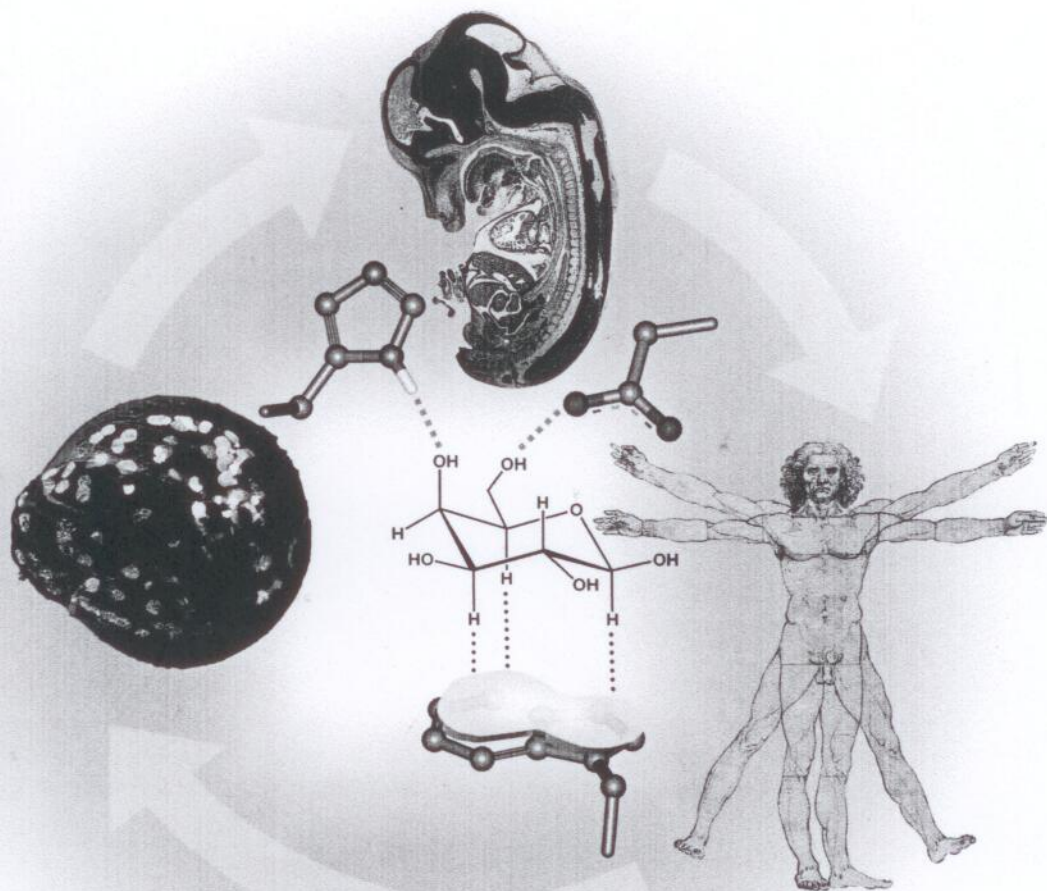


Edited by Hans-Joachim Gabius

 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., Sinaý, 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**  
Fundamentals of  
Editea  
Hans-

  
WILEY  
BLACKWELL  
WILEY

ored post-  
magnify the  
rief discus-  
play pivotal  
o teach this

y Schachter

## Contents

Forewords V

Preface XXI

List of Contributors XXIII

### Part One Chemical Basis

- 1 The Biochemical Basis and Coding Capacity of the Sugar Code 3**  
*Harold Rüdiger and Hans-Joachim Gabius*
- 1.1 Etymological Roots 3  
1.2 What Projection Formulas Tell Us 4  
1.3 The Coding Capacity of the Sugar Code 8  
1.4 Conclusions 12  
References 13
- 2 Three-Dimensional Aspects of the Sugar Code 15**  
*Tibor Kožár, Sabine André, Jozef Uličný, and Hans-Joachim Gabius*
- 2.1 How to Obtain Information about Carbohydrate Conformation 15  
2.2 Complexity of Carbohydrate Flexibility 16  
2.3 How to Describe the Shape of Monosaccharides 17  
2.4 How to Describe the Shape of Di- and Oligosaccharides 18  
2.5 Additional Factors Influencing the Shape of Oligo- and Polysaccharides 22  
2.6 Examples of Di- and Oligosaccharide Conformations 23  
2.7 Carbohydrate-Protein Intermolecular Interactions and Reaction Mechanisms 26  
2.8 How to Perform Molecular Modeling of Large Glycans 28  
2.9 Conclusions 28  
References 29

<b>3</b>	<b>The Chemist's Way to Synthesize Glycosides</b>	<b>31</b>	6.2
	<i>Stefan Oscarson</i>		6.3
3.1	Synthesis of Oligosaccharides: Strategies	32	
3.2	Glycosidic Bond Formation	33	6.4
3.3	Fischer Glycosylations	35	
3.4	Glycosyl Donors	36	6.5
3.5	Anomeric Configuration: Stereoselectivity	38	6.6
3.5.1	Formation of 1,2- <i>Trans</i> -Linkages	39	6.6.1
3.5.2	Formation of 1,2- <i>Cis</i> -Linkages	40	6.6.2
3.6	Neuraminic Acid and Kdo-Glycoside Synthesis	42	
3.7	Formation of Building Blocks: Orthogonal Glycosylations	44	6.6.3
3.8	Protecting Group Manipulations	46	6.7
3.9	An Example	46	
3.10	Conclusions	49	6.8
	References	51	6.8.1
			6.8.2
			6.8.3
<b>4</b>	<b>The Chemist's Way to Prepare Multivalency</b>	<b>53</b>	6.8.4
	<i>Yoann M. Chabre and René Roy</i>		6.8.5
4.1	Blocking Viral/Bacterial Adhesion	54	6.8.6
4.2	How to Prepare Multivalent Carbohydrates?	56	6.8.7
4.3	Neoglycoproteins	57	6.9
4.4	Neoglycolipids and Liposomes	58	6.10
4.5	Glycopolymers	60	
4.6	Glycodendrimers	62	
4.7	Glycodendrimer Syntheses	64	
4.8	Conclusions	66	<b>7</b>
	References	69	
			7.1
			7.2
<b>5</b>	<b>Analytical Aspects: Analysis of Protein-Bound Glycans</b>	<b>71</b>	7.3
	<i>Hiroaki Nakagawa</i>		7.3.1
5.1	Detection of Glycans on Glycoproteins	71	7.3.2
5.2	Release of Glycans from Glycoproteins	73	7.3.3
5.3	Glycan Purification	74	7.3.4
5.4	Detailed Structural Analysis Using HPLC	76	7.4
5.5	Detailed Structural Analysis Using MS	78	7.4.1
5.6	Glycomic Analysis Using MS	78	7.4.1.1
5.7	Other Methods of Analysis	80	7.4.1.2
5.8	Glycopeptide Analysis Using MS	81	7.4.1.3
5.9	Conclusions	82	7.4.1.4
	References	82	7.4.1.5
			7.4.2
	<b>Part Two Natural Glycosylation—Glycoproteins</b>		7.4.2.1
			7.4.2.2
<b>6</b>	<b>N-Glycosylation</b>	<b>87</b>	7.4.3
	<i>Christian Zuber and Jürgen Roth</i>		7.4.3.1
6.1	NCAM1	87	

6.2	Initial Steps in Asparagine-Linked Glycosylation	89
6.3	Trimming Reactions by $\alpha$ -Glucosidases and Interactions with ER Lectins	92
6.4	Quality Control of Protein Folding and Assembly: Machinery and Principal Mechanism	95
6.5	ER Exit—Facing a Crucial Decision and What Mannose Has to Do	96
6.6	How to Become a Mature <i>N</i> -Glycan?	99
6.6.1	Golgi Mannose Trimming as the Start for <i>N</i> -Glycan Elongation	100
6.6.2	Nucleotide Sugar Transporters Import the Fuel for Oligosaccharide Elongation	100
6.6.3	Glycosyltransferases: The Orderly Maturation Reactions	102
6.7	Structure Building by <i>N</i> -Acetylglucosaminyltransferase-I and Fucosyltransferase-VIII	103
6.8	Branching and Elongation Reactions	103
6.8.1	Mannosyl $\beta$ - <i>N</i> -Acetylglucosaminyltransferases	103
6.8.2	<i>N</i> -Acetylglucosaminyl- $\beta$ -Galactosyltransferases	104
6.8.3	Capping Sugars Provide Functions	104
6.8.4	Sialyltransferases	105
6.8.5	Fucosyltransferases	106
6.8.6	Glucuronyltransferases	106
6.8.7	Sulfotransferases	107
6.9	Diversity of <i>N</i> -Glycans: Structural and Functional Implications	107
6.10	Conclusions	109
	References	109
<b>7</b>	<b>O-Glycosylation: Structural Diversity and Functions</b>	<b>111</b>
	<i>Georgios Patsos and Anthony Corfield</i>	
7.1	Structure of <i>O</i> -Linked Glycans	111
7.2	Biosynthetic Routes for <i>O</i> -Glycans	121
7.3	Regulation of <i>O</i> -Glycosylation and Glycan Processing	125
7.3.1	<i>O</i> -GalNAc, Mucin Type	125
7.3.2	$\beta$ - <i>O</i> -GlcNAc	127
7.3.3	<i>O</i> -Man	128
7.3.4	<i>O</i> -Fuc and <i>O</i> -Glc	128
7.4	Functions of <i>O</i> -Linked Glycosylation	129
7.4.1	<i>O</i> -GalNAc, Mucin Type	129
7.4.1.1	Protein Structure and Stability	129
7.4.1.2	Protein Conformation and Tertiary Structure	129
7.4.1.3	Protein Quaternary Structure and Molecular Association	130
7.4.1.4	Protein Stability: Protease and Heat Resistance	130
7.4.1.5	Recognition Phenomena	130
7.4.2	$\beta$ - <i>O</i> -GlcNAc	130
7.4.2.1	Protein Structure and Stability	132
7.4.2.2	Recognition Phenomena and Disease	132
7.4.3	<i>O</i> -Man	132
7.4.3.1	Protein Structure and Stability	132

7.4.3.2	Recognition Phenomena	132	10.6
7.4.4	O-Fuc and O-Glc	132	10.7
7.4.4.1	Protein Structure and Stability	134	10.8
7.4.4.2	Recognition Phenomena	134	10.9
7.5	Mucins: A Major Group of O-Glycosylated Proteins	134	10.10
7.6	Conclusions	136	10.11
	References	137	10.12
<b>8</b>	<b>Glycosylation of Model and 'Lower' Organisms</b>	<b>139</b>	<b>10.13</b>
	<i>Iain B. H. Wilson, Katharina Paschinger, and Dubravko Rendić</i>		10.14
8.1	Bacterial Glycosylation	139	10.15
8.2	Yeast Glycosylation	141	
8.3	Plant Glycosylation	143	<b>11</b>
8.4	Insect Glycosylation	146	
8.5	Worm Glycosylation	148	11.1
8.6	Protozoan Glycosylation	150	11.1.1
8.7	Fish Glycosylation	151	11.1.2
8.8	Conclusions	152	11.1.3
	References	153	11.2
			11.3
			11.3.1
<b>9</b>	<b>Glycosylphosphatidylinositol Anchors: Structure, Biosynthesis and Functions</b>	<b>155</b>	11.3.2
	<i>Hosam Shams-Eldin, Françoise Debierre-Grockiego, and Ralph T. Schwarz</i>		11.3.3
9.1	Structure of GPI Anchors	155	11.4
9.1.1	Detection and Isolation of GPI-Anchored Proteins	158	11.5
9.1.2	Biosynthesis of GPI Anchors	160	11.6
9.2	Remodeling of Lipid Moieties of GPI Proteins	166	11.6.1
9.3	Chemical Synthesis of GPIs	167	11.6.1.1
9.3.1	Mutant Cells Lead the Way to Identification of Complementation Classes Involved in GPI Biosynthesis	167	11.6.1.2
9.3.2	Defects in GPI Anchor Biosynthesis	168	11.6.2
9.3.3	Function	169	11.7
9.4	Conclusions	171	
	References	173	<b>12</b>
	<b>Part Three Natural Glycosylation—Glycolipids, Proteoglycans and Chitin</b>		12.1
			12.2
			12.3
			12.3.1
<b>10</b>	<b>Glycolipids</b>	<b>177</b>	12.3.2
	<i>Jürgen Kopitz</i>		12.3.3
10.1	Classification and General Structures of Glycolipids	177	12.3.4
10.2	Glycoglycerolipids in Thylakoid Membranes	180	12.4
10.3	Glycolipids in Non-photosynthetic Bacteria	180	12.5
10.4	Bacterial Glycolipids in T-Cell Activation	182	
10.5	Glycosphingolipids (GSLs)	183	

10.6	Complex Neutral GSLs	183
10.7	Complex Acidic (Anionic) GSLs	185
10.8	Survey of GSL Functions	187
10.9	GSL Microdomains	188
10.10	GSLs as Attachment Sites for Viruses, Bacteria and Toxins	192
10.11	GSLs as Developmental or Differentiation Markers	193
10.12	Tumor-Associated GSL Antigens	193
10.13	Gangliosides in Neural Tissue	195
10.14	GSL Degradation and GSL Storage Disorders	195
10.15	Conclusions	196
	References	197
<b>11</b>	<b>Proteoglycans</b>	<b>199</b>
	<i>Eckhart Buddecke</i>	
11.1	Glycosaminoglycans: Components of Proteoglycans (PGs)	199
11.1.1	Structure	199
11.1.2	Biosynthesis	201
11.1.3	Catabolism	202
11.2	PGs	204
11.3	Large Aggregating (Hyaluronan-Binding) PGs	204
11.3.1	Aggrecan	205
11.3.2	Versican	206
11.3.3	Neurocan, Brevican	207
11.4	Small Leucine-Rich PGs	207
11.5	Basement Membrane PGs	209
11.6	Cell-Surface (Transmembrane) PGs	211
11.6.1	Syndecans	211
11.6.1.1	Structure	211
11.6.1.2	Functions	212
11.6.2	Glypicans	213
11.7	Conclusions	214
	References	215
<b>12</b>	<b>Chitin</b>	<b>217</b>
	<i>Hans Merzendorfer</i>	
12.1	Occurrence	218
12.2	Structure	219
12.3	Function	221
12.3.1	Fungal Cell Walls	222
12.3.2	Arthropod Cuticles and Shells	222
12.3.3	Peritrophic Matrices and Cocoons	223
12.3.4	Other Functions	224
12.4	Metabolism	224
12.5	Conclusions	228
	References	228



	<b>Part Four Protein–Carbohydrate Interactions</b>	<b>17</b>
<b>13</b>	<b>Protein–Carbohydrate Interactions: Basic Concepts and Methods for Analysis</b> 233 <i>Dolores Solís, Antonio Romero, Margarita Menéndez, and Jesús Jiménez-Barbero</i>	17.1 17.1.1 17.1.1.1 17.1.1.2 17.1.1.3
13.1	Atomic Features of Protein–Sugar Interactions 233	17.1.1.4
13.2	Role of Water in Protein–Sugar Interactions 237	17.1.2
13.3	Selection of Carbohydrate Conformers by Proteins 238	17.1.2.1
13.4	Thermodynamics of Protein–Carbohydrate Interactions 241	17.1.2.2
13.5	Conclusions 244	17.1.3
	References 245	17.1.3.1 17.1.3.2 17.1.3.3 17.1.3.4
<b>14</b>	<b>How to Determine Specificity: From Lectin Profiling to Glycan Mapping and Arrays</b> 247 <i>Hiroaki Tateno, Atsushi Kuno, and Jun Hirabayashi</i>	17.2 17.2.1 17.2.1.1 17.2.1.2 17.2.1.3
14.1	Quantitative Aspects of Lectin Affinity 248	17.2.2
14.2	Frontal Affinity Chromatography (FAC) for Sugar–Protein Interactions 250	17.2.3
14.3	Automated FAC-FD System 252	17.2.4
14.4	From ‘Lectin Profiling’ to ‘Glycan Mapping’ 254	17.2.5
14.5	Lectin Microarray Enables Multiplexed Lectin–Glycan Interaction Analysis 254	17.2.6
14.6	Practice in Differential Glycan Profiling: Approaches and Applications 257	17.2.7
14.7	Conclusions 258	17.3 17.3.1 17.3.2 17.3.3
	References 259	17.4
<b>15</b>	<b>The History of Lectinology</b> 261 <i>Harold Rüdiger and Hans-Joachim Gabius</i>	
15.1	How Lectinology Started 261	
15.2	Early Definitions 264	
15.3	The Current Definition of the Term ‘Lectin’ 265	
15.4	Recent Developments 266	
15.5	Conclusions 267	
	References 268	<b>18</b>
<b>16</b>	<b>Ca<sup>2+</sup>: Mastermind and Active Player for Lectin Activity (Including a Gallery of Lectin Folds)</b> 269 <i>Hans-Joachim Gabius</i>	18.1 18.2 18.3 18.4 18.5 18.6
16.1	Ca <sup>2+</sup> : Organizing the Active Site 269	
16.2	Ca <sup>2+</sup> : Contacting Charged Ligands 272	
16.3	Ca <sup>2+</sup> : Neutralizing Negative Charges and Contacting Neutral Ligands 275	
16.4	Conclusions 277	
	References 278	

**17 Bacterial and Viral Lectins 279***Jan Holgersson, Anki Gustafsson, and Stefan Gaunitz*

- 17.1 Bacterial Lectins 279
  - 17.1.1 Fimbriae/Pili 282
    - 17.1.1.1 Type 1 Fimbriae 283
    - 17.1.1.2 Type P Fimbriae 283
    - 17.1.1.3 Type S Fimbriae 284
    - 17.1.1.4 Type IV Pili 284
  - 17.1.2 Bacterial Surface Lectins 284
    - 17.1.2.1 BabA and SabA 285
    - 17.1.2.2 LecA and LecB 285
  - 17.1.3 Toxins 285
    - 17.1.3.1 Toxin A of *Clostridium difficile* 285
    - 17.1.3.2 Cholera Toxin 286
    - 17.1.3.3 Heat-Labile and Heat-Stable Toxins 286
    - 17.1.3.4 Shiga and Shiga-Like Toxins 286
- 17.2 Virus Binding 287
  - 17.2.1 Influenza Virus 290
    - 17.2.1.1 Influenza Virus Surface Proteins 290
    - 17.2.1.2 Epidemiology 290
    - 17.2.1.3 Influenza Virus Species and Tissue Tropism 291
  - 17.2.2 Rotavirus 291
  - 17.2.3 Human Immunodeficiency Virus 1 292
  - 17.2.4 Norovirus 292
  - 17.2.5 Herpes viruses 293
  - 17.2.6 Hepatitis C Virus 293
  - 17.2.7 Paramyxoviridae 294
- 17.3 Carbohydrate-Based Antiinfectives 295
  - 17.3.1 Neuraminidase Inhibitors as Drugs Against Influenza 295
  - 17.3.2 Oligosaccharides as Inhibitors of Microbial Adhesion 295
  - 17.3.3 A New Generation of Multivalent, Carbohydrate-Based Inhibitors of Microbial Adhesion 297
- 17.4 Conclusions 298
- References 299

**18 Plant Lectins 301***Harold Rüdiger and Hans-Joachim Gabius*

- 18.1 Nomenclature 301
- 18.2 Folding Patterns and Occurrence 305
- 18.3 Purification 309
- 18.4 Applications 311
- 18.5 Biological Functions 311
- 18.6 Conclusions 314
- References 315

**19 Animal and Human Lectins 317**  
*Hans-Joachim Gabius*

19.1 Protein Folds with Lectin Activity 318

19.2 Functions of Animal and Human Lectins 320

19.3 Lectin Ligands and Affinity Regulation 326

19.4 Conclusions 327

References 328

**20 Routes in Lectin Evolution: Case Study on the C-Type Lectin-Like Domains 329**  
*Jill E. Gready and Alex N. Zelensky*

20.1 C-Type Lectin (CTL) Evolution as a Case Study 329

20.2 CTL Superfamily: Structures and Groups 330

20.3 Mechanism of Carbohydrate Binding 335

20.4 CTLs in the Genome Era 337

20.5 CTL Domain-Containing Proteins (CTLDcps) in Metazoans from Whole-Genome Analysis 339

20.6 Non-Metazoan CTLDs: From Viruses, Bacteria and Protozoa 343

20.7 CTLDcps in Genomes of Pre-Metazoans and Plants 344

20.8 Conclusions 344

References 345

**21 Carbohydrate–Carbohydrate Interactions 347**  
*Iwona Bucior, Max M. Burger, and Xavier Fernández-Busquets*

21.1 Molecular Basis of Carbohydrate–Carbohydrate Interactions 347

21.2 Carbohydrate–Carbohydrate Interactions in Cell Recognition 349

21.2.1 Proteoglycans 349

21.2.1.1 Carbohydrate Self-Interactions in Sponge Proteoglycans 349

21.2.1.2 Glycosaminoglycan Self-Interactions 351

21.2.2 Glycolipids 352

21.2.2.1 Le<sup>x</sup>–Le<sup>x</sup> Interactions 352

21.2.2.2 Gb4-Dependent Adhesion 352

21.2.2.3 GM3-Dependent Adhesion 353

21.3 Carbohydrates as DNA-Binding Motifs 354

21.4 New Strategies to Study Multivalent Carbohydrate–Carbohydrate Interactions 355

21.4.1 Analytical Ultracentrifugation 355

21.4.2 2D/3D Polyvalent Model Systems 356

21.4.2.1 Surface Plasmon Resonance (SPR) 356

21.4.2.2 Quartz Crystal Microbalance (QCM) 357

21.4.3 Single-Molecule Detection and Manipulation 358

21.4.3.1 Single-Molecule Force Spectroscopy (SMFS) 358

21.5 Conclusions 359

References 361

**Part F**

**22 Disease-Associated Glycans 365**  
*Thierry Thoden*

22.1 N-Glycans 365

22.2 O-Glycans 365

22.2.1 O-Glycans 365

22.2.2 O-Mannosylated Glycans 365

22.2.3 O-Fucose 365

22.3 Glycans in Disease 365

22.4 Glycans in Disease 365

22.5 Glycans in Disease 365

22.6 Disease-Associated Glycans 365

22.7 Transferrin 365

22.8 Conclusions 365

References 365

**23 Antigen Presentation by Glycans 365**  
*Kenneth Wilson*

23.1 Introduction 365

23.2 Specificity 365

23.2.1 Bacteria 365

23.2.2 Fungi 365

23.2.3 Viruses 365

23.2.4 Eukaryotes 365

23.2.5 Unconventional Antigen Presentation 365

23.2.6 Immunomodulation 365

23.3 Outlook 365

23.4 Conclusions 365

References 365

**24 Glycans in Cell Signaling 365**  
*Felix Beutelmann*

24.1 Primary Signaling 365

24.2 The Cell Surface 365

24.3 The Cell Surface 365

24.4 Glycans in Cell Signaling 365

24.5 The Cell Surface 365

24.6 Ligand Binding 365

24.7 The Cell Surface 365

24.8 Surface Glycans 365

24.9 Conclusions 365

References 365

## Part Five Biomedical Aspects and Case Studies

- 22 Diseases of Glycosylation 365**  
*Thierry Hennet*
- 22.1 N-Glycosylation 366
- 22.2 O-Glycosylation 372
- 22.2.1 O-GalNAc Glycosylation 372
- 22.2.2 O-Man Glycosylation (O-Mannosylation) 374
- 22.2.3 O-Fuc Glycosylation (O-Fucosylation) 376
- 22.3 Glycosaminoglycans 377
- 22.4 Glycosphingolipids 379
- 22.5 Glycosylphosphatidylinositol Anchor 379
- 22.6 Defects Affecting Multiple Classes of Glycosylation 380
- 22.7 Trafficking Disorders 381
- 22.8 Conclusions 382
- References 383
- 23 Animal Models to Delineate Glycan Functionality 385**  
*Koichi Honke and Naoyuki Taniguchi*
- 23.1 Knockout Mouse 385
- 23.2 Specific Features of Glycogene KO 387
- 23.2.1 Relationship between Glycogenes and Related Glycans 387
- 23.2.2 Functional Association of Glycogenes 390
- 23.2.3 Which Are Essential—Glycans or Their Carriers? 392
- 23.2.4 Effects of Elimination of the Core and Terminal Structures of Glycans 393
- 23.2.5 Unexpected Findings Provide New Insights into Glycan Functions 398
- 23.2.6 Insights into Human Diseases 399
- 23.3 Other Gene Manipulation Techniques 400
- 23.4 Conclusions 401
- References 401
- 24 Glycobiology of Fertilization and Early Embryonic Development 403**  
*Felix A. Habermann and Fred Sinowatz*
- 24.1 Primer to Mammalian Fertilization 403
- 24.2 The Functional Morphology of the Zona Pellucida (ZP) 405
- 24.3 The Glycoproteins of the ZP and Their Encoding Genes 406
- 24.4 Glycan Structures of ZP Glycoproteins 407
- 24.5 The Synthesis of ZP Glycoproteins 409
- 24.6 Ligand Properties of ZP Glycans 410
- 24.7 The Glycoprotein Shell of Mammalian Embryos 413
- 24.8 Surface Glycans of Stem Cells 414
- 24.9 Conclusions 416
- References 416

<b>25</b>	<b>Glycans as Functional Markers in Malignancy?</b> 419	28.3	Dis
	<i>Sabine André, Jürgen Kopitz, Herbert Kaltner, Antonio Villalobo, and Hans-Joachim Gabius</i>	28.4	Car
25.1	The Past 420	28.5	Car
25.2	The Present 421	28.6	Car
25.3	The Future 430		Re
25.4	Conclusions 430		
	References 431	<b>29</b>	F
<b>26</b>	<b>Small Is Beautiful: Mini-Lectins in Host Defense</b> 433		K
	<i>Robert I. Lehrer</i>	29.1	P
26.1	Meet the Families 433	29.2	C
26.2	Where Do $\alpha$ - and $\beta$ -Defensins Reside? 435	29.3	C
26.3	Introducing $\theta$ -Defensins 436	29.4	L
26.4	Introducing Retrocyclins 437		C
26.5	Hemagglutination 438	29.5	P
26.6	How HIV-1 Enters Target Cells 440	29.6	C
26.7	Studies with Influenza A Virus 441		R
26.8	A Toxic Side-Trip 442	<b>30</b>	F
26.9	And Now, the Surprise 442		P
26.10	It Takes Two to Tangle 443	30.1	C
26.11	Which Human $\alpha$ - and $\theta$ -Defensins Are Lectins? 443	30.2	C
26.12	Conclusions 445	30.3	C
	References 445	30.4	C
		30.5	C
		30.6	C
		30.7	C
		30.8	C
		30.9	C
<b>27</b>	<b>Inflammation and Glycosciences</b> 447	30.10	C
	<i>Reinhard Schwartz-Albiez</i>	30.11	C
27.1	Sequence of Events 448		
27.2	Where Do Carbohydrate-Lectin Interactions Play a Role During Acute Inflammation? 449		
27.3	Selectins 451		
27.4	Selectin Carbohydrate Ligands and Their Carrier Glycoproteins 451		
27.5	Galectins 456		
27.6	Siglecs 459		
27.7	Other Lectins Involved in Antigen Recognition and Inflammatory Processes 462		
27.8	Glycans Involved in Bacteria-Host Interactions 463		
27.9	Glycosylation in Inflammatory Bowel Diseases 464		
27.10	Conclusions 466		
	References 466		
<b>28</b>	<b>Sugars as Pharmaceuticals</b> 469		
	<i>Helen M. I. Osborn and Andrea Turkson</i>		
28.1	Cancer Therapeutics 469		
28.2	Viral Infections: HIV-1 and Influenza 473		

28.3	Diabetes	476
28.4	Carbohydrate-Based Antibacterial Agents	477
28.5	Carbohydrate-Based Antithrombotic Agents	480
28.6	Conclusions	482
	References	482
<b>29</b>	<b>Platelet Glycoproteins as Lectins in Hematology</b>	<b>485</b>
	<i>Karin Hoffmeister and Hervé Falet</i>	
29.1	Platelet Physiology	485
29.2	GPIb-IX-V Complex	487
29.3	Cold-Induced Platelet Clearance	488
29.4	Long-Term Platelet Refrigeration May Reveal New Insights into Platelet Clearance	490
29.5	P-Selectin and PSGL-1	491
29.6	Conclusions	492
	References	493
<b>30</b>	<b>Neurobiology Meets Glycosciences</b>	<b>495</b>
	<i>Robert W. Ledeen and Gusheng Wu</i>	
30.1	Glucose and Glycogen as Energy Sources	497
30.2	Gangliosides as Primary Glycans of the Nervous System	497
30.3	Ganglioside Metabolism	499
30.4	Gangliosides of the Peripheral Nervous System	502
30.5	Ganglioside Functional Activities	503
30.6	Neural Glycoproteins: Overview	506
30.7	Neural Recognition Glycoproteins	507
30.8	Glycoproteins of the Synapse	510
30.9	Glycoproteins of Myelin	511
30.10	Proteoglycans and Extracellular Matrix of the Nervous System	512
30.11	Conclusions	514
	References	515
	<b>Glossary</b>	<b>517</b>
	<b>Index</b>	<b>549</b>

## 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  $\text{Ca}^{2+}$  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 (CTLDD). As the proteins containing the domain (CTLDD-containing proteins (CTLDDcps)) 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 glycoenzyme 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 CTLDDcps 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 CTLDD fold to evolve new specificities and functions. Definition of the full repertoire of CTLDDcps 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

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

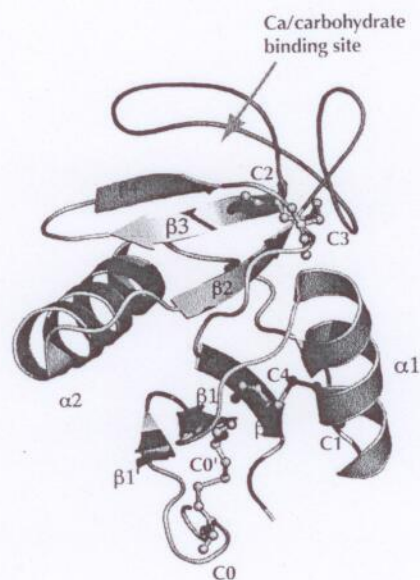
**Figure 20.1** A typical CTL cartoon representation, which harbors the primary binding site (position arrow). Disulfide bridges are shown.

defense, intercellular communication (see Chapters 19 and 20). It may be the ancestral function of defense CTLDcps characterized. CTLDs have evolved into various proteins, but also lipases, proteases and calcium-binding proteins.

As summarized in Figure 20.1, C-type referring to the carbohydrate binding site and distinguished from other dependent lectins (for example, the ‘lectin-like domain’—referring to it became clear that the greater functionality of C-type lectin domains (CRDs) also bind proteins (see Figure 19.3).

In Chapter 19 we discussed the design (see Figure 15.1)





**Figure 20.1** A typical CTLD structure shown in cartoon representation. The long-loop region which harbors the primary  $\text{Ca}^{2+}$ /carbohydrate-binding site (position arrowed) is shown in red. Disulfide bridges are shown as yellow sticks. The archetypal 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  $\text{Ca}^{2+}$ -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

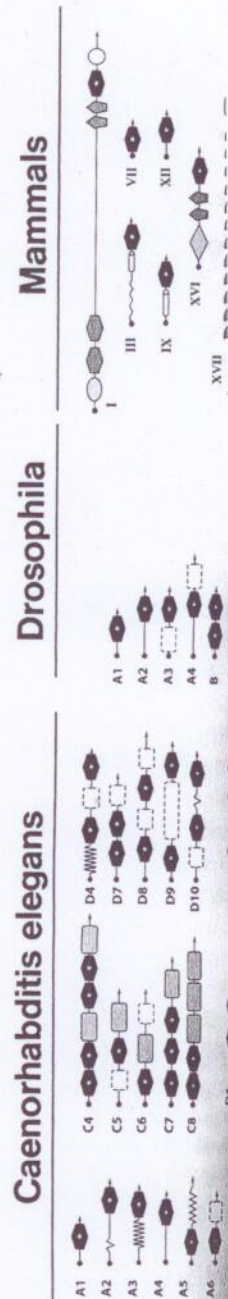
principle, a domain has a compact three-dimensional structure that may fold independently of the rest of the protein, and may function and evolve semi-independently. The sequence of a protein domain is typically 100–200 amino acids long. A given CTLDcp by definition contains an operative lectin domain, 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  $\text{Ca}^{2+}$ -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.

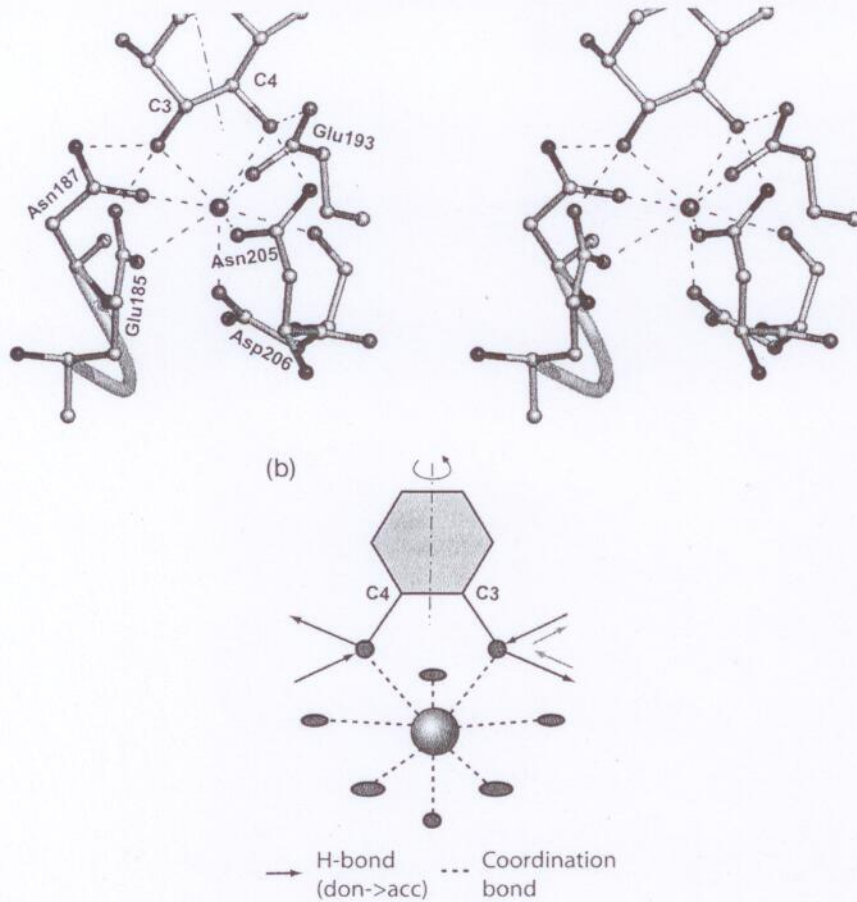
Understanding of the structural basis of the sequence signature came with solution of the first X-ray structure of a CTLD—of rat mannan-binding protein (MBP)-A—in 1991 and a little later by a structure of this CTLD with a bound mannose molecule. Comparison of the residues then identified as the sequence signature—12 totally conserved and 18 conservatively conserved—against the three-dimensional structure allowed the roles for most of them in  $\text{Ca}^{2+}$ - 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  $\beta$ -sheets and two  $\alpha$ -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 20.3**  $\text{Ca}^{2+}$ -dependent monosaccharide binding by CTLDs. (a) Stereo depiction of the structure of rat MBL (formerly called mannose-binding protein, MBP-A) complexed with  $\text{Ca}^{2+}$  (blue sphere) and mannose. The coordination bonds to  $\text{Ca}^{2+}$  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  $\text{Ca}^{2+}$ -pyranose-CTLD complex. Two hydroxyl oxygens and the pyranose ring are shown with the  $\text{Ca}^{2+}$  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

CTLD and whether the

### 20.3 Mechanism of

In this section ligands have re the main binding site that interacts with mannose-type and this main site but selectins (Chapter 27.2 for structure note that it is this CTLDs to select on developing view the

The architecture of ligands is illustrated. CTLD of rat MBP-A (Figures 6 and 8 for N-glycosidic). The complex is formed by oxygen atoms from the sugar forming coordination bonds with the  $\text{Ca}^{2+}$ . Asn187 (EPN), Asn205, and Asp206 form the  $\text{Ca}^{2+}$ -binding site. WND motifs are in the binding site respectively (Figure 16.1h,i).

The arrangement of bonds in the binding site features. First, it depends on the symmetry axis relative to the binding site, rotated by  $180^\circ$  with respect to the examples are now known. It is shown that a galactose-binding site in echinoderm sea cu



the pyranose ring are large blue sphere, and small grey ovals. Arrows show hydrogen bonds in mannose-type sugars, and grey arrows indicate hydrogen bonds in galactose-specific sugars. The orientation of the sugar is shown in the inset. See also Figures 19.1 and 27.1.

and experimentally in the rat MBP-A (1) have been added. The identification has facilitated the discovery of new properties for newly identified lectin 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  $\text{Ca}^{2+}$  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<sup>x</sup> 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  $\text{Man}_6\text{-GlcNAc}_2\text{-Asn}$ ; 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  $\text{Ca}^{2+}$  ion and four hydrogen bonds with residues—Glu185 and Asn187 (EPN), Asn205 (WND) and Glu193—whose carbonyl side-chains coordinate the  $\text{Ca}^{2+}$ -binding site. This bonding pattern is fundamental for CTLD/ $\text{Ca}^{2+}$ /monosaccharide complexes, and is observed in all known structures. The EPN and WND motifs are in the long-loop region and  $\beta$ 4 strand of the CTLD structure, respectively (Figure 20.1). Asp206 of the WND motif contributes another  $\text{Ca}^{2+}$  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  $180^\circ$  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  $\text{Ca}^{2+}$ -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  $\text{Ca}^{2+}$ -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  $\text{Ca}^{2+}$ , 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.

In the presence of experimental data from diverse species, the specificity of CTLDcps is apparent, the fact that evolution of the

The advent of bases such as C the way biological of the procedure in Figure 20.4. HMMER (a protein genome sequence CTLD. A simple but this is heavy digmatic mechanism

The sensitivity can be improved as we did in our the reliability of logs and spurious of invalid findings (negatives).

Such factors analyses report of a newly sequenced query sequence with specific optimum query and deduced data. This point is *C. elegans* and *F. coli* genome-sequenced was the seventh specific study a containing 183 CTLDcps [11]. Instances of mispredicted by the web site (Ensembl structures [8].

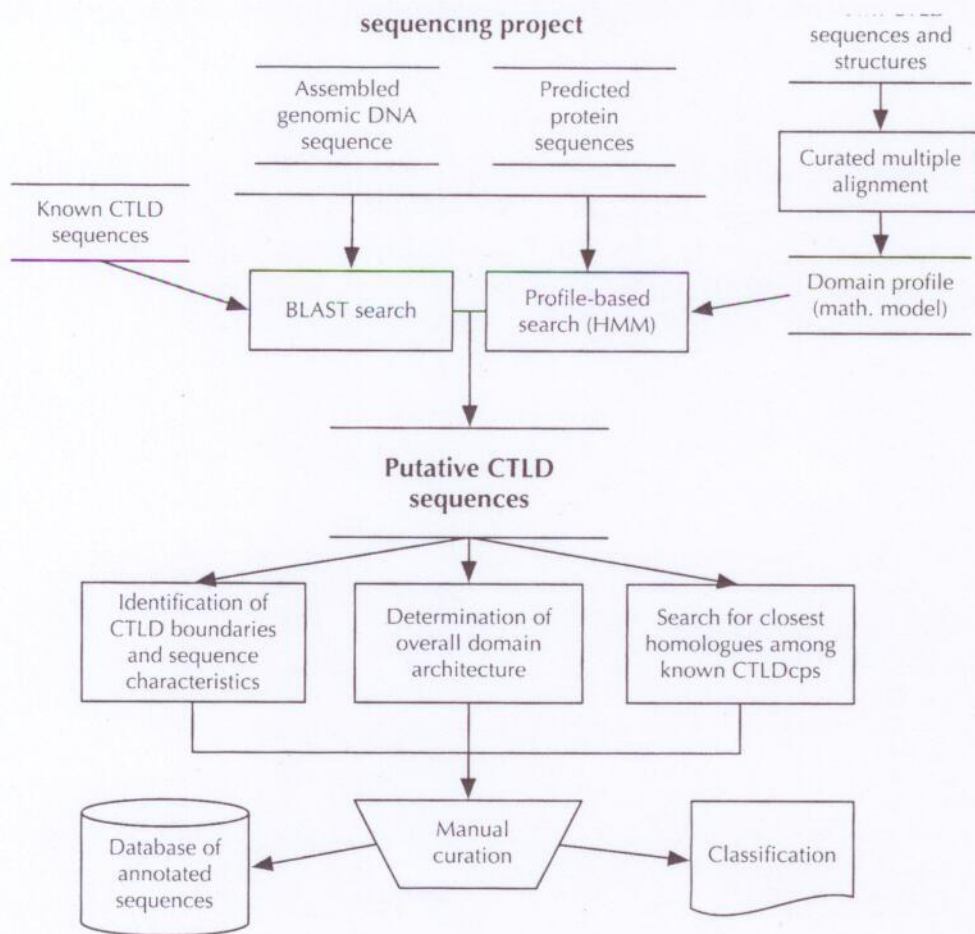
## 20.4 CTLs in the Genome Era

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

**20.5**  
**CTL Domain-Containing Analysis**

In the previous section (worm) and *F. rubi* common this abundant whole-genome analysis proteins are 'put to method is called shown in Figure 20



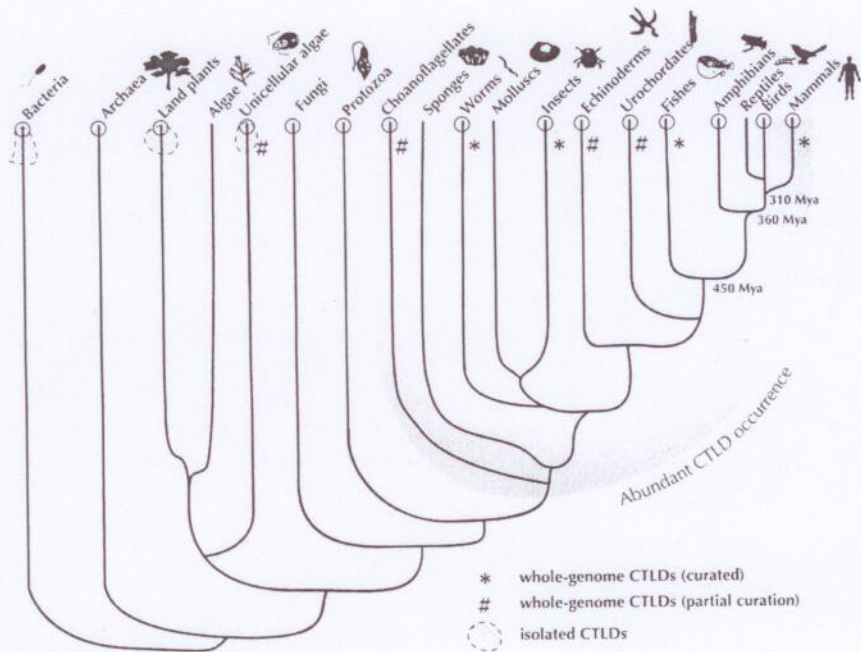
**Figure 20.5** The tree of divergence of CTLDcps. B genomes are shown in CTLDcps have been in partially curated analysis '#', respectively. The M



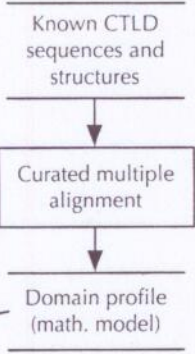
considered as 'work in progress', we will proceed to consider the main findings of the major genome studies and what they can tell us about the evolution of the superfamily.

**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-dashed circles. Evolution times are not drawn to scale.



for closest  
clades among  
CTLDCps

Classification

in summary form in  
analysis of the sequences  
initial assessment of  
carbohydrate and, if so,  
N/QPD and WND mo-  
20.3). This information  
of the annotation.

Assessing published  
annotating of genome  
and domain archi-  
evolutionary con-  
genomes should be

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 example of the hevein-like domain of the molecular structure that similar 'solu

- Evolutionary divergence and regulatory interactions. There are examples of events leading to development of r

We will illustrate the first mechanism in families in groups I CTLDcps (30 or 45' group II and V genes of group II and V  $\xi$  and 27. These clusters of the CTLD example is illustrated of genes is greatly structure and funct

An example of genes by CTLDcps with conserved proteins and fish a construction of the evolution which allows us to of the only carbohydrate Reg proteins of group lution is given in I

#### Info Box 1

An intriguing example of the recent reanalysis of 278. Most of the gene duplications CTLDs. It has been response towards sponse). Much of or 22% of the total The majority (74% not only in pathogen large set of CTLDcps 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.

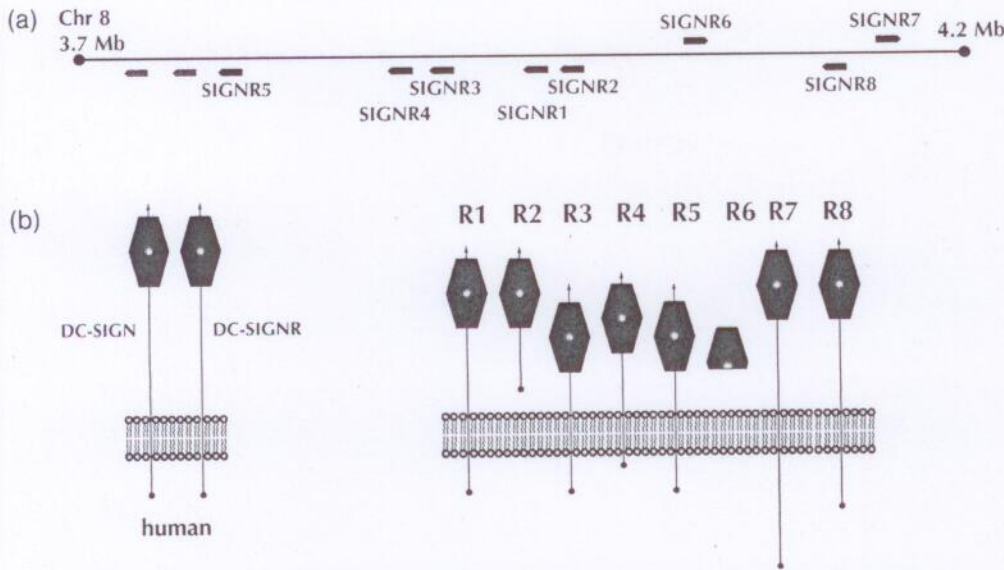
- Evolutionary diversification of function of CTLDcps has occurred by both genetic and regulatory mechanisms, consistent with the classic theory of Ohta [15]. There are examples showing gene duplication within groups, recombination events leading to new groups and mutational adaptation of CTLDs, but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

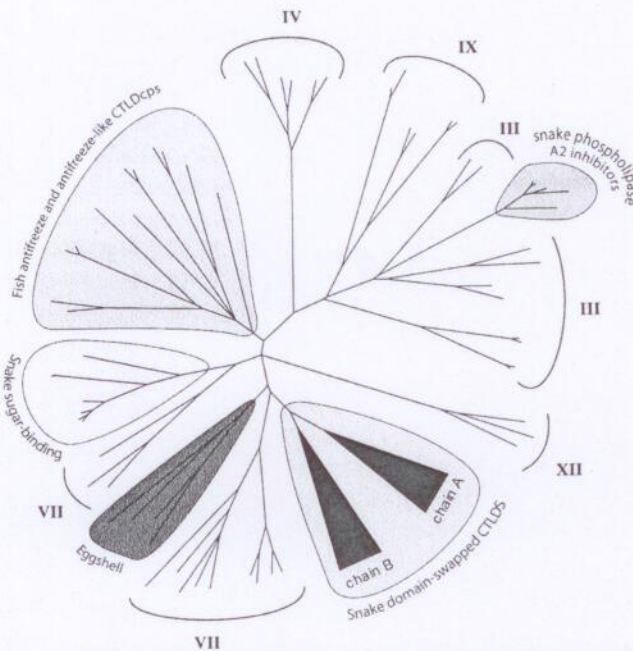
An example of gene recombination events in vertebrates is well demonstrated by CTLDcps with clade-specific functions, such as snake venoms, bird egg-shell proteins and fish antifreeze proteins. Figure 20.7 presents a phylogenetic reconstruction of the evolutionary history of the CTLDs of selected vertebrate 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 what CTLD had originated binding function. were consistent with but absent in non-metazoa (*Yersinia pseudotuberculosis*). Notwithstanding non-Metazoan CTL

Many of these involved in interactions with the immune system. [2] and African swine fever [2]. Well-conserved protozoan. The beta-proaerolysin) and from *Yersinia pseudotuberculosis* presented in Chapter their structures shown Info Box 2 for example contain well-conserved although some lack

How might these simultaneous explanations homologous to the hijacked host protein and structural divergence their origin. They

**Info Box 2**

Examples of free *Leeuwenhoekiella* a marine photosynthetic proteobacterium; rium. A particular genome sequence protein in the G and cadherin domain Metazoa (for further could be the function 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  $\text{Ca}^{2+}$ /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*?

SIGNR7 4.2 Mb  
SIGNR8

7 R8

ization and mem-  
s. Adapted from

phalipase

fish antifreeze  
ence similarity  
oup IV, IX, III,



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<sup>2+</sup>-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.

ins  
s mimicry of host

ing plant *Arabidopsis*  
ative CTLDcps also  
very recent reports  
Metazoan, the cho-  
nicellular green alga  
on of the Metazoan

eing reminiscent of  
genomes we have  
have evolved within  
containing a total of  
h (SRCR) domains,  
s for worm, fly and  
ess the conserved  
<sup>h</sup>-dependent carbo-

llection of CTLDcps  
etails of the branch-  
l an ongoing debate  
pparent absence of  
lthough Wheeler *et*  
es or divergence of  
e plausible explana-  
TLD genes by hori-  
atterns of CTLDcps  
volution starting at  
s to perform novel  
structurally flexible  
consistent with the  
evolutionarily very  
planation as to why

how understanding  
unfolded. We have

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

### 21.1 Molecular Basis of C

Carbohydrate–carb  
satile mechanism fi  
ticity of glycan cha  
sites, and to the cap  
cation forces (for  
through surfaces di  
covalent bonds (see  
Waals contacts, hyd  
(Figure 21.1a) [2].  
well defined by the  
dimensional (3D) s  
polar patches of car  
to the initial select  
and amino groups t  
solvating water mo  
13.1 for illustratio  
provide electrostat



#### 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  $\alpha$ 2,6-sialylated N-glycans of a bovine blastocyst (together with DNA and F-actin staining; please see Fig. 24.4 for details) and for  $\alpha$ 2,8-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.

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

**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>.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

**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-er  
teins cal  
mentary  
differen  
pharma-  
science.  
The m  
organic  
mainly v  
last two  
sugars a  
Sugar C  
welcome  
around t  
mentals  
nomena.  
treated in  
carbohyc  
lectinolo

The Sugar C  
Copyright ©  
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](http://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.*

[www.wiley-vch.de](http://www.wiley-vch.de)

ISBN 978-3-527-32089-9



9 783527 320899