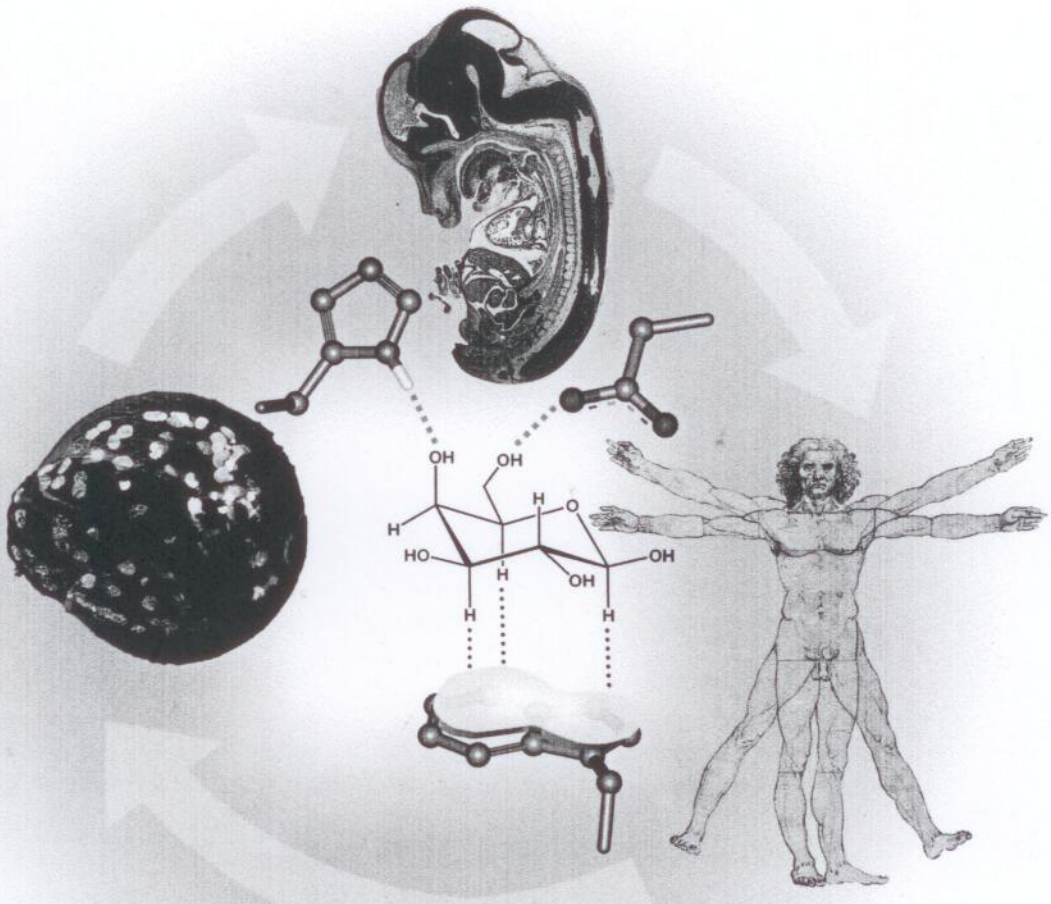
Edited by Hans-Joachim Gabius

WILEY-BLACKWELL

# The Sugar Code

## Fundamentals of Glycosciences



# The Sugar Code

Fundamentals of Glycosciences

*Edited by*
*Hans-Joachim Gabius*

WILEY-
BLACKWELL

WILEY-VCH Verlag GmbH & Co. KGaA

## *Further Reading*

Demchenko, A. V. (ed.)

**Handbook of Chemical Glycosylation**

**Advances in Stereoselectivity and Therapeutic Relevance**

524 pages in 1 volumes with 492 figures and 14 tables
2008
Hardcover
ISBN: 978-3-527-31780-6

Wong, C.-H. (ed.)

**Carbohydrate-based Drug Discovery**

980 pages in 2 volumes with 296 figures and 42 tables
2003
Hardcover
ISBN: 978-3-527-30632-9

Ernst, B., Hart, G. W., Sinaÿ, P. (eds.)

**Carbohydrates in Chemistry and Biology**

2578 pages in 4 volumes with 949 figures and 111 tables
2000
Hardcover
ISBN: 978-3-527-29511-1

The

Funda

Editea
Hans-

WILE
BLACK

WILEY

## Contents



*The Sugar Code.* Edited by Hans-Joachim Gabius

ISBN: 978-3-527-32089-9

## Part Five Biomedical Aspects and Case Studies

# 20
# Routes in Lectin Evolution: Case Study on the C-Type Lectin-Like Domains

*Jill E. Gready and Alex N. Zelensky*

The preceding chapters have introduced a common protein fold found in animal lectins, that is the C-type lectin domain. As illustrated in Figure 16.1h, which shows the X-ray structure of this fold in human P-selectin (for further details on selectins and their functions, see Chapters 19 and 27), the fold is characterized by a unique combination of individual features suited for stabilizing the structure and the primary sugar-binding site. The term originates from the essential presence of a $Ca^{2+}$ which interacts with sugar ligands via coordination bonds. The domain shows ligand preferences for different types of carbohydrates, but also noncarbohydrate ligands. This has led to general use of the term C-type lectin-like domain (CTLD). As the proteins containing the domain (CTLD-containing proteins (CTLDcps)) comprise a large heterogeneous superfamily with diverse functions, they are an excellent model to study its evolution. Our method of analysis is also relevant to other lectin folds (given in Chapters 16 and listed in Tables 18.2 and 19.1), and to glycosyltransferases and other multimember glycogene groups (given for example in Chapters 6 and 7, with relevance for disease in Tables 22.1 and 23.1).

In this chapter we will show how systematic comparative analysis and data mining have created a strong framework for deciphering the complex functions of CTLDcps and studying their evolution. First, linking of protein sequence with three-dimensional structure provided a framework for interpretation of sugar specificity and binding. Second, tracking how these binding mechanisms have evolved in model organisms such as worm and human has revealed the exceptional capacity of the CTLD fold to evolve new specificities and functions. Definition of the full repertoire of CTLDcps in evolutionary branches by whole-genome sequence analysis has 'laid open' the field, posed unexpected questions and provided novel directions for further study.

## 20.1
## C-Type Lectin (CTL) Evolution as a Case Study

In this chapter we will illustrate how the evolution of C-type lectins [1,2] (see Chapters 16, 19 and 27 for further information) demonstrates key principles of

*The Sugar Code*. Edited by Hans-Joachim Gabius

ISBN: 978-3-527-32089-9

routes for diversification, using these proteins as an exciting model. The following issues will be addressed in a stepwise manner, emphasizing how these insights have been obtained.

- How a protein scaffold – in this case the CTLD [3, 4] – with superior stability and versatility has the capacity to adapt to bind many types of ligands – not only carbohydrate – specifically. This allows it to be very successful evolutionarily, in the case of the CTLD becoming one of the most abundant protein domains in multicellular animals (Metazoa) [5–8].
- How this abundance of CTLDcps has developed differently in major animal lineages, such as worm [5], fly [6] and vertebrates [7, 8]. This has apparently occurred by independent evolution of novel CTLDcps with lineage-adapted functions, by recruiting the CTLD into the new proteins, usually in combination with other domains.
- How the likely earliest functions of C-type lectins as sugar-binding proteins as part of an innate immune response have been preserved in all animal lineages studied so far.
- How application of systematic whole-genome analyses can provide a sudden revelation of the extent of the CTLDcps repertoire in particular animal lineages, and how this can transform thinking and approaches to defining functions.
- How Nature is always full of surprises, with recent reports of CTLDcps in pre-Metazoa [9], and non-Metazoa, including plants [10], requiring a rethink of the earliest origin of the 'C-type lectin' with sugar-binding function.
- How generation of this knowledge has come from scientists working from diverse perspectives – genes and proteins, structure and function, genomics and proteomics, model organisms, and individual species.
- How this richness of acquired knowledge can be integrated into some pithy lessons on the evolution of C-type lectins, which also points the way to how evolution of other lectin domains might be studied.

## 20.2 CTL Superfamily: Structures and Groups

C-type lectins comprise a large and heterogeneous superfamily of proteins containing CTLDs (Figure 20.1). They have diverse functions and are among the first animal lectins discovered, with conglutinin being detected in 1906, as listed in Table 15.1 which gives a historical overview of lectinology. Until recently, they were regarded as exclusively extracellular proteins that originated in the Metazoan era to fulfill new needs in multicellular organisms for intercellular communication, immune defense and response. Carbohydrate binding is the most common CTLD function in vertebrates, and in this role C-type lectins function either as membrane-anchored or extracellular-matrix proteins in adhesion or immune

**Figure 20.1** A typical CTL
cartoon representation.
which harbors the prima
binding site (position arr
Disulfide bridges are sh

defense, intercellular
lism (see Chapters 19
be the ancestral func
defense CTLDcps cha
CTLDs have evolved
proteins, but also lip
proteins and calcium

As summarized in
capable of binding ca
C-type referring to th
site and distinguish
dependent lectins (fo
lectin-like domain' – r
it became clear that t
greater functionality t
(please see Chapters
tion domains (CRDs)
also bind proteins (se
19.3).

In Chapter 19 we
design (see Figure 1

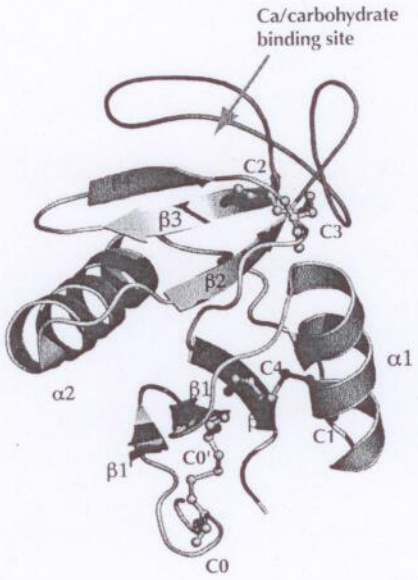



**Figure 20.1** A typical CTLD structure shown in cartoon representation. The long-loop region which harbors the primary $Ca^{2+}$/carbohydrate-binding site (position arrowed) is shown in red. Disulfide bridges are shown as yellow sticks. The archetypical disulfide bridges of the fold are between C1 and C4, and C2 and C3, whereas that between C0 and C0′ is present only in long-form CTLDs which have an N-terminal extension.

defense, intercellular communication and integration, and glycoprotein metabolism (see Chapters 19 and 27 for details). Carbohydrate binding is also thought to be the ancestral function of the superfamily, as evidenced by the many humoral defense CTLDcps characterized in insects and other invertebrates. However, many CTLDs have evolved to recognize ligands other than carbohydrates, particularly proteins, but also lipids and inorganic substances such as ice in fish antifreeze proteins and calcium carbonate in bird egg-shell proteins.

As summarized in Chapters 16, 18 and 19, there are a number of protein folds capable of binding carbohydrate, that is 'lectin domains'. The CTLD is one of them, C-type referring to the presence of a calcium ion in the main carbohydrate binding site and distinguished by its sequence signature from other classes of $Ca^{2+}$-dependent lectins (for an overview, see Chapter 16). The nomenclature 'C-type lectin-like domain'–rather than just 'C-type lectin domain'–was introduced when it became clear that this domain could bind other ligands [3]. The CTLD, thus, has greater functionality than the lectin domains characteristic of the other lectin types (please see Chapters 18 and 19). They function primarily as carbohydrate recognition domains (CRDs) although I- and P-type lectin-like domains and galectins can also bind proteins (see Chapter 19; protein ligands of galectins are listed in Table 19.3).

In Chapter 19 we saw that a salient feature of many lectins is their modular design (see Figure 19.1). This feature is frequently encountered in CTLDcps. In

principle, a domain has a compact three-dimensional structure that may fold independently of the rest of the protein, and may function and evolve semi-independently. The sequence of a protein domain is typically 100–200 amino acids long. A given CTLDcp by definition contains an operative lectin domain, but in addition may contain multiple diverged copies of it with different sugar or other ligand-binding specificity as well as other domains. These features are shown in Figure 20.2.

The development of the C-type lectin field benefited greatly from the early efforts of researchers, notably Kurt Drickamer, to systematize the disparate biological data [1, 2]. This led to their characterization as lectins with CRDs of length 110–140 residues, which bound carbohydrates in a $Ca^{2+}$-dependent manner. Furthermore, alignment of their sequences showed this protein domain had a characteristic conserved sequence signature. This feature has been of enormous value in advancing C-type lectin research as it has allowed initial identification as a potential CTLDcp directly from analysis of the protein sequence. A further helpful finding from this early sequence analysis was that carbohydrate specificity is often correlated with a particular tripeptide sequence motif within the sequence signature – **EPN** (Glu-Pro-Asn) for mannose-type ligands and **QPD** (Gln-Pro-Asp) for galactose-type ligands. This permits initial functional predictions.

Understanding of the structural basis of the sequence signature came with solution of the first X-ray structure of a CTLD – of rat mannan-binding protein (MBP)-A – in 1991 and a little later by a structure of this CTLD with a bound mannose molecule. Comparison of the residues then identified as the sequence signature – 12 totally conserved and 18 conservatively conserved – against the three-dimensional structure allowed the roles for most of them in $Ca^{2+}$- and carbohydrate-binding and stabilizing the protein fold to be defined [1].

The basic protein fold is shown in Figure 20.1 and the main details of carbohydrate binding in Figure 20.3. The key characteristics of the CTLD fold are two antiparallel β-sheets and two α-helices, and two disulfide bridges and a hydrophobic core stabilizing the long-loop region which contains the primary sugar-binding site. Examination of the large number of CTLD structures that have been obtained by crystallography shows that this unique loop-in-a-loop structure, in which the large flexible long-loop region is maintained on a stable core, allows the fold to tolerate substantial variation in the shape of the primary ligand-binding site and adjacent regions [4]. This allows specific binding of large multivalent ligands such as complex-type oligosaccharides and mannose-rich structures (triantennary *N*-glycans and mannans are examples given in Chapter 19), non-carbohydrate ligands, and even both. Formation of quaternary complexes, for example in the trimers of the group III vertebrate collectins (for explanation of this term, see Chapter 19), further increases both specificity and affinity of carbohydrate recognition by C-type lectins [3].

In 1993, Drickamer classified mammalian CTLDcps into seven groups (I–VII). This was based primarily on their domain architecture, but the grouping also appeared to reflect evolutionary history as it correlated well with the results of phylogenetic analysis of the CTLD sequences [1]. The classification was revised in

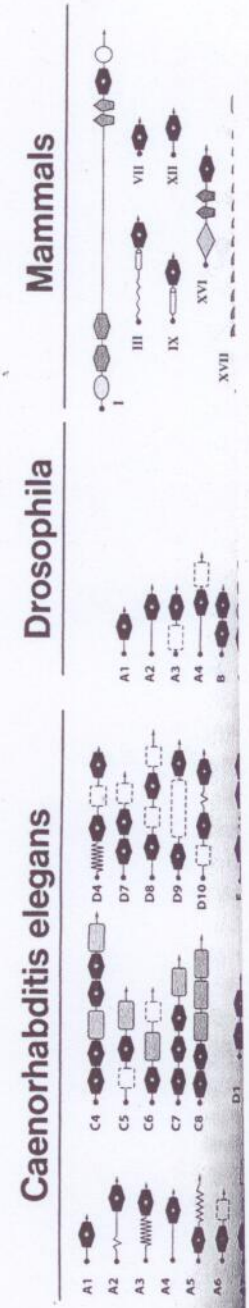


**Figure 20.2** Summary of domain architectures in CTLDcps of worm [5], fly [6] and mammals [7, 8] as revealed by whole-genome analysis.

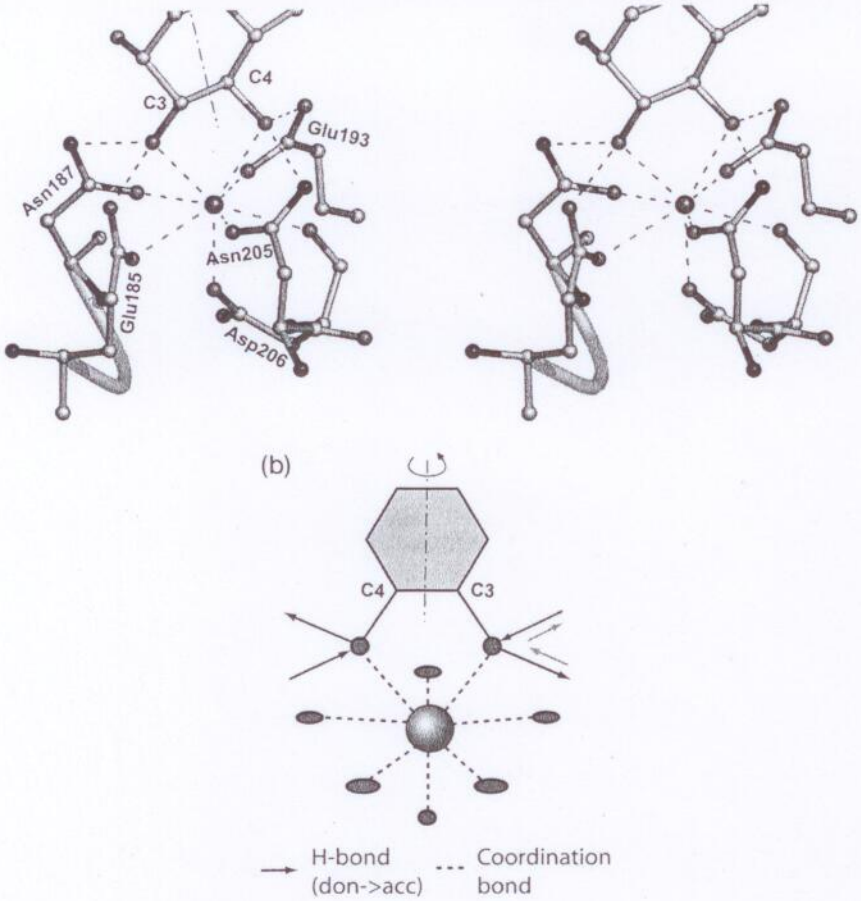



**Figure 20.3** $Ca^{2+}$-dependent monosaccharide binding by CTLDs. (a) Stereo depiction of the structure of rat MBL (formerly called mannose-binding protein, MBP-A) complexed with $Ca^{2+}$ (blue sphere) and mannose. The coordination bonds to $Ca^{2+}$ are shown in orange. Hydrogen bonds where the 3′ and 4′ sugar hydroxyls act as acceptor and donor are shown as red and blue, respectively. (b) Schematic representation of a $Ca^{2+}$–pyranose–CTLD complex. Two hydroxyl oxygens and the pyranose ring are shown with the $Ca^{2+}$ as a large blue sphere, and oxygens as red circles and ovals. Arrows show the direction of hydrogen bonds in mannose-specific CTLDs, while light-grey arrows indicate the changed directions in galactose-specific CTLDs. The C3 and C4 atoms of the sugar, and the orientation of a rotation axis are shown in both (a) and (b) (please see also Figures 16.1h,i).

2002 with the addition of seven new groups (VIII–XIV) found experimentally in the interim [7]. Subsequently, a further three groups (XV–XVII) have been added based on findings from whole-genome analysis [8]. This classification has facilitated prediction of the oligomerization and ligand-binding properties for newly found members [3]. The domain architectures of the mammalian CTLDcps are

e pyranose ring are
arge blue sphere, and
d ovals. Arrows show
bonds in mannose-
t-grey arrows indicate
in galactose-specific
oms of the sugar, and
ion axis are shown in
se see also Figures

d experimentally in
II) have been added
ification has facili-
roperties for newly
dian CTLDcps are

summarized in Figure 20.2. This shows the variation in the number and order of CTLD and non-CTLD protein domains in multidomain CTLDcps, as well as whether they are secreted or anchored in the membrane.

## 20.3 Mechanism of Carbohydrate Binding

In this section we will examine how crystal structures of CTLDs with bound sugar ligands have revealed a general molecular mechanism of carbohydrate binding at the main binding site. Specifically we will look at the roles of the $Ca^{2+}$ and two groups of residues (the 'EPN'/'QPD' and 'WND' motifs) in forming the binding site that interacts with the sugar, and at how the **EPN** and **QPD** motifs discriminate mannose-type and galactose-type sugars, respectively. We confine our attention to this main site but note that some C-type lectins, such as the vertebrate group IV selectins (Chapters 19 and 27), bind additional monosaccharide units of oligosaccharide ligands such as the sialylated and sulfated Lewis$^x$ epitope (see Tables 7.4 and 27.2 for structures) at auxiliary sites of the CTLD (see Figure 16.1h). We also note that it is this site that has been modified in many non-carbohydrate-binding CTLDs to select other ligands (protein, ice, calcium carbonate), reinforcing our developing view that this site is unusually adaptable by evolution.

The architecture of the primary monosaccharide-binding site for mannose-type ligands is illustrated in Figure 20.3a. This is based on the crystal structure for the CTLD of rat MBP-A in complex with the *N*-glycan $Man_6$-$GlcNAc_2$-Asn; see Chapters 6 and 8 for *N*-glycan structures and Figure 1.6 for structures of monosaccharides. The complex is stabilized by a network of coordination and hydrogen bonds. Oxygen atoms from the 3′ and 4′-hydroxyls of the mannose form two coordination bonds with the $Ca^{2+}$ ion and four hydrogen bonds with residues–Glu185 and Asn187 (**EPN**), Asn205 (**WND**) and Glu193–whose carbonyl side-chains coordinate the $Ca^{2+}$-binding site. This bonding pattern is fundamental for CTLD/$Ca^{2+}$/monosaccharide complexes, and is observed in all known structures. The **EPN** and **WND** motifs are in the long-loop region and β4 strand of the CTLD structure, respectively (Figure 20.1). Asp206 of the **WND** motif contributes another $Ca^{2+}$ coordination bond, while the Trp204 residue is highly conserved and contributes to the hydrophobic core [1, 4].

The arrangement of the hydrogen-bond donors and acceptors and coordination bonds in the binding site, as summarized in Figure 20.3b, has two important features. First, it determines the overall positioning and orientation of the sugar in the binding site. However, as shown in Figure 20.3b, the site has a 2-fold symmetry axis relating the sugar hydroxyls which would allow the sugar to be rotated by 180° without introducing any changes to the bonding scheme. Indeed, examples are now known. The structure of the rat MBP-C complex with mannose showed this hexapyranose bound in the opposite orientation. Also, structures of a galactose-binding mutant of MBP-A and CEL-I, a C-type lectin from the echinoderm sea cucumber, showed galactose bound in the opposite orientation

to that observed in the complex of 'TC-14', a lectin found in a tunicate urochordate [2].

Second, constraints imposed by the structure of the $Ca^{2+}$-coordination site determine the properties of the carbohydrate hydroxyls that the site can accept. This is best demonstrated by the mechanism by which the CRD discriminates between the mannose-type and galactose-type monosaccharides. Crystallographic analysis of the galactose-specific MBP-A mutant in which the **EPN** motif was mutated to **QPD** showed little restructuring of the $Ca^{2+}$-binding site, suggesting that the key switch for specificity was swapping the hydrogen-bond donor and acceptor across the monosaccharide-binding plane. This changed the hydrogen-bonding pattern from asymmetrical mannose-type (Figure 20.3b; dark-grey arrows) to symmetrical galactose-type (light-grey arrows). The theory is nicely supported by the finding of the same hydrogen-bond distribution in the structure of the TC-14 lectin complex with galactose, even though the details are rather complex as the TC-14 CRD contains an unusual **EPS** (not **QPD**) motif [2].

Although many of the determinants of monosaccharide-binding specificity have been established experimentally by numerous examples, a convincing explanation of the underlying mechanism is still wanting. Although mutual spatial disposition of bonded hydroxyls was initially suggested to be the main determinant of specificity, a growing number of crystal structures of CRDs with the MBP-A-like ('asymmetrical') distribution of hydrogen bonds have shown the binding site is compatible with configurations other than two equatorial hydroxyls (for example 3- and 4-OH of mannose and glucose, 2- and 3-OH of fucose). For example, a combination of axial and equatorial hydroxyls (3- and 4-OH of fucose) have been found in E- and P-selectin structures; Figure 16.1h) [2].

From a comparative study of different lectin–carbohydrate complexes, Elgavish and Shaanan suggested that additional stereochemical factors need to be considered [2] to understand the determinants of specificity. A detailed electronic description of binding of the sugar hydroxyls, which are also coordinated to $Ca^{2+}$, may be necessary to understand the relative stabilities of possible hydrogen-bonding patterns. Improved understanding is certainly needed as very many sequenced CTLDs – especially the multitude coming from genome sequencing projects – contain atypical or apparently incomplete versions of the carbohydrate-binding motifs we have discussed. Currently, most of these are classified as 'noncarbohydrate binding' but the reliability of these predictions is questionable. A better understanding of specificity would also be useful in assessing whether carbohydrate-binding patterns from known examples, particularly the numerous invertebrate CTLDcps, represent divergent or convergent evolution. The importance of this question will become clearer in the next section where we will see that the repertoire of CTLDcps appears to have been created anew in each major Metazoan branch. This confounds attempts to generate phylogenetic trees, making chemical insights into specificity of great value.

20.4
CTLs in the C

In the pregen
of experiment
diverse specie
sity of CTLDcp
apparent, the fi
Evolution of th

The advent o
bases such as
the way biologi
of the procedur
in Figure 20.4.
HMMER (a pro
genome sequen
CTLD. A simpl
but this is heavi
digmatic mech

The sensitivi
can be improve
as we did in ou
the reliability o
logs and spurio
of invalid find
negatives).

Such factors
analyses report
of a newly sequ
query sequenc
with specific
optimum query
and deduced d
This point is
*elegans* and *Fu*
genome-sequen
was the seventh
specific study a
containing 183
CTLDcps [11].
instances of mis
predicted by th
web site (Ense
structures [8].

## 20.4
## CTLs in the Genome Era

In the pregenome era, knowledge of C-type lectins/CTLDcps reflected the interests of experimental researchers who found, isolated and characterized them from diverse species, but was heavily skewed to mammalian CTLDcps. While the diversity of CTLDcp domain architectures and ligands the CTLD fold could bind was apparent, the full extent of the variety of the CTLDcp superfamily was unknown. Evolution of the superfamily was a puzzle.

The advent of large-scale DNA sequencing, free public access to sequence databases such as GenBank or EMBL and search tools such as BLAST has changed the way biologists work and the way new genes are discovered. A simplified view of the procedures used to search the sequence of a genome for CTLDcps is shown in Figure 20.4. In the first step, a tool such as the profile-based PSI-BLAST or HMMER (a program using hidden Markov models) is used to look for regions of genome sequence which are similar to that of a supplied query sequence for the CTLD. A simple example is the Drickamer sequence signature discussed above, but this is heavily weighted towards CTLDs which bind carbohydrate by the paradigmatic mechanism discussed in the last section.

The sensitivity of the search to detect weak but significant sequence similarity can be improved by incorporating structural information into the query sequence, as we did in our analysis of CTLDcps in the *Fugu* genome [8]. This also improved the reliability of the search by improving the discrimination between true homologs and spurious sequence similarities. This is critical as it minimizes the chances of invalid findings (false positives) and of missing valid occurrences (false negatives).

Such factors highlight a major problem with the results of automated domain analyses reported in the many-authored papers which announce the availability of a newly sequenced model-organism genome. Such audits using nonoptimized query sequences for each domain are less reliable than studies by researchers with specific knowledge of particular domains. These researchers construct optimum query sequences, and manually check ('curate') the gene identifications and deduced domain architecture of the proteins to make sure they are sensible. This point is well illustrated by the CTLDcps audits of the *Caenorhabditis elegans* and *Fugu rubripes* genomes. The 1998 *Science* paper by the *C. elegans* genome-sequencing consortium reported the exciting result that the CTLD was the seventh most common domain with 120 of them. A subsequent CTLD-specific study a little later produced a larger estimate of at least 125 CTLDcps containing 183 CTLDs [5]. In a recent study the number was updated to 278 CTLDcps [11]. In our study of CTLDcps in the *Fugu* genome we found many instances of mis-prediction of genes. Overall, we verified 32 of the gene structures predicted by the *Fugu* genome sequencing team and available on the public web site (Ensembl) but predicted or revised predictions of a further 63 gene structures [8].

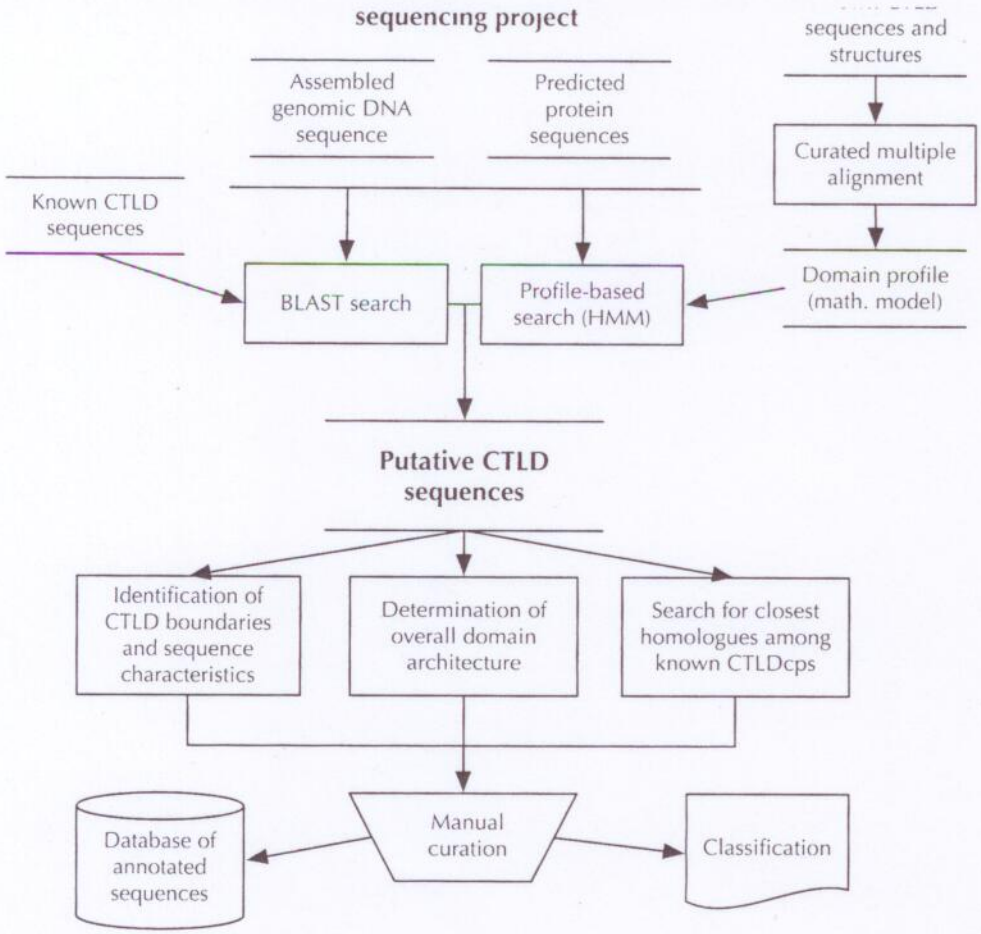



**Figure 20.4** Schematic flowchart illustrating the procedure for auditing CTLDcps of a genome. First, the putative CTLD sequences are found by whole-genome analysis. Next, the genes of CTLDcps containing these CTLDs are analyzed to define their domain architectures and closest homologs, and to classify them into groups, illustrated in summary form in Figure 20.2. Further analysis of the sequences of the CTLDs allows initial assessment of whether they may bind carbohydrate and, if so, the likely specificity (**EPN**/**QPD** and **WND** motifs; see text and Figure 20.3). This information can be included as part of the annotation.

In summary, the biologist needs to exercise caution in assessing published genome statistics. Much careful finding, checking and annotating of genome sequences is necessary before whole-genome CTLDcp statistics and domain architectures can be regarded as sufficiently robust to start drawing evolutionary conclusions. With this caveat that CTLDcp statistics of most genomes should be

superfamily.

## 20.5
## CTL Domain-Conta
## Analysis

In the previous sect
(worm) and *F. rub*
common this abun
whole-genome ana
proteins are 'put to
method is called c
shown in Figure 20


**Figure 20.5** The tree o
rence of CTLDcps. Br
genomes are shown ci
CTLDcps have been i
partially curated analys
'#', respectively. The N

considered as 'work in progress', we will proceed to consider the main findings of the major genome studies and what they can tell us about the evolution of the superfamily.

## 20.5 CTL Domain-Containing Proteins (CTLDcps) in Metazoans from Whole-Genome Analysis

In the previous section we have learnt that CTLDcps are abundant in the *C. elegans* (worm) and *F. rubripes* (fish) genomes. In this section we will examine how common this abundance is in the main Metazoan clades by comparing results of whole-genome analyses of model organisms. We will also examine how these proteins are 'put together' (that is their domain architectures). This investigative method is called comparative genomics. These major Metazoan branches are shown in Figure 20.5.

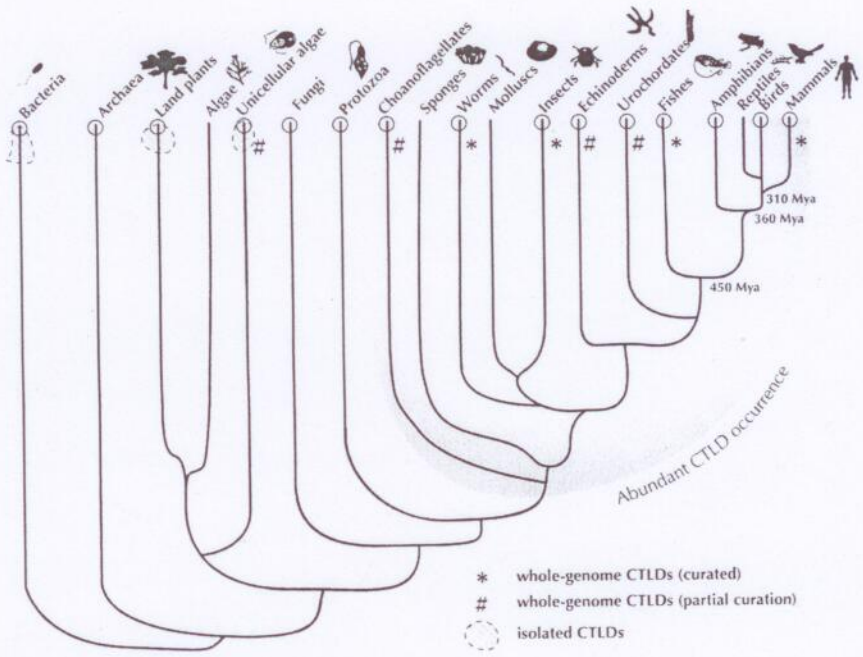



**Figure 20.5** The tree of life showing the occurrence of CTLDcps. Branches with sequenced genomes are shown circled. Branches in which CTLDcps have been identified by curated or partially curated analysis are marked by '*' and '#', respectively. The Metazoan and pre-Metazoan (choanoflagellates) branches are shown shaded in grey. The branching region at the Metazoan 'explosion' is tentative only. Non-Metazoan branches in which CTLDcps have been identified are shown by dotted-hashed circles. Evolution times are not drawn to scale.

Systematic audits of CTLDcps are reported for model organisms representing invertebrates (worm: *Caenorhabditis elegans* [5, 11]; fly/insect: *Drosophila melanogaster* [6]) and vertebrates (human [7]; fish: *Fugu rubripes* [8]). More restricted analyses focused on immune system CTLDcps have been reported for a urochordate (sea squirt: *Ciona intestinalis* [12]) and a protochordate (echinoderm/purple sea urchin: *Strongylocentrotus purpuratus* [13]). Also for sea urchin, a distinctive suite of genes having in common a CTLD evolved to construct the unique biomineral structure of the endoskeletal tissue called the stereom has been analyzed [14]. Description of glycosylation in these model organisms which complements the lectin analysis is given in Chapter 7. Vertebrate genome sequences are also available for a representative bird, amphibian, monotreme and marsupial, and increasingly multiple genome sequences for some branches, especially mammals (mouse, rat, dog, cow) and other animals of commercial importance (for example honey bee).

These studies have confirmed that only the 'tip of the iceberg' of the variety of CTLDcps had been gleaned by the traditional experimental approach. The main conclusions are:

- The CTLD is indeed very common but its relative and absolute abundance varies. For example, it is particularly abundant in *C. elegans* with 125 CTLDcps ([5] and more recently 278 CTLDcps [11]), 52 of them with more than one CTLD, whereas in *Drosophila* only 32 CTLDcps were found, all but one containing only one CTLD [6]. In a typical vertebrate (human), 66 CTLDcps with 96 CTLDs were found [7].
- A high proportion of the CTLDs in the invertebrate CTLDcps lack the sequence signature correlated with carbohydrate-binding capacity discussed above (85% in *C. elegans* and 81% in *Drosophila* [2], whereas about half of the vertebrate CTLDs are classed as CRDs (that is predicted to bind carbohydrate [7]).
- Whereas there is strong conservation of the groups within the vertebrate lineage, there is little or no similarity between vertebrate and invertebrate CTLDcps in their domain organization. This is illustrated in Figure 20.2, where it may also be seen that vertebrate CTLDcps contain a greater variety of other domains than do invertebrate CTLDcps.
- Furthermore, attempts to construct phylogenetic trees from sequence analysis of CTLDs of CTLDcps from evolutionarily distant Metazoan branches (for example human, worm and fly CTLDcps) has been unsuccessful. This has led to the conclusion that the repertoire of CTLDcps has evolved independently in the main Metazoan lineages starting from one or a small number of primordial CTLDs. These have been 'crafted' to create the repertoire of CTLDcps with CTLD specificities for lineage-adapted functions, as observed from whole-genome analysis [2, 3].
- A consequence of this evolutionary complexity is that in cases where a similar carbohydrate-binding function can be attributed to CTLDcps from widely distant Metazoan branches it is difficult to discriminate divergent from convergent

evolution. An exa
the hevein-like d
of the molecular
that similar 'solu

- Evolutionary dive
and regulatory r
There are examp
events leading t
development of r

We will illustrate
first mechanism is
families in groups I
CTLDcps (30 or 45
group II and V gen
of group II and V g
and 27. These clust
gence of the CTLI
example is illustrat
of genes is greatly
structure and funct

An example of g
by CTLDcps with c
proteins and fish a
struction of the ev
which allows us to
of the only carbohy
Reg proteins of gro
lution is given in In

**Info Box 1**

An intriguing exa
the recent reanaly
bered at 278. Mos
gene duplications
CTLDs. It has bee
response towards
sponse). Much of
or 22% of the total
The majority (749
not only in pathog
large set of CTLDc
by evolution in *C.*

evolution. An example of this issue has been presented in Chapters 17 and 19 for the hevein-like domain in plants and animals. As already discussed, case studies of the molecular mechanisms for carbohydrate binding and specificity indicate that similar 'solutions' have likely been found several times by evolution.

- Evolutionary diversification of function of CTLDcps has occurred by both genetic and regulatory mechanisms, consistent with the classic theory of Ohta [15]. There are examples showing gene duplication within groups, recombination events leading to new groups and mutational adaptation of CTLDs, but also development of regulation of expression of CTLDcps genes.

We will illustrate the last point with some examples. In higher vertebrates the first mechanism is particularly common in the adaptive immune system gene families in groups II and V, which constitute a significant proportion of vertebrate CTLDcps (30 or 45%). These proteins bind carbohydrate or protein or both. Most group II and V genes are clustered on the chromosomes, including mixed clusters of group II and V genes. For details of these immune receptors, see Chapters 19 and 27. These clusters clearly result from gene duplication with subsequent divergence of the CTLD sequence to provide varied ligand-binding specificity. An example is illustrated in Figure 20.6 for the SIGNR proteins in mouse; the number of genes is greatly expanded compared with human [16]. For information on the structure and function of DC-SIGN see Figure 16.1i and Chapters 19 and 25.

An example of gene recombination events in vertebrates is well demonstrated by CTLDcps with clade-specific functions, such as snake venoms, bird egg-shell proteins and fish antifreeze proteins. Figure 20.7 presents a phylogenetic reconstruction of the evolutionary history of the CTLDs of selected vertebrate groups which allows us to see the likely origin of the clade-specific CTLDs [2]. The CTLD of the only carbohydrate-binding snake venom class most resembles those of the Reg proteins of group VII. An example involving both genetic and regulatory evolution is given in Info Box 1.

**Info Box 1**

An intriguing example of both genetic and regulatory evolution is provided by the recent reanalysis by Schulenberg *et al.* of *C. elegans* CTLDcps, now numbered at 278. Most of these CLTDcp genes are found in clusters resulting from gene duplications, an interpretation supported by phylogenetic analysis of the CTLDs. It has been found that the nematode is able to mount a distinct defense response towards different pathogens (that is a *specific* innate immune response). Much of this is attributed to pathogen-induced CTLDcps; 61 of them or 22% of the total can be upregulated but very few by more than two pathogens. The majority (74%) are secreted proteins and it is thought that they may act not only in pathogen recognition, but as opsonizing factors. It appears that the large set of CTLDcps are key players in an intricate innate immune system built by evolution in *C. elegans*.

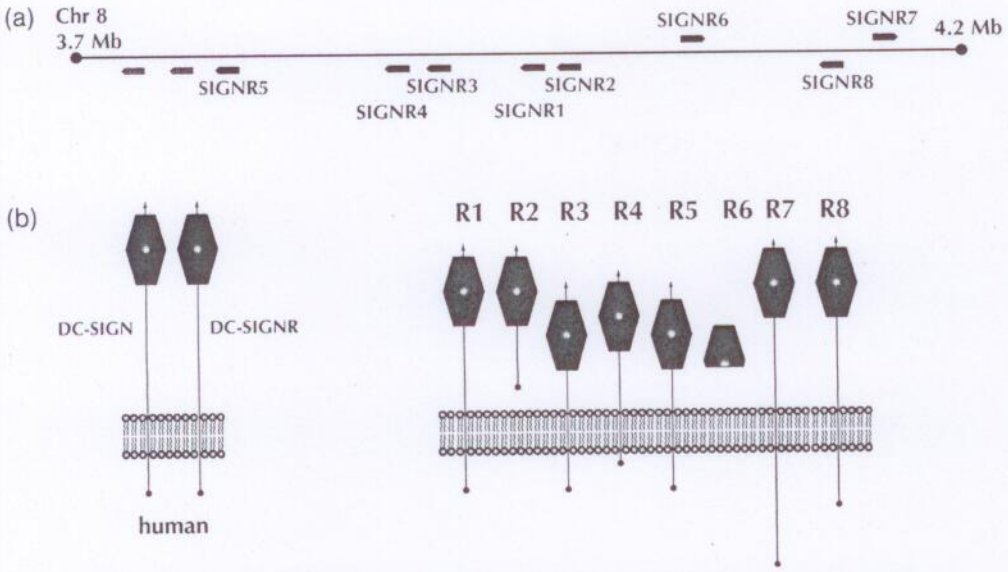



**Figure 20.6** Example of a genomic region showing duplicated CTLDcps genes. The eight SIGNR genes with adjacent CTLDcps genes in mouse are shown in (a) gene order and strand orientation, and (b) domain organization and membrane orientation of the proteins. Adapted from [16].

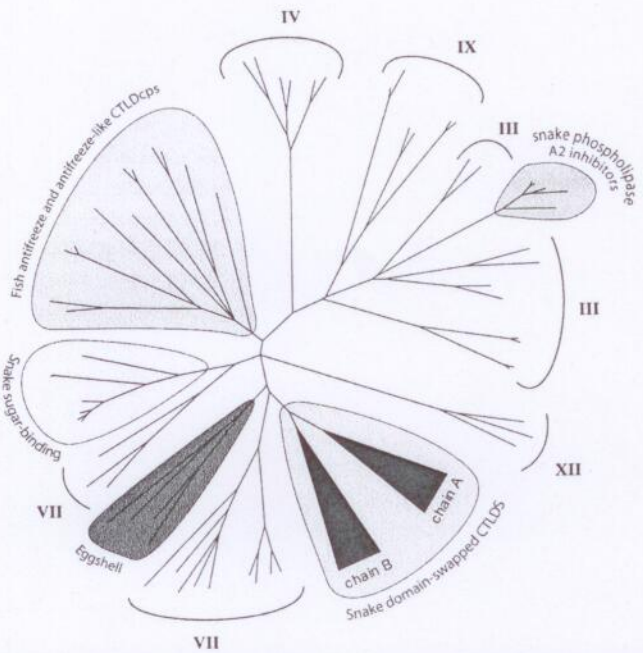



**Figure 20.7** Phylogenetic relationships of clade-specific vertebrate CTLDs to the main vertebrate group CTLDs. The positions of the branches of the CTLDs for the three classes of snake venoms, and bird egg-shell and fish antifreeze proteins show their relative sequence similarity to the CTLDs of mammalian group IV, IX, III, XII and VII CTLDcps [2].

## 20.6 Non-Metazoan CTL

To recapitulate wh
CTLD had originat
binding function.
were consistent wi
but absent in nor
*cerevisiae*). Notwith
non-Metazoan CTI

Many of these
involved in interac
immune system. T
and African swine f
[2]. Well-conserved
protozoan. The be
proaerolysin) and
from *Yersinia pseu*
sented in Chapter
their structures sh
Info Box 2 for exa
contain well-conse
although some lac

How might thes
simonious explana
homologous to th
hijacked host prot
and structural dive
their origin. They

### Info Box 2

Examples of free
*Leeuwenhoekiella*
a marine photosy
proteobacterium;
rium. A particula
genome sequenc
protein in the ge
G and cadherin d
Metazoa (for furth
could be the func
as *Pirellula*?

## 20.6
## Non-Metazoan CTLDs: From Viruses, Bacteria and Protozoa

To recapitulate what we have learnt so far: it had been thought that the primordial CTLD had originated at the beginning of the Metazoan era and had a carbohydrate-binding function. As shown in Figure 20.5, pre-genome and early-genome studies were consistent with this model showing that CTLDcps were abundant in Metazoa but absent in non-Metazoa such as fungi (genome of the yeast *Saccharomyces cerevisiae*). Notwithstanding these general findings, many interesting examples of non-Metazoan CTLDcps have been reported [2].

Many of these CTLDcps are from parasitic viruses and bacteria which are involved in interactions with the animal host, often as mechanisms to defeat its immune system. The CTLDs of the viral proteins (for example fowlpox, vaccinia and African swine fever viruses) contain sequences similar to mammalian CTLDcps [2]. Well-conserved CTLD sequences are also present in *Trypanosoma*, a parasitic protozoan. The best-characterized bacterial group are toxins (pertussis toxin and proaerolysin) and adhesion proteins (intimin from *Escherichia coli* and invasin from *Yersinia pseudotuberculosis*). Details on bacterial toxins and adhesins are presented in Chapter 18. Their CTLDs cannot be identified by sequence analysis and their structures show a more compact fold [2, 4]. However, several CTLDs (see Info Box 2 for examples) found in sequenced genomes from free-living bacteria contain well-conserved **EPN/WND** motifs suggesting they are $Ca^{2+}$/sugar binding, although some lack conserved cysteines.

How might these CTLDcps in bacteria and viruses have arisen? The most parsimonious explanation of the presence of the viral, protozoan and bacterial CTLDs homologous to those in Metazoans is horizontal gene transfer (that is they are hijacked host proteins in viruses or otherwise acquired) [2]. The high sequence and structural divergence of the CTLD of the parasitic bacterial proteins obscure their origin. They may also have been acquired by horizontal gene transfer or,

### Info Box 2

Examples of free-living bacteria with CTLDs found by genome analysis are: *Leeuwenhoekiella blandensis*, a marine flavobacterium; *Synechococcus* sp. RS9917, a marine photosynthetic cyanobacterium; *Marinomonas* sp. MED121, a marine proteobacterium; and *Stigmatella aurantiaca*, a gliding, Gram-negative bacterium. A particularly intriguing example is a putative CTLDcp deduced from the genome sequence of a marine planctomycete *Pirellula* sp. This is the largest protein in the genome (7716 residues) and it contains several CTLD, laminin G and cadherin domains, all of which are domains almost exclusively found in Metazoa (for further information on laminin G domains, see Chapter 16). What could be the function of such a complex protein in a free-living species such as *Pirellula*?

alternatively, may have arisen by convergent evolution, as mimicry of host proteins.

## 20.7
## CTLDcps in Genomes of Pre-Metazoans and Plants

Early findings of CTLD sequences in the genome of the flowering plant *Arabidopsis thaliana*, with a well-conserved ortholog for one of the putative CTLDcps also present in the rice genome, were very perplexing [2]. However, very recent reports of more numerous CTLDcps in genome sequences of a pre-Metazoan, the choanoflagellate *Monosiga brevicollis* (12 CTLDcps) [9], and the unicellular green alga *Chlamydomonas reinhardtii* [10] require a serious reconsideration of the Metazoan hypothesis.

The statistics of the *C. reinhardtii* analysis are compelling, being reminiscent of the general conclusions from the invertebrate and vertebrate genomes we have already outlined. They strongly argue that these CTLDcps have evolved within *Chlamydomonas* for specific roles. Thus, of the 11 CTLDcps containing a total of 67 CTLDs, seven also contain scavenger-receptor cysteine-rich (SRCR) domains, a combination rarely seen in Metazoans and not in CTLDcps for worm, fly and human shown in Figure 20.2. Also, 18 of the CTLDs possess the conserved sequence motifs (EPN/QPD and WND) associated with $Ca^{2+}$-dependent carbohydrate binding.

How can the presence of this novel, and likely functional, collection of CTLDcps in a distant non-Metazoan branch be explained? Although the details of the branching pattern around the time of the Metazoan 'explosion' are still an ongoing debate (please see Figure 20.5), such ambiguity cannot explain the apparent absence of CTLDs in fungi, but variable abundance in plants and algae. Although Wheeler *et al.* [10] interpret their findings in terms of loss of CTLDcp genes or divergence of the CTLD beyond recognition as the eukaryotes evolved, a more plausible explanation may be that the plant/algae branch acquired primordial CTLD genes by horizontal transfer, possibly multiple times. Under this model, the patterns of CTLDcps in subbranches such as *Chlamydomonas* simply represent evolution starting at punctuated times to happily build anew groups of CTLDcps to perform novel needed roles, using one of its best domain building blocks – the structurally flexible and functionally versatile CTLD. This explanation is more consistent with the knowledge we have acquired in this chapter of the CTLD as an evolutionarily very successful domain in Metazoans as it does not require an explanation as to why the CTLD was not equally successful in fungi and plants!

## 20.8
## Conclusions

In this chapter we have taken a historical approach to showing how understanding of the structure, function and evolution of C-type lectins has unfolded. We have

seen how C-type lec
tematic methods of
how such studies h
tance of C-type lectin
tive approaches. We
strong framework fo
have expedited the re
So we have learnt so
beginning of the cha

However, new cha
lessons, particularly
proteins evolutionari
of CTLDcps in wide
subject to different
worms. The plenitud
an improved unders
and specificity in or
sugar code using se
questions for futur
branches, and defin
obvious carbohydrat

**Summary Box**

Our journey track
C-type lectins, ha
protein domains i
usually with othe
essential for mul
communication a
genome sequence
diversity in the m
CTLDcps with pre
and algae. The su
bohydrates and n
success.

### References

1 Drickamer K. Evolu
animal lectins. *Pro
Biol* 1993;45:207–3
2 Zelensky AN, Gr
lectin-like domain
2005;272:6179–217

seen how C-type lectin research has been greatly advanced by application of systematic methods of structural biology and whole-genome sequence analysis, and how such studies have provided a different perspective on the variety and importance of C-type lectins than can be provided by traditional experimental investigative approaches. We have seen how these systematic approaches have provided a strong framework for linking molecular mechanisms to biological functions and have expedited the resolution of many evolutionary questions for the superfamily. So we have learnt some pithy lessons with general relevance, as promised at the beginning of the chapter.

However, new challenges have appeared which follow from some of these lessons, particularly the notion that CTLDcps are an unusually dynamic set of proteins evolutionarily. It appears that we cannot expect to find a paradigmatic set of CTLDcps in widely diverged branches or even for closely related organisms subject to different physiological conditions, such as free-living and parasitic worms. The plenitude of CTLDcps in life highlights the importance of developing an improved understanding of the chemical mechanisms of carbohydrate binding and specificity in order to provide an improved tool for initial deciphering their sugar code using sequence analysis combined with homology modeling. Other questions for future study are whether CTLDcps are present in some fungal branches, and definition of the roles of CTLDcps in plants, especially those with obvious carbohydrate-binding capacity.

**Summary Box**

**Our journey tracking the evolution of the CLTD, the defining protein fold of C-type lectins, has provided many insights. It is one of the most abundant protein domains in Metazoa having been recruited by evolution into proteins, usually with other domains, which carry out the full spectrum of functions essential for multicellular life–cell adhesion, immune defense, intercellular communication and integration, and glycoprotein metabolism. Analyses of genome sequences of model organisms have defined the full extent of this diversity in the major branches of life, including the unexpected presence of CTLDcps with predicted sugar-binding function in non-Metazoa such as plants and algae. The superior stability and versatility of the fold to adapt to bind carbohydrates and many other types of ligands are the keys to its evolutionary success.**

## References


1 Drickamer K. Evolution of $Ca^{2+}$-dependent animal lectins. *Prog Nucleic Acid Res Mol Biol* 1993;45:207–32.

2 Zelensky AN, Gready, JE. The C-type lectin-like domain superfamily. *FEBS J* 2005;272:6179–217.

3 Drickamer K. C-type lectin-like domains. *Curr Opin Struct Biol* 1999;9:585–90.

4 Zelensky AN, Gready JE. Comparative analysis of structural properties of the C-type-lectin-like domain (CTLD). *PROTEINS* 2003;52:466–77.

5 Drickamer K, Dodd RB. C-type lectin-like domains in *Caenorhabditis elegans*: predictions from the complete genome sequence. *Glycobiology* 1999;9:1357–69.
6 Dodd RB, Drickamer K. Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity. *Glycobiology* 2001;11:71R–9R.
7 Drickamer K, Fadden AJ. Genomic analysis of C-type lectins. *Biochem Soc Symp* 2002;69:59–72.
8 Zelensky AN, Gready JE. C-type lectin-like domains in *Fugu rubripes*. *BMC Genomics* 2004;5:51.
9 King N *et al.* The genome of the choanoflagellate *Monosiga brevicollis* and the origin of Metazoans. *Nature* 2008;451:783–8.
10 Wheeler GL *et al.* Genome analysis of the unicellular green alga *Chlamydomonas reinhardtii* indicates an ancient evolutionary origin for key pattern recognition and cell-signaling protein families. *Genetics* 2008;179:193–7.
11 Schulenburg H *et al.* Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode *Caenorhabditis elegans*. *Immunobiology* 2008;213:237–50.
12 Azumi K *et al.* Genomic analysis of immunity in a Urochordate and the emergence of the vertebrate immune system: 'waiting for Godot'. *Immunogenetics* 2003;55:570–81.
13 Hibino T *et al.* The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol* 2006;300:349–65.
14 Bottjer DJ *et al.* Paleogenomics of echinoderms. *Science* 2006;314:956–60.
15 Ohta T. Evolution by gene duplication revisited: differentiation of regulatory elements versus proteins. *Genetica* 2003;118:209–16.
16 Powlesland AS *et al.* Widely divergent biochemical properties of the complete set of mouse DC-SIGN-related proteins. *J Biol Chem* 2006;281:20440–9.

# 21
# Carbohydrate–C

*Iwona Bucior, Max M*

Carbohydrates are
binding proteins, ty
19). However, broa
including direct carl
steps in these mul
carbohydrate intera
stems from the fact
are likely involved i

## 21.1
## Molecular Basis of C

Carbohydrate–carb
satile mechanism f
ticity of glycan cha
sites, and to the cap
ciation forces (for
through surfaces d
covalent bonds (see
Waals contacts, hyd
(Figure 21.1a) [2].
well defined by th
dimensional (3D) s
polar patches of carl
to the initial select
and amino groups
solvating water mo
13.1 for illustration
provide electrostati

*The Sugar Code*. Edited b

ISBN: 978-3-527-32089-9

**The Editor**

***Prof. Dr. Hans-Joachim Gabius***
Ludwig-Maximilians-University Munich
Faculty of Veterinary Medicine
Department of Veterinary Sciences
Chair of Physiological Chemistry
Veterinärstrasse 13
80539 München
Germany

***Cover***
Staining for α2,6-sialylated N-glycans of a bovine blastocyst (together with DNA and F-actin staining; please see Fig. 24.4 for details) and for α2,8-linked polysialic acid of a rat embryo (please see Fig. 6.1 and Chapter 30.7 for details) is exemplarily illustrated to document the importance of glycosylation and protein-carbohydrate recognition, shown in the center (please see Fig. 13.1 for details), from fertilization and different stages of embryogenesis to reach the adult and enter the new cycle for progeny.



**Library of Congress Card No.:** applied for

**British Library Cataloguing-in-Publication Data**
A catalogue record for this book is available from the British Library.

**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.d-nb.de.



**Cover** Adam Design, Weinheim
**Typesetting** SNP Best-set Typesetter Ltd., Hong Kong
**Printing and Binding** Strauss GmbH, Mörlenbach

Printed in the Federal Republic of Germany
Printed on acid-free paper

**ISBN:** 978-3-527-32089-9

**Forew**

The bo
letters
chemis
'*molecu*
etc'. Fo
accept t
in mol
interact
trafficki
as clear
agents
metasta

To en
posal a
sugar-e
teins cal
mentary
differen
pharma
science.

The m
organic
mainly v
last two
sugars a
*Sugar Co*
welcome
around t
mentals
nomena.
treated i
carbohyc
lectinolo


ISBN: 978-3

A reader friendly, strategically structured introduction to glycosciences, guiding from the basics of the sugar alphabet and pertinent aspects of chemical/natural glycosylation to functional carbohydrate recognition (sugar code), with an eye on emerging medical relevance. All chapters, intimately connected by cross-referencing, share the same interdisciplinary level to make the information readily digestible. Story-telling info boxes add an entertaining touch.

Written by a team of renowned glycoscientists from the forefront of research with the clear intention to convey their enthusiasm for teaching key lessons and edited by a leading figure of the field this book is a perfect primer for students in life and medical sciences and for specialists looking for a concise overview. With its remarkable set of summary illustrations and its focus on landmark references it is ideally suited as practical resource to teach classes, run courses and get prepared for contributing to the dynamic development of glycosciences.

**Supplementary Material is available at**
**www.wiley-vch.de/home/thesugarcode**

*Hans-Joachim Gabius studied biochemistry at the University of Hannover and the University of California San Diego. He received his PhD degree in 1982 for investigating proofreading mechanisms by phenylalanyl-tRNA synthetases, then starting his contribution to glycosciences in 1983 at the Max-Planck-Institute for Experimental Medicine in Göttingen with his discovery and characterization of lectins in tumors. Following an appointment as associate professor for pharmaceutical chemistry from 1991–1993 at the Philipps-University Marburg he became chairman of the Institute for Physiological Chemistry in the Faculty of Veterinary Medicine of the Ludwig-Maximilians-University Munich. Hans-Joachim Gabius received numerous research awards including the Otto-Hahn-Medal of the Max-Planck-Society, and awards of the Dr. Carl-Duisberg Foundation and the Paul-Martini Foundation. He has published over 500 research articles in peer-reviewed journals with over 13,000 citations and an h-factor of 59, this resource for teaching and self-study being the fifth book for which he served as editor.*

ISBN 978-3-527-32089-9
9 783527 320899
www.wiley-vch.de