automated acquisition and analysis of Small Angle X-ray Scattering data*. Nucl Instrum Methods Phys Res A, 689:52-59. **doi: 10.1016/j.nima.2012.06.008**.

23. Friedrich W, Knipping P and Laue M. (1912) *Sitzungsberg*. K Bayer Akad Wiss, 303-322.

24. Frigeri LG, Zuberi RI and Liu F-T. (**1993**) *Epsilon BP, a beta-galactoside-binding animal lectin, recognizes IgE receptor (Fc epsilon RI) and activates mast cells*. Biochemistry, 32(30):7644–7649. **PMID: 8347574**.

25. Funasaka T, Raz A and Nangia-Makker P. (**2014a**) *Nuclear transport of galectin-3 and its therapeutic implications*. Sem Cancer Bio, 27:30-38. **doi: 10.1016/j.semcancer.22014.03.004**.

26. Funasaka T, Raz A and Nangia-Makker P. (**2014b**) *Galectin-3 in angiogenesis and metastasis*. Glycobiology, 24(10):886–891. **doi: 10.1093/glycob/cwu086**.

27. Gabius H-J. (**2000**) *Biological information transfer beyond the genetic code: The sugar code*. Naturwissenschaften 87:108. **doi: 10.1007/s001140050687**.

28. Gabius HJ, Siebert HC, André S, Jiménez-Barbero J and Rüdiger H. (**2004**) *Chemical biology of the sugar code*. Chembiochem, 5(6):740-764. **doi: 10.1002/cbic.200300753**.

29. Gabius H-J. (**2009**) *The Sugar Sode: Fundamentals of Glycosciences*. Wiley-VCH, pp 53-69.

30. Gabius H-J, André S, Jiménez-Barbero J, Romero A and Solís D. (**2011**) *From lectin structure to functional glycomics: principles of the sugar code*. Trends Biochem Sci, 36(6):298-313. **doi:10.1016/j.tibs.2011.01.005**.

31. Gao X, Liu D, Fan Y, Li X, Xue H, Ma Y, Zhou Y and Tai G. (**2012**) *The Two Endocytic Pathways Mediated by the Carbohydrate Recognition Domain and Regulated by the Collagen-like Domain of Galectin-3 in Vascular Endothelial Cells*. PLoS ONE, 7(12). **doi: 10.1371/journal.pone.0052430**.

32. Gao X, Balan V, Tai G and Raz A. (**2014**) *Galectin-3 induces cell migration via a calcium-sensitive MAPK*. Oncotarget, 5(8):2077–2084. **doi: 10.18632/oncotarget.1786**.

33. Gao X*, Liu J*, Liu X, Li L and Zheng J. (**2017**) *Cleavage and phosphorylation: important post-translational modifications of galectin-3*. *Authors contributed equally. Cancer Metastatic Rev, online. **doi: 10.1007/s10555-017-9666-0**.

34. García C G*, Flores-Ibarra A*, Michalak M*, Khasbiullina N, Bovin NV, André S, Manning JC, Vértesy S, Ruiz FM, Kaltner H, Kopitz J, Romero A and Gabius H-J. (**2016**) *Galectin related protein: an integral member of the network of chicken galectins. 1. From strong sequence conservation of the gene confined to vertebrates to biochemicaal characteristics of the chicken protein and its crystal structure*. *Authors contributed equally. Biochim Biophys Acta – Gen Subj, 1860(10):2285-2297. **doi: 10.1016/j.bbagen.2016.06.001**.

35. Glinsky VV, Glinsky GV, Glinskii OV, Huxley VH, Turk JR, Mossine VV, Deutscher SL, Pienta KJ and Quinn TP. (**2003**) *Intravascular metastatic cancer cell homotypic aggregation at the sites of primary attachment to the endothelium*. Cancer Res, 63(13):3805–3811. **PMID: 12839977**.

36. Gong HC, Honjo Y, Nangia-Makker P, Hogan V, Mazurak N, Bresalier RS and Raz A. (**1999**) *The NH_2 terminus of galectin-3 governs cellular compartmentalization and functions in cancer cells*. Cancer Res, 59(24):6239–6245. **PMID: 10626818**.

37. Grant TD, Luft JR, Wolfley JR, Tsuruta H, Martel a, Montelione GT and Snell EH. (2011) *Small Angle X-ray Scattering as a complementary tool for high-throughput sturctural studies*. Biopomlymers, 95(8):517-530. doi: 10.1002/bip.21630.

38. Guinier A and Foumet F. (**1955**) *Small Angle Scattering of X-rays*. Wiley, New York.

39. Halekotte J, Witt L, Ianes C, Kruger M, Buhrmann M, Rauh D, Pichlo C, Brunstein E, Luxenburger A, Baumann U, Knippschild U, Bischof J and Peifer C. (**2017**) *Optimized 4,5-diarylimidazoles as potent/selective inhibitors of Protein Kinase CKIδ and their structural relation to p38α MAPK*. Molecules, 22(4):E522. **doi: 10.3390/molecules22040522**.

40. Hernández JD and Baum LG. (**2002**) *Ah, sweet mystery of death! Galectins and control of cell fate*. Glycobiology, 12(10):127R–136R. **doi: 10.1093/glycob/cwf081**.

41. Hirabayashi J, Hashidate T, Arata Y, Nishi N, Nakamura T, Hirashima M, Urashima T, Oka T, Futai M, Muller WE, Yagi F and Kasai K. (**2002**) *Oligosaccharide specificity of galectins: a search by frontal affinity chromatography*. Biochim Biophys Acta - Gen Subj, 1572:232–254. **PMID: 12223272**.

42. Ippel H, Miller MC, Vértesy S, Zhang Y, Cañada FJ, Suylen D, Umemoto K, Romanò C, Hackeng T, Tai G, Leffler H, Kopitz J, André S, Kübler D, Jiménez-Barbero J, Oscarson S, Gabius H-J and Mayo KH. (**2016**) *Intra- and intermolecular interactions of human galectin-3: assessment by full-assignment-based NMR*. Glycobiology, Advanced access. **doi: 10.1093/glycob/cww021**.

43. Jacques DA and Trewhella J. (**2010**) *Small-angle scattering for structural biology--expanding the frontier while avoiding the pitfalls*. Protein Sci, 19:642-657. **doi: 10.1002/pro.351**.

44. Kabsch W. (**2010**) *XDS*. Acta Cryst D, 66:125-32. **doi: 10.1107/S0907444909047337**.

45. Kaltner H, Kübler D, López-Merino L, Lohr M, Manning JC, Lensch M, Seidler J, Lehmann WD, André S, Solís S and Gabius H-J. (**2011**) *Toward comprehensive analysis of the galectin network in chicken: unique diversity of galectin-3 and comparison of its localization profile in organs of adult animals to the other four members of this lectin family*. Anatomical Rec, 294(3):427–44. **doi: 10.1002/ar.21341**.

46. Karplus PA and Diederichs K. (**2012**) *Linking crystallographic model and data quality*. Science, 336:1030-3, 2012. **doi: 10.1126/science.1218231**.

47. Katayama Y, Battista M, Wei-Ming K, Hidalgo A, Peired AJ, Thomas SA and Frenette PS. (**2006**) *Signals from the sympathetic nervous system regulate hematopoietic stem cell egress from bone marrow*. Cell 124:407–421. **doi: 10.1016/j.cell.2005.10.041**.

48. Kopitz J, Vértesy S, André S, Fiedler S, Schnölzer M and Gabius H-J. (**2014**) *Human chimera-type galectin-3: Defining the critical tail length for high-affinity glycoprotein/cell surface binding*. Biochimie, 104:90–99. **doi: 10.1016/j.biochi.2014.05.010**.

49. Le Maire M, Thauvette L, de Foresta B, Viel A, Beauregard G and Potier M. (**1990**) *Effects of ionizing radiations on proteins. Evidence of non-random fragmentations and a caution in the use of the method for determination of molecular mass*. Biochem J, 267(2):431-439. **PMID: 2334402**.

50. Leffler H, Carlsson S, Hedlund M, Qian Y and Poirier F. (**2002**) *Introduction to galectins*. Glycoconj J, 19(7–9):433–440. **doi: 10.1023/B:GLYC.0000014072.34840.04**.

51. Lepur A, Salomonsson E, Nilsson UJ and Leffler H. (**2012**) *Ligand induced galectin-3 protein self-association*. J Biol Chem, 287(26):21751-21756. **doi: 10.1074/jbc.C112.358002**.

52. Lichtenstein RG and Rabinovich GA. (**2013**) *Glycobiology of cell death: when glycans and lectins govern cell fate*. Cell Death Diff, 20(8): 976–86. **doi:10.1038/cdd.2013.50**.

53. Liu F-T, Hsu DK, Zuberi RI, Kuwabara I, Chi EY and Henderson WR Jr. (**1995**) *Expression and function of galectin-3, a beta-galactoside-binding lectin, in human monocytes and macrophages*. Am J Pathol, 147(4):1016–1028. **PMID: 7573347**.

54. Liu FT, Patterson RJ and Wang JL. (**2002**) *Intracellular functions of galectins*. Biochim Biophys Acta – Gen Subj, 1572:263–273. **doi:10.1016/S0304-4165(02)003136**.

55. Liu F-T and Rabinovich GA. (**2005**) *Galectins as modulators of tumour progression*. Nature Reviews - Cancer, 5(1):29–41. **doi:10.1038/nrc1527**.

56. López-Lucendo M, Solís D, André S, Hirabayashi J, Kasai K, Kaltner H, Gabius H-J and Romero A. (**2004**) *Growth regulatory human galectin-1: crystallographic characterization of the structural changes induced by single-site mutations and their impact on the thermodynamics of ligand binding*. J Mol Biol, 343(4):957-970. **doi: 10.10167j.jmb.204.08.078**.

57. López-Lucendo MF, Solís D, Sáiz JL, Kaltner H, Russwurm R, André S, Gabius H-J, Romero A. (**2009**) *Homodimeric chicken galectin CG-1B (C-14): crystal structure and detection of unique redox-dependent shape changes involving inter- and intrasubunit disulfide bridges by gel filtration, ultracentrifugation, site-directed mutagenesis, and peptide mass fingerprinting*, J Mol Biol, 386:366–378. **doi: 10.1016/j.jmb.2008.09.054**.

58. Machado CML, Andrade LNS, Teixeira VR, Costa FF, Melo CM, dos Santos SN, Nonogaki S, Liu F-T, Bernardes ES, Camargo AA and Chammas R. (**2014**) *Galectin-3 disruption impared tumoral angiogenesis by reducing VEGF secretion from TGF1-induced macrophages.* Cancer medicine, 3(2):201-214. **doi: 10.1002/cam4.173**.

59. Markowska AI, Jefferies KC and Panjwani N. (**2011**) *Galectin-3 protein modulates cell surface expression and activation of vascular endothelial growth factor receptor 2 in human endothelial cells*. J Biol Chem, 286:29913–29921. **doi: 10.1074/jbc.M111.226423**.

60. Matthews BW. (**1968**) *Solvent content of protein crystals*. J Mol Biol, 33(2):491-497. **PMID: 5700707**.

61. Mendoza ME and Moreno A. (**2015**) *Cristalogénesis biológica y fundamentos de difracción con rayos X*. B. Universidad Autónoma de Puebla, Dirección de Fomento Editorial.

62. Messerschmidt A. (**2007**) X-ray crystallography of biomacromolecules. Wiley-VCH Verlag GmbH & Co. Weinheim.

63. Miroux B and Walker JE. (**1996**) *Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels*. J Mol Biol, 260(3):289-98. **doi: 10.1006/jmbi.1996.0399**.

64. Nahalka J. (**2012**) *Glycocodon theory—the first table of glycocodons*. J Theor Biol, 307:193–204. **doi:10.1016/j.jtbi.2012.05.010**.

65. Nakahara S, Oka N, Wang Y, Hong V, Inohara H and Raz A. (**2006**) *Characterization of the nuclear import pathways of galectin-3*. Cancer Res, 66(20):9995-10006. **doi: 10.1158/0008-5472**.

66. Nangia-Makker P, Raz T, Tait L, Hogan V and Raz A. (**2007**). *Galectin-3 cleavage: a novel surrogate marker for matrix metalloproteinase activity in growing breast cancers*. Cancer Res, 67:11760–11768. **doi: 10.1158/0008-5472.CAN-07-3233**.

67. Nangia-Makker P, Wang Y, Raz T, Tait L, Balan V, Hogan V and Raz A. (**2010**) *Cleavage of galectin-3 by matrix metalloproteinase induces angiogenesis in breast cancer*. Int J Cancer, 127:2530-2541. **doi: 10.1002/ijc.25254**.

68. Nieminen J, St-Pierre C and Sato S. (**2005**) *Galectin-3 interacts with naive and primed neutrophils, inducing innate immune responses*. Leukocyte Biol, 78:1127–1135. **doi: 10.1189/jlb.1204702**.

69. Nieminen J, Kuno A, Hirabayashi J and Sato S. (**2007**) *Visualization of galectin-3 oligomerization on the surface of neutrophils and endothelial cells using fluorescence resonance energy transfer*. J Biol Chem, 282(2):1374–1383. **doi: 10.1074/jbc.M604506200**.

70. Ochieng J, Fridman R, Nangia-Makker P, Kleiner DE, Liotta LA, Stetler-Stevenson WG and Raz A. (**1994**) *Galectin-3 is a novel substrate for human matrix metalloproteinases-2 and -9*. Biochemistry, 33(47):14109-14114. **PMID: 7947821**.

71. Oda Y and Kasai K. (**1983**) *Purification and characterization of β-galactoside-binding lectin from chick embryonic skin*. Biochim Biophys Acta 761:237–245. **PMID: 6197094**.

72. Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HD, Konarev PV and Svergun DI. (2012) *New developments in the ATSAS program package for small-angle scattering data analysis*. J Appl Crystallog, 45(2):342-350. **doi: 10.1107/S0021889812007662**.

73. Putnam CD, Hammel M, Hura GL and Tainer JA. (**2007**) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys, 40:191-285. **doi: 10.1017/S0033583507004635**.

74. Rabinovich GA, Rubinstein N and Fainboim L. (**2002a**) *Unlocking the secrets of galectins: a challenge at the frontier of glyco-immunology*. J Leuk Biol, 71(5):741–752. **PMID: 11994498**.

75. Rabinovich GA, Baum LG, Tinari N, Paganelli R, Natoli C, Liu F-T and Iacobelli S. (**2002b**) *Galectins and their ligands: amplifiers, silencers or tuners of the inflammatory response?* Trends Immunol, 23(6): 313-320. **PMID: 12072371**.

76. Rabinovich GA, Hirashima M, Liu F-T and Anderson A. (**2007**) *An emerging role for galectins in tuning the immune response: lessons from experimental models of inflammatory disease, autoimmunity and cancer*. Scand J Immunol, 66:143-158. **doi: 10.1111/j.1365-3083.2007.01986.x**.

77. Rabinovich GA and Toscano MA. (**2009**) *Turning 'sweet' on immunity: galectin-glycan interactions in immune tolerance and inflammation*. Nature Reviews – Immunol, 9:338-352. **doi: 10.1038/nri2536**.

78. Rabinovich GA and Thijssen VL. (**2014**) *Galectins go with the flow*. Glycobiology, 24(10):885. **doi: 10.1093/glycob/cwu081**.

79. Ramachandran GN and Sasisekharan V. (**1968**) Conformation of polypeptides and proteins. Adv Protein Chem 23:283-438. **PMID: 4882249**.

80. Rapoport EM, Kurmyshkina OV and Bovin NV. (**2008**) *Mammalian galectins: structure, carbohydrate specificity, and functions*. Biochemistry (Mosc), 73:483–497. **PMID: 18457568**.

81. Rosano GL and Ceccarelli EA. (**2014**) *Recombinant protein expression in Escherichia coli: advances and challenges*. Front Microbiol 5(172):1-17. **doi: 10.3389/fmicb.2014.00172**.

82. Rossmann MG and Blow DM. (**1962**) *The detection of sub-units within the crystallographic asymmetric unit*. Acta Cryst 15:24-31.

83. Ruiz FM, Fernández IS, López-Merino L, Lagartera L, Kaltner H, Menéndez M, André S, Solís D, Gabius H-J and Romero A. (**2013**) *Fine-tuning of prototype chicken galectins: structure of CG-2 and structure-activity correlation.* Acta Cryst D, 69(9):1665-1676. **doi: 10.1107/S0907444913011773**.

84. Ruiz FM, Scholz BA, Buzamet E, Kopitz J, André S, Menéndez M, Romero A, Solís D and Gabius H.J. (**2014**) *Natural single aminoacid polymorphism (F19Y) in human galectin-8: detection of structural alterations and increased growth-regulatory activity on tumor cells*. FEBS J, 281:1446-1464. **doi: 10.1111/febs.12716**.

85. Ruiz FM, Gilles U, Lindner I, André S, Romero A, Reusch D and Gabius H-J. (**2015**) *Combining crystallography and hydrogen-deuterium exchange to study galectin-ligand complexes*. Chem Eur J, 21:13558–13568. **doi: 10.1002/chem.201501961**.

86. Sato S, Ouellet N, Pelletier I, Simard M, Rancourt A and Bergeron MG. (**2002**) *Role of galectin-3 as an adhesion molecule for neutrophile extravasation during streptococcal pneumonia*. J Immunol, 168:1813–1822. **PMID: 11823514**.

87. Schuck P, Zhao H, Bräutigam CA and Ghirlanda R. (**2015**) *Basic Principles of Analytical Ultracentrifugation*. CRC Press, Boca Raton, USA.

88. Seetharaman J, Kanigsberg A, Slaaby R, Leffler H, Barondes SH and Rini JM. (**1998**) *X-ray crystal structure of the human galectin-3 carbohydrate recognition domain at 2.1-Å resolution*. J Biol Chem, 273(21): 13047–13052. **doi: 10.1074/jbc.273.21.13047**.

89. Solís D, Maté MJ, Lohr M, Ribeiro JP, López-Merino L, André S, Buzamet E, Cañada FJ, Kaltner H, Lensch M, Ruiz FM, Haroske G, Wollina U, Kloor M, Kopitz J, Sáiz JL, Menéndez M, Jiménez-Barbero J, Romero A and Gabius H-J. (**2010**) *N-domain of human adhesion/growth-regulatory galectin-9: preference for distinct conformers and non-sialylated N-glycans and detection of ligand-induced structural changes in crystal and solution*. Int J Biochem Cell Biol, 42(6):1019–29. **doi: 10.1016/j.biocel.2010.03.007**.

90. Solís D, Bovin NV, Davis AP, Jiménez-Barbero J, Romero A, Roy R, Smetana K and Gabius H-J. (**2015**) *A guide into glycosciences: how chemistry, biochemistry and biology cooperate to crack the sugar code*. Biochim Biophys Acta, 1850:186-235. **doi: 10.1016/j.bbagen.2014.03.016**.

91. Sorme P, Arnoux P, Kahl-Knutsson B, Leffler H, Rini JM and Nilsson UJ. (**2005**) *Structural and thermodynamic studies on cation-Pi interactions in lectin-ligand complexes: high-affinity galectin-3 inhibitors through fine-tuning of an arginine-arene interaction*. J Am Chem Soc, 127(6):1737-1743. **doi: 10.1021/ja043475p**.

92. Svergun DI. (**1992**) *Determination of the regularization parameter in indirect-transform methods using perceptual criteria*. J Appl Crystallog, 25(4):495-503. **doi: 10.1107/S0021889892001663**.

93. Teichberg VI, Silman I, Beitsch DD and Resheff G. (**1975**) *A β-D-galactoside binding protein from electric organ tissue of E. electricus*. PNAS, 72(4):1383-1987.

94. Than NG, Romero R, Kim CJ, McGowen MR, Papp Z and Wildman DE. (**2012**) *Galectins: guardians of eutherian pregnancy at the maternal-fetal interface*. Trends Endocrin Metab 23(1):23-31. **doi: 10.1016/j.tem.2011.09.003**.

95. Thiemann S and Baum LG. (**2016**) *Galectins and immune responses — just how they do those things they do*. Annu Rev Immunol, 34:243-264. **doi: 10.1146/annurev.immunol.041015-055402**.

96. Thijssen VL, Rabinovich GA and Griffioen AW. (**2013**) *Vascular galectins: regulators of tumor progression and targets for cancer therapy*. Cytok Growth Fact Rev, 24(6):547–558. **doi: 10.1016/j.cytogfr.2013.07.003**.

97. Umemoto K, Leffler H, Venot A, Valafar H and Prestegard JH. (**2003**) *Conformational differences in liganded and unliganded states of Galectin-3*. Biochemistry, 42(13):3688–3695. **doi: 10.1021/bi026671m**.

98. Vagin A, Teplyakov A. (**2010**) Molecular replacement with MOLREP. Acta Cryst D Biol Crystallog, 66(1):22-25. **doi: 10.1107/S0907444909042589**.

99. Varela PF, Solís D, Díaz-Mauriño T, Kaltner H, Gabius H-J, Romero A. (**1999**) *The 2.15 Å crystal structure of CG-16, the developmentally regulated homodimeric chicken galectin*, J Mol Biol, 294(2):537–549. **doi: 10.1006/jmbi.1999.3273**.

100. Varki A, Cummings RD, Esko JD, editors. (**2009**) *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, New York, online version.

101. Vasta GR, Ahmed H, Nita-Lazar M, Banerjee A, Pasek A, Shridhar S, Guha P and Fernández-Robledo JA. (**2012**) *Galectins as self/non-self recognition receptors in innate and adaptive immunity: an unresolved paradox*. Front Immunol, 3(199):1-14. **doi: 10.3389/fimmu.2012.00199**.

102. Voet D, Voet J and Pratt C. (**2013**) *Principles of Biochemistry*. Wiley & sons, Hoboken, pp 236-240.

103. Wälti MA, Thore S, Aebi M and Künzler M. (**2008a**) *Crystal structure of the putative carbohydrate recognition domain of human galectin-related protein*. Proteins, 72(2):804–808. **doi: 10.1002/prot.22078**.

104. Wälti MA, Walser PJ, Thore S, Grunler A, Bednar M, Künzler M and Aebi M. (**2008b**) *Structural basis for chitotetraose coordination by CGL3, a novel galectin-related protein from* Coprinopsis cinerea. J Mol Biol, 379:146-149. **doi: 10.1016/j.jmb.2008.03.062**.

105. Watanabe M, Nakamura O, Muramoto K and Ogawa T. (**2012**) *Allosteric regulation of the carbohydrate-binding ability of a novel conger eel galectin by d-mannoside*. J Biol Chem, 287:31031-31072. **doi: 10.1074/jbc.M112.346213**.

106. Wright DE, Bowman EP, Wagers AJ, Butcher EC and Weissman IL. (**2002**) *Hematopoietic stem cells are uniquely selective in their migratory response to chemokines*. J Exp Med 195:1145–1154. **PMID: 11994419**.

107. Yang RY, Rabinovich GA and Liu F-T. (**2008**) *Galectins: structure, function and therapeutic potential*. Expert Rev Mol Med, 10:e17. **doi: 10.1017/S1462399408000719**.

108. Yoshida H, Teraoka M, Nishi N, Nakakita S, Nakamura T, Hirashima M and Kamitori S. (**2010**) *X-ray structures of human galectin-9 C-terminal domain in complexes with a biantennary oligosaccharide and sialyllactose*. J Biol Chem, 285:36969-36976. **doi: 10.1074/jbc.M110.163402**.

109. Zhou D, Ge H, Sun J, Gao Y, Teng M and Niu L. (**2008**) *Crystal structure of the C-terminal conserved domain of human GRP, a galectin-related protein, reveals a function mode different from those of galectins*. Proteins, 71(3):1582–1588. **doi: 10.1002/prot.22003**.