

Hunting down hydrogen: applying neutron macromolecular crystallography to galectins

Derek T. Logan

(Lund University, Sweden)

Galectins are a medically important family of small, soluble carbohydrate-binding proteins that have been implicated in a wide variety of diseases, and are therefore targets for the development of new drugs. Many X-ray crystal structures of small molecules in complex with galectins have been determined and have given invaluable information on how natural ligands and synthetic inhibitors are recognized. However, until now the hydrogen-bonding patterns involved in ligand recognition have been a matter of informed guesswork, as X-ray crystallography can't reliably tell us where hydrogen atoms are positioned. We describe how neutron macromolecular crystallography has been applied to galectins for the first time to shed light on this question.

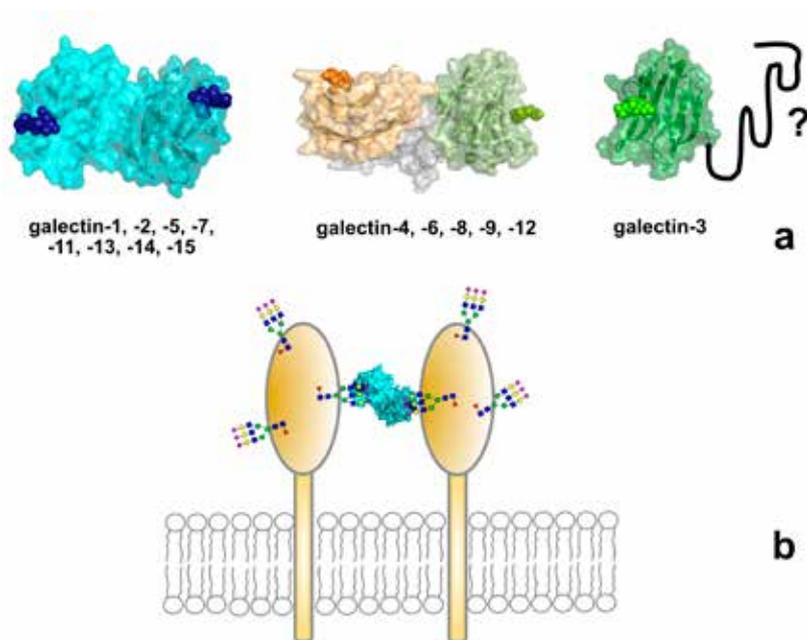


Figure 1. Galectin subfamilies **a**) The different families of galectin. The carbohydrate recognition domains (CRDs) are shown in surface representation. Spheres indicate the centre of the carbohydrate-binding site. Colours represent different binding specificities. The linker in the galectin-4-like family and the N-terminus of galectin-3 are represented as lines as their exact conformation is unknown. However, it is known that the N-terminus of galectin-3 can have local order and interact transiently with the CRD. **b**) A homodimeric galectin (e.g., galectin-1) bringing together two glycosylated cell-surface proteins, which may result in a signal being transmitted into the cell, or alternatively the blocking of a signal.

Galectins and their role in disease

Galectins are small, soluble proteins with many functional roles, the majority of which involve binding to β -D-galactoside-containing sugar moieties on the surface of other proteins. Galectins are synthesized in the cytoplasm and spend most of their time either there or in the nucleus. However, they are secreted by a non-classical pathway to the extracellular space, where they exert many of their biological effects via the sequestration of proteins through glycosylation and the formation of glycoprotein lattices. Myriad biological functions have been attributed to galectins, and deciphering these is the subject of intensive study, with publications implicating galectins in one way or another appearing almost daily. For more detail, we recommend the excellent recent overview by Johannes et al. (see 'Further reading') and publications therein.

To date, 15 galectins have been found in mammals. These fall into three subfamilies (Figure 1): a) those having a single type of carbohydrate recognition domain (CRD; e.g. galectins -1, -2 and -7), which are largely monomeric but which dimerize at higher concentrations, b) galectins in which two different CRDs are joined in tandem in the same polypeptide (galectins -4, -6, -8, -9 and -12) and c) galectin-3, which is unique in having a largely disordered N-terminal segment that may modulate self-association. Because of their different make-up and specificity, these three families can mediate the formation of different kinds of lattice: homotypic or heterotypic.

Structure of galectins

Galectins are characterized by a common CRD of about 150 amino acids that folds into a β -sandwich (Figure 2). One face of the sandwich has loops that form a shallow, hydrophilic groove in which the carbohydrate recognition site is found. This site can be divided into five subsites, A–E. The groove is lined with largely hydrophilic amino acids that contribute hydrogen bonds to hydroxyl groups on the bound carbohydrate.

Drug design against galectin-3

In recent years, human galectin-3 has emerged as an interesting drug target, owing to its involvement in various diseases, including inflammation, cancer proliferation and metastasis, and diabetes. The design of inhibitors of galectin function is an area of considerable activity, particularly against galectin-3. For example, the galectin-3 inhibitor TD139 is in phase II clinical trials for the treatment of idiopathic pulmonary fibrosis, and another, GB1107, is showing promising results against lung adenocarcinoma. All successful inhibitors to date, including TD139, have been based on a central mono- or disaccharide building block (Figure 2). In one of the natural substrates, lactose, the disaccharide consists of galactose followed by glucose. Such ligands have very weak affinities, in the mM range, but this has been increased to single digit nM by adding non-sugar groups to the ends of the disaccharide core without modifying the core itself. As a contribution to drug design against galectins, we wanted to know how specific the hydrogen-bonding patterns to galectin-3 are in a disaccharide. If some were found to be relatively non-specific, we could imagine different building blocks for the central core, or different substituents on the sugar rings to the usual hydroxyl groups, as these make the sugar moieties polar and can lower the oral bioavailability of candidate drugs. Furthermore, experimentally determined hydrogen atom positions would provide an unambiguous foundation for future theoretical calculations on ligand binding.

X-ray crystallography and its blind spot for the smallest of atoms

Macromolecular crystallography using X-rays (MX) is the dominant technique for probing the structures of proteins and other biological macromolecules. MX has contributed over 133,000 structures, or around 90% of all depositions in the structural biologists' favourite repository, the Protein Data Bank. However, MX has one major achilles heel, and that is in visualizing hydrogen atoms. These

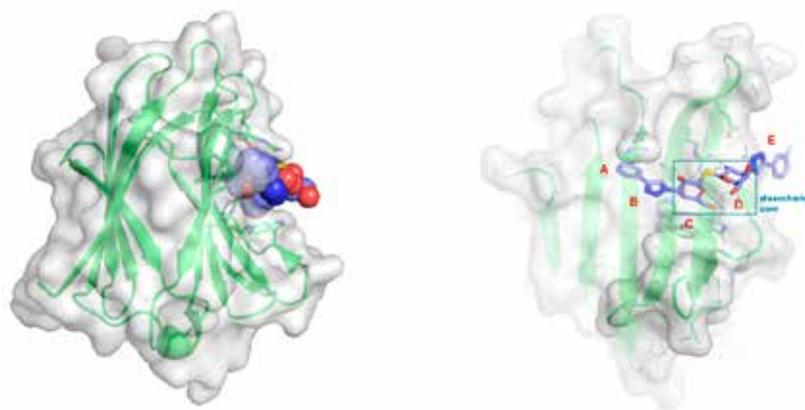


Figure 2. Two views of the C-terminal carbohydrate-binding domain (CRD) of galectin-3C in complex with the drug candidate TD139. The left-hand view shows the β -sandwich and the binding groove formed by loops on one face. The right-hand view shows the five potential binding subsites, of which C and D are the central ones. Hydrogen bonds between TD139 and the CRD are shown as dotted lines.

lightest of atoms make up about half of those in a protein. In fairness, it must be admitted that most of them are relatively uninteresting, being unreactive hydrogen atoms attached to parts of a protein far from the site of interest. However, knowledge of the positions of a few hydrogen atoms can be critical to understand an enzymatic reaction mechanism, or as in our case, the recognition of a designed inhibitor in the process of drug design. These are usually the polar hydrogens, exchangeable with solvent, often mobile and harder to predict the location of.

How can we hunt down hydrogen?

This is where neutron macromolecular crystallography (NMX) comes into its own. MX exploits the scattering of X-rays by the electrons in atoms, proportional to their atomic number. Hydrogen has only one electron, whereas, for example, carbon has six. Thus, while it is relatively easy to 'see' the other typical atoms in a protein – carbon, nitrogen, oxygen and sulphur – hydrogens are essentially invisible. They begin to show up at higher resolution (defined as how much detail we can see in a crystal structure, limited by how well-ordered our crystals are), but even at the highest resolutions obtained (for less than 1% of all structures) we can at best only see about half of them, and unfortunately these tend to be the less interesting ones mentioned above. In contrast, neutron crystallography involves scattering of neutrons by atomic nuclei, with no obvious relation to atomic number. In effect, hydrogen nuclei, particularly deuterium, become just as 'visible' as C, N or O, even at limited resolution.

Only 0.1% of the structures in the PDB have been determined using NMX, despite the method having been around for several decades. Why is it not more popular? Both MX and NMX involve focusing a beam onto a crystalline sample. The diffraction of X-rays or neutrons from that sample gives us (after some computational work) the structure. The difficulties of NMX are mainly due to the very low flux of neutron sources compared with modern X-ray sources, such as synchrotrons, and the requirement to compensate for this by making very large samples, usually at least 1 mm³. This may not seem like a big effort, but modern biochemists are more used to making crystals at least a thousand times smaller in volume. However, the NMX community is slowly growing, with half of all the structures deposited over the past 3 years equivalent to the total output for the previous 30 years. For further details on the technology and history of NMX, the reader is directed to the excellent 2014 article in *The Biochemist* by Matthew Blakeley (see 'Further reading').

NMX has become more accessible

Happily, our study benefited from recent developments, both in the wet lab and at the neutron source. We were able to exploit the fact that perdeuteration of the sample, i.e., replacement of every hydrogen atom in the protein with deuterium, significantly improves the signal-to-noise ratio in NMX by reducing so-called 'incoherent scattering'. In addition, the emergence of so-called spallation sources and improved instrumentation at reactor-based sources has led to an improvement in the neutron flux in recent years. Both of these developments meant that we were able to use smaller crystals or reduce the experimental time for a crystal of a given size.

Our first crystals of perdeuterated galectin-3C (the C-terminal carbohydrate recognition domain of galectin-3), at 0.4 mm³ were gigantic by X-ray standards, but were still relatively small for neutron work, and our first data took over 2 weeks to collect. This can be compared to a few seconds for an X-ray dataset at the same resolution. We painstakingly enlarged our crystals to about four times that volume, finding new ways to keep them growing, dragging them around the world, when ready, to neutron facilities at Los Alamos, Oak Ridge, Grenoble and Munich, until the point where we could collect data of high quality in only a few days. An example of one of the beautiful 'pseudo-Laue' diffraction patterns typical of NMX is shown in Figure 3.

The fruits of these efforts, so-called 'nuclear density maps', clearly showed the positions of all the important polar hydrogen atoms on the natural substrate, lactose, and how they are oriented towards the protein (Figure 4a). Each one of these had a very well-defined geometry, demonstrating that the arrangement of protein atoms

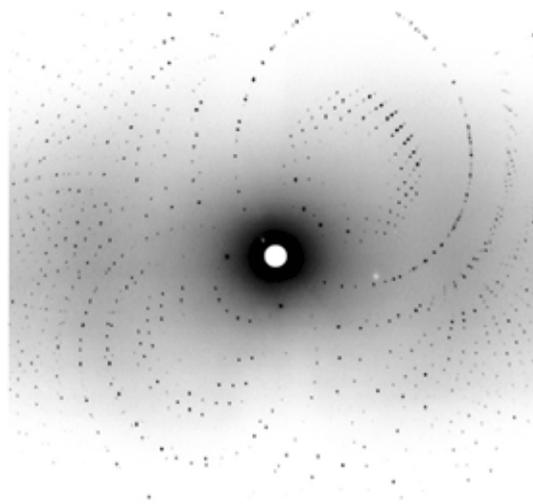


Figure 3. A typical diffraction pattern from galectin-3C, a collection of which can, with the help of computer calculations, reveal the positions of hydrogen atoms in the structure. These data were collected at the LADI-III instrument of the ILL in Grenoble.

in the binding site is extremely intricately engineered by nature to recognize the exact pattern of hydroxyl groups on a disaccharide. Thus, it may be difficult to find anything non-sugar-like in our search for new molecular scaffolds that would do the job just as well, and drug design is wise to remain focused on improving interactions outside that area.

NMX is also very powerful for visualizing the exact hydrogen-bonding patterns of water molecules. As the hydrogen atoms are just as 'visible' as C, N or O, well-ordered water molecules have a characteristic 'boomerang' shape in nuclear density maps, revealing exactly how they H-bond. In an electron density map from X-ray crystallography, by contrast, one can only see water molecules as spheres, and H-bonding geometry has to be guessed. Our results showed that, when there was no lactose bound to galectin-3C, three water molecules that imitate the hydroxyl groups on lactose generally obeyed the same H-bonding patterns, but only one of them had a clear 'boomerang' shape due to the higher mobility of those water molecules (Figure 4b).

A future challenge is to study questions of specificity between galectins. Designed inhibitors often present quite different affinities towards galectin-1 and galectin-3, despite the fact that most of the key residues in the central binding site are identical. Thus differences in affinity could be related to water-mediated hydrogen-bonding networks involving non-conserved residues. The selectivity of inhibitors for galectin-1 over galectin-3 is important, because they, as the two proteins, have complex but specific phenotypes in inflammation and oncology and the function of one should not be disturbed by inhibition of the other.

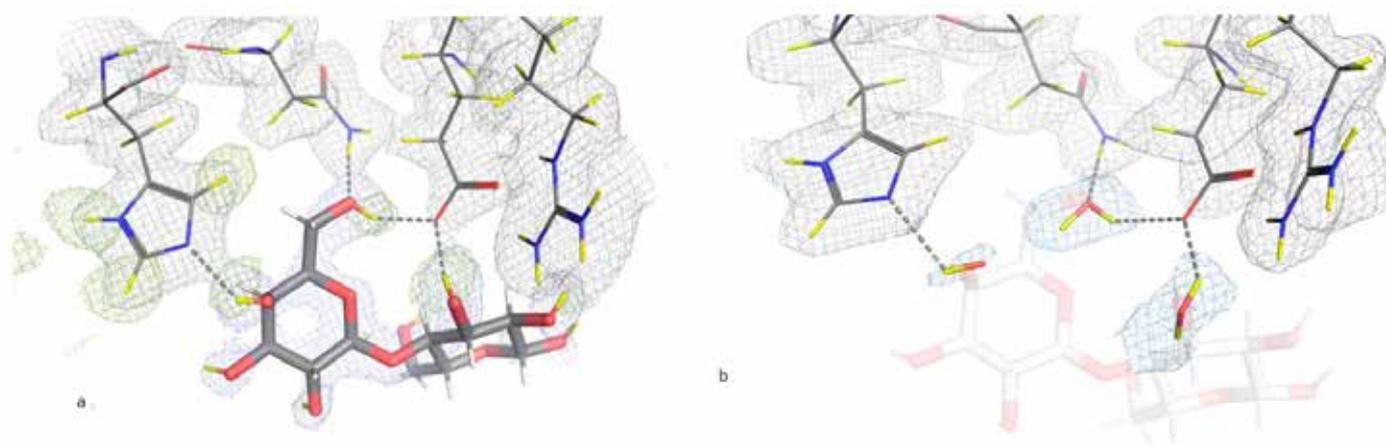


Figure 4. Nuclear density maps that show the positions of key hydrogen atoms in galectin-3 **a**) In complex with lactose and **b**) in the protein on its own. The grey mesh reveals the overall structure while the green blobs show the exact positions of hydrogens. An unusual protonation state was confirmed for the histidine on the left. In the right-hand panel, water molecules take the place of key hydroxyl groups on the inner-facing side of lactose, which is shown as a 'ghost' for comparison.

In summary, our first adventures in neutron crystallography have provided valuable information on hydrogen bonding in molecular recognition in galectins and pave the way for a deeper investigation of the whole family.

Acknowledgements

This work would not have been possible without the efforts of many colleagues, especially my former PhD student Francesco Manzoni, who sadly passed away in 2017 at the age of 27. We are also grateful to our colleagues at the Institut Laue Langevin, Oak Ridge National Laboratory, Los Alamos National Laboratory and Meier-Leibniz Zentrum. Finally, we thank Joane Rustiguel for the coordinates of galectin-4 used in Figure 1. ■

Further reading

- Johannes, L., Jacob, R. and Leffler, L. (2018) Galectins at a glance. *J. Cell. Sci.* **131**: jcs208884. (Contains an excellent 'poster' summarizing galectin structure and function.)
- Blakeley, M.P. (2014) Macromolecular crystallography using neutrons. *The Biochemist*. <http://www.biochemist.org/bio/03603/0040/036030040.pdf>
- Manzoni, F., Wallerstein, J., Schrader, T.E. et al. (2018) Elucidation of hydrogen bonding patterns in ligand-free, lactose- and glycerol-bound galectin-3C by neutron crystallography to guide drug design. *J. Med. Chem.* **61**, 4412–4420



Derek Logan is an associate professor in Biochemistry and Structural Biology at Lund University in Sweden. He graduated from Glasgow University in 1988 with a BSc Hons (first class) degree in Chemistry with Computer Applications. Derek then carried out his DPhil studies in virus crystallography at Oxford University, supervised by David I. Stuart, achieving his degree in 1992. From there, Derek moved to the IBMC du CNRS in Strasbourg, France, and determined the first structure of a glycyI-tRNA synthetase. Following a further postdoc at Stockholm University, Derek has been at Lund University since 2001. For 15 years he was a part-time beamline scientist at the MAX-lab synchrotron. In 2006, he co-founded and is still the Chief Scientific Officer of the structural biology contract research organization SARomics Biostructures. Email: derek.logan@biochemistry.lu.se