Glycoconjugate Journal (1994) **11**:400–413

Molecular modelling of the *Dolichos biflorus* seed lectin and its specific interactions with carbohydrates: α-D-*N*-acetyl-galactosamine, Forssman disaccharide and blood group A trisaccharide

ANNE IMBERTY[1]*, FLORENCE CASSET[2], COLIN V. GEGG[3], MARILYNN E. ETZLER[3] and SERGE PÉREZ[2]

[1] *Laboratoire de Synthèse Organique-CNRS, Faculté des Sciences et Techniques, 2 rue de la Houssinière, 44072 Nantes cedex 03, France*

[2] *Ingénierie Moléculaire, INRA, BP 527, 44026 Nantes cedex 03, Nantes, France*

[3] *Department of Biochemistry and Biophysics, University of California, Davis, California 95616, USA*

Received 31 March 1994, revised 23 June 1994

The three-dimensional structure of *Dolichos biflorus* seed lectin has been constructed using five legume lectins for which high resolution crystal structures were available. The validity of the resulting model has been thoroughly investigated. Final structure optimization was conducted for the lectin complexed with αGalNAc, providing thereby the first three-dimensional structure of lectin/GalNAc complex. The role of the *N*-acetyl group was clearly evidenced by the occurrence of a strong hydrogen bond between the protein and the carbonyl oxygen of the carbohydrate and by hydrophobic interaction between the methyl group and aromatic amino acids. Since the lectin specificity is maximum for the Forssman disaccharide αGalNAc(1-3)βGalNAc-*O*-Me and the blood group A trisaccharide αGalNAc(1-3)[αFuc(1-2)]βGal-*O*-Me, the complexes with these oligosaccharides have been also modelled.

Keywords: molecular modelling; lectin; *Dolichos biflorus*; blood group

Introduction

The *Dolichos biflorus* lectins belong to a multigene family including the seed lectin (DBL), stem and leaf lectins (DB57 and DB58), and a root lectin (DB46) which also has a counterpart in the stem and leaves [1–4]. The spatial and temporal differential expression of these lectins provides a model for studying the role of these proteins in the plant [5].

The seed lectin from *Dolichos biflorus* was isolated and characterized more than 20 years ago [1] following reports that extracts from the seeds of this plant agglutinate blood group A erythrocytes [6] and precipitate blood group A substance [7]. The lectin is a 110 000 M_r tetramer [8] in which two of the subunits have been post-translationally altered by the proteolytic removal of approximately 10 amino acids from their carboxyl termini [9]. The lectin has two carbohydrate binding sites per tetramer and carbohydrate binding activity has been found only with the larger subunit [10]. The complete primary structure of the lectin was deduced from its cDNA sequence [11] and corrections to this structure were made after sequencing the lectin geonomic DNA [12].

The *Dolichos biflorus* seed lectin displays a very narrow specificity against α-D-GalNAc [1] and has little if any ability to recognize galactose [13]. The A-active disaccharide, α-D-GalNAc(1-3)D-Gal, and trisaccharide, α-D-GalNAc(1-3)β-D-Gal(1-3)D-GlcNAc, are equal to α-D-GalNAc-*O*-Me in their ability to inhibit the precipitation of blood group A substance with the lectin [1]. The A-type-2 pentasaccharide is a 1.7-fold better inhibitor than the above oligosaccharides, possibly due to the presence of the hydrophobic fucose on the subterminal Gal [1]. Subsequent studies showed the Forssman antigens, either as a pentasaccharide or the α-D-GalNAc(1-3)β-D-GalNAc disaccharide, to be much stronger inhibitors than the A-active pentasaccharide [14, 15]. The lectin has thus been classified as a blood group A specific and Forssman specific lectin [16].

Knowledge about the binding site of DBL has been limited because of the difficulty in obtaining crystals of this lectin, possibly due to the fact that it is a glycoprotein with

* To whom correspondence should be addressed.

0282-0080 © 1994 Chapman & Hall

one *N*-linked carbohydrate unit per subunit [17]. An alternate way to understand the binding of GalNAc and GalNAc-containing oligosaccharides by legume lectins is molecular modelling. Thanks to the well-established methodology for modelling antibodies, some antibody combining sites complexed with oligo- or polysaccharides have been studied by means of computer methods [18–20]. Meanwhile, modelling studies on lectins have been limited to the docking of monosaccharides in binding sites of known three-dimensional structures [21, 22], and more recently to the study of conformational preferences of complexed di- and trisaccharides in known binding sites [23, 24].

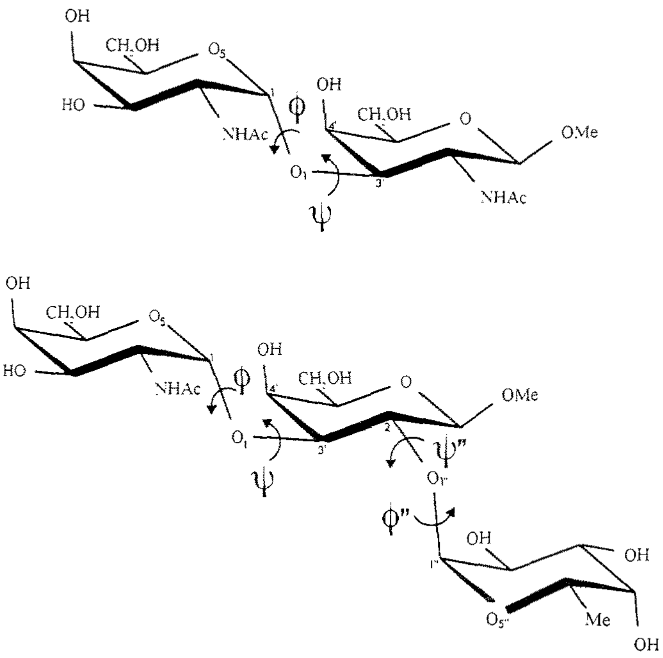
Despite their similar folding and reasonable range of homology, molecular modelling of legume lectin binding sites presents some peculiar difficulties [25] due: (i) to the large variations in the length and sequence of the loops in the perimeter of the binding site; and (ii) to the unusual conformations of these loops, mainly due to the presence of cations, which do not allow the search of canonical forms in other types of proteins of the Brookhaven Database. Because of this second difficulty, computer engineering of a new lectin has to rely only on other lectin crystal structures. Such a goal can now be envisaged thanks to: (i) the several crystal structures of legume lectins which have been published in recent years; and (ii) our progress in understanding and modelling interactions between proteins and carbohydrates [22, 24, 26].

The present paper deals with the modelling of DBL with the use of knowledge-based model building methods. Full optimisation of the protein, with α-D-GalNAc-*O*-Me docked in the binding site has been performed and insights into the role of the *N*-acetyl group have been gained. Complexes between DBL and two oligosaccharides, the Forsmann disaccharide and the blood group A trisaccharide (see Scheme 1) have also been optimized. Implications in the understanding of the fine specificity towards oligosaccharides are discussed.

Materials and methods

Lectin sequences and coordinates

The DBL sequence used in this study was deduced from the cDNA sequence [11], with inclusion of some recent corrections after sequencing the genomic DNA [12]. All the other lectin sequences were directly derived from their crystalline structures. The following crystal structures were extracted from the Protein Data Bank [27]: lens lectin (LCL) [28], *Griffonia simplicifolia* isolectin IV (GS4) complexed with Lewis b tetrasaccharide [29], *Erythrina corallodendron* lectin (ECorL) complexed with lactose [30]. Coordinates of concanavalin A (ConA) [31] and *Lathyrus ochrus* isolectin I (LOL) complexed with glucose [32] were kindly provided by their authors.



Scheme 1.

Hydrophobic cluster analysis and alignment of sequences

Hydrophobic Cluster Analysis (HCA) is a method to compare amino acid sequence [33] which is derived from the theory of Lim [34]. The method involves the drawing of the sequence of a theoretical α-helix where the hydrophobic residues form clusters. The shape, size and relative position of the clusters can be compared and the sequence similarity, when it exists, may be readily revealed. Conversion of the amino acid sequences into the 2D-helical plot required by the method was made using the HCA-Plot software [35].

Knowledge-based model building methods

The COMPOSER program, within the SYBYL software (SYBYL, TRIPOS), is a tool for knowledge-based homology modelling [36, 37]. Starting from the model protein sequence, the program checks a library of 3D structures for homologous proteins. In the second step, the program aligns the 3D-structures of the selected proteins, defining the so-called Structurally Conserved Region (SCR). An averaged structure of the SCRs is determined to serve as a framework for building the unknown structure [38]. The program then decides which of the homologues will be used to construct each SCR of the model and creates a framework for the unknown protein. The backbone of the model SCRs is generated and the side-chains are built using a rule-based procedure [39] with no manual intervention. The last step involves linking the different SCRs by building the variable regions. For each loop of the model, the program looks for a loop of the same length in one of the homologues, or, if necessary, for loop fragments with compatible length from other libraries of known structures [40].

Energy calculations and optimisation

Three-dimensional structures of GalNAc were taken from the monosaccharide data bank [41]. Several low energy conformations of the GalNAc containing di- and trisaccharides were constructed using the program MM3 [42, 43] (see Scheme 1). Glycan structures were then transferred into the SYBYL software [44] for all subsequent calculations. All energy calculations were performed with the TRIPOS force-field [45, 46]. New atom types have been added for carbohydrates which allow for the inclusion of energy parameters appropriate for protein/carbohydrate complexes [22]. Charges are calculated using the Pullman method for the protein part [47] in which each atomic charge is evaluated as the sum of the σ component calculated by Del Re method and the π component calculated by Hückel method. The carbohydrate atoms were given appropriate charges [22] and partial charges of 2e were allotted to the two cations. A distance dependent dielectric function was used for electrostatic interaction with a distance cut-off of 8 Å. Hydrogen bonding is taken into account in an implicit way by neglecting the van der Waals interactions between atoms which can act as hydrogen donor and acceptor.

The Powell method [48], which belongs to the conjugate gradient family, was used as energy minimizer in the MAXIMIN2 procedure of the SYBYL software. All minimizations were conducted up to a convergence gradient of 0.1.

Validation of the model

Stereochemical validations of the model were performed using the PROCHECK suite of programs [49]. This procedure is widely used to test the quality for protein crystal structures. Some parameters, which are indicators of stereochemical quality [50] were calculated for the model and for the homologous lectins and compared.

Knowledge-based molecular modelling of DBL

Revisiting the alignment of leguminous lectins

The knowledge-based design of proteins requires a very precise sequence alignment between the known structures, as well as between the unknown structure and the known ones. Alignment of leguminous lectins has been the topic of many investigations and, due to the occurence of some very conserved regions, the proposed alignments are very similar [25, 51, 52]. However, some discrepancies remain in the proposed locations for short insertions and deletions. Such discrepancies could be of importance in our molecular modelling of lectins since it has been recently shown that the areas of greatest variability are located in the carbohydrate-binding site regions, forming a perimeter around a well-conserved core [25]. It was also necessary to revisit the alignment of DBL with related lectins, since the sequence used here had received some recent corrections [12].

Table 1. Similarity matrix (% of identity) for six leguminous lectins as calculated from the alignment of Fig. 2.

	LOL	*LCL*	*GS4*	*ECorL*	*DBL*
ConA	38.3	39.9	39.7	42.2	39.6
LOL		88.5	41.4	47.1	50.2
LCL			41.7	37.4	50.4
GS4				38.0	42.5
ECorL					45.5

Alignment was then reconsidered by the HCA method [33], which allows for a rapid visual identification of the cluster and an easy alignment (Fig. 1). The shapes and locations of clusters are absolutely identical for LOL and LCL which share more than 80% identity. For the other lectins, the identification of the clusters is straightforward even if there are some variations in cluster shapes. The alignment was checked on the crystalline structure for the first five lectins of Fig. 1 and the extensions of the β-strands are indicated on the HCA plots. For the HCA plot of DBL, the extension of the β-strands was deduced from the other plots. It should be noted that the HCA plot of DBL is not identical to any other, but, depending on the sequence fragment, it displays some strong similarities with one of ConA, LOL and LCL, or ECorL.

The alignment deduced from the HCA plots is displayed in Fig. 2. This alignment differs only slightly from published data and the differences are mainly located in close proximity to the binding site, that is in the hypervariable regions [25]. The amino-acids directly involved in the binding of carbohydrates have been contoured in Fig. 2. This clearly shows that the loop from Gly217 to Asp221 is particularly involved in the galactose specificity of ECorL.

The identity matrix calculated from this alignment is displayed in Table 1. The two Viciae lectins, LOL and LCL, exhibit almost 90% identity. For the other ones, the identity scores range between 37% and 50%. DBL displays between 40% and 50% identity with all the other lectins, which indicates good conditions for model building. It has to be noted, that although DBL belongs to the same Phaseoleae tribe as *Erythrina* and both lectins are specific for the *galacto* configuration, the homology is higher with LOL and LCL. This indicates that the use of several homologous known structures should be a great help for building a model.

Building and refining the DBL model

Constructing the conserved regions of DBL

The crystal structures of five lectins: LOL/mannose, ConA, LCL, ECorL/lactose and GS4/Le^b tetrasaccharide, were gathered in a library. The knowledge-based molecular builder COMPOSER [36, 37] first searches for 'seed' residues; these are residues which are identical in the sequences of the lectins included in the library. These

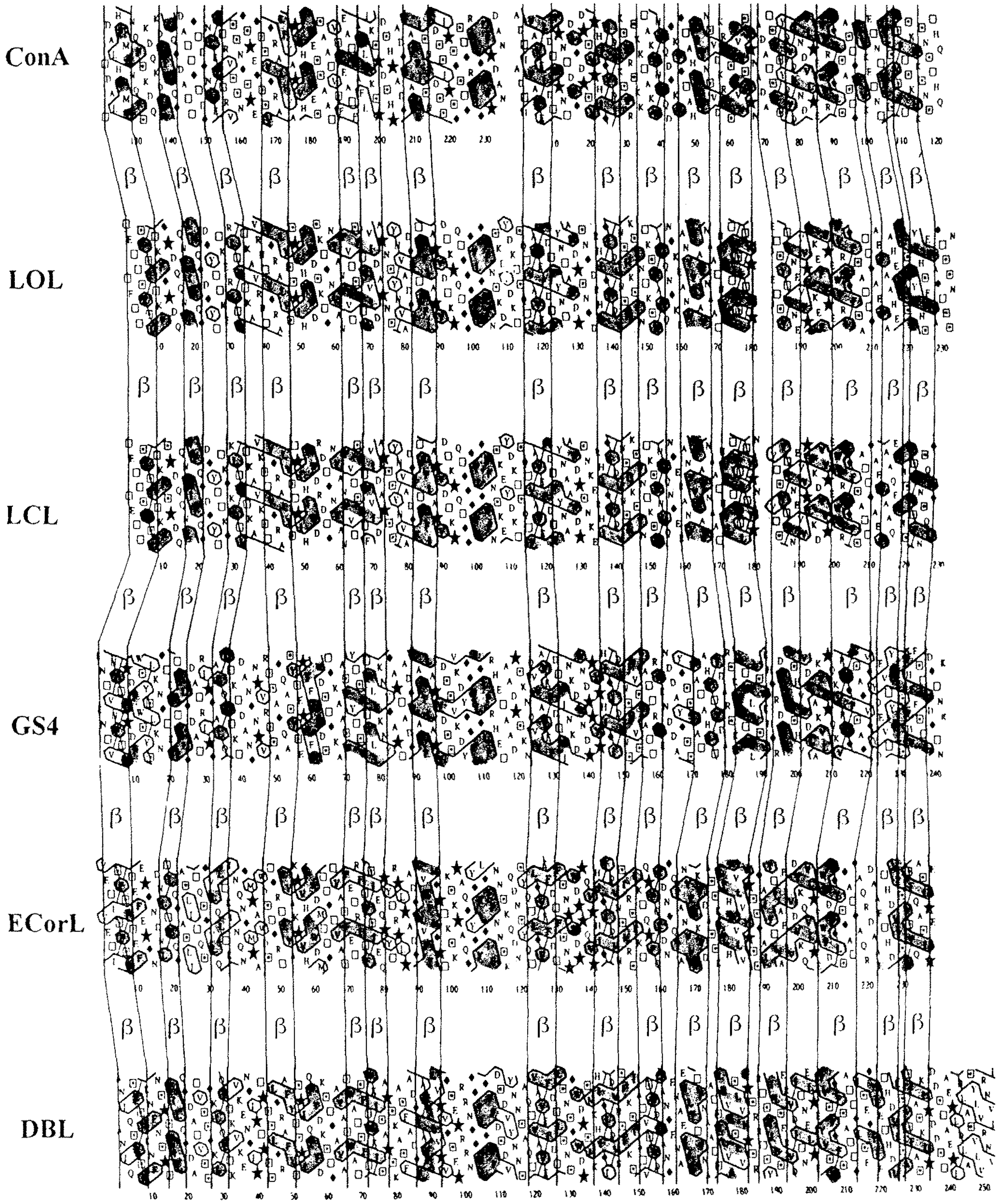


Figure 1. HCA plot of six lectins. Four amino acids are conventionally represented by special symbols (★: proline; ◆: glycine; □: threonine; ⊡: serine). Vertical lines have been inserted to delimit the extension of β-strands. The most conserved hydrophobic clusters have been shadowed for better visualization. For the sake of plot alignment, ConA sequence has been cut and rearranged.

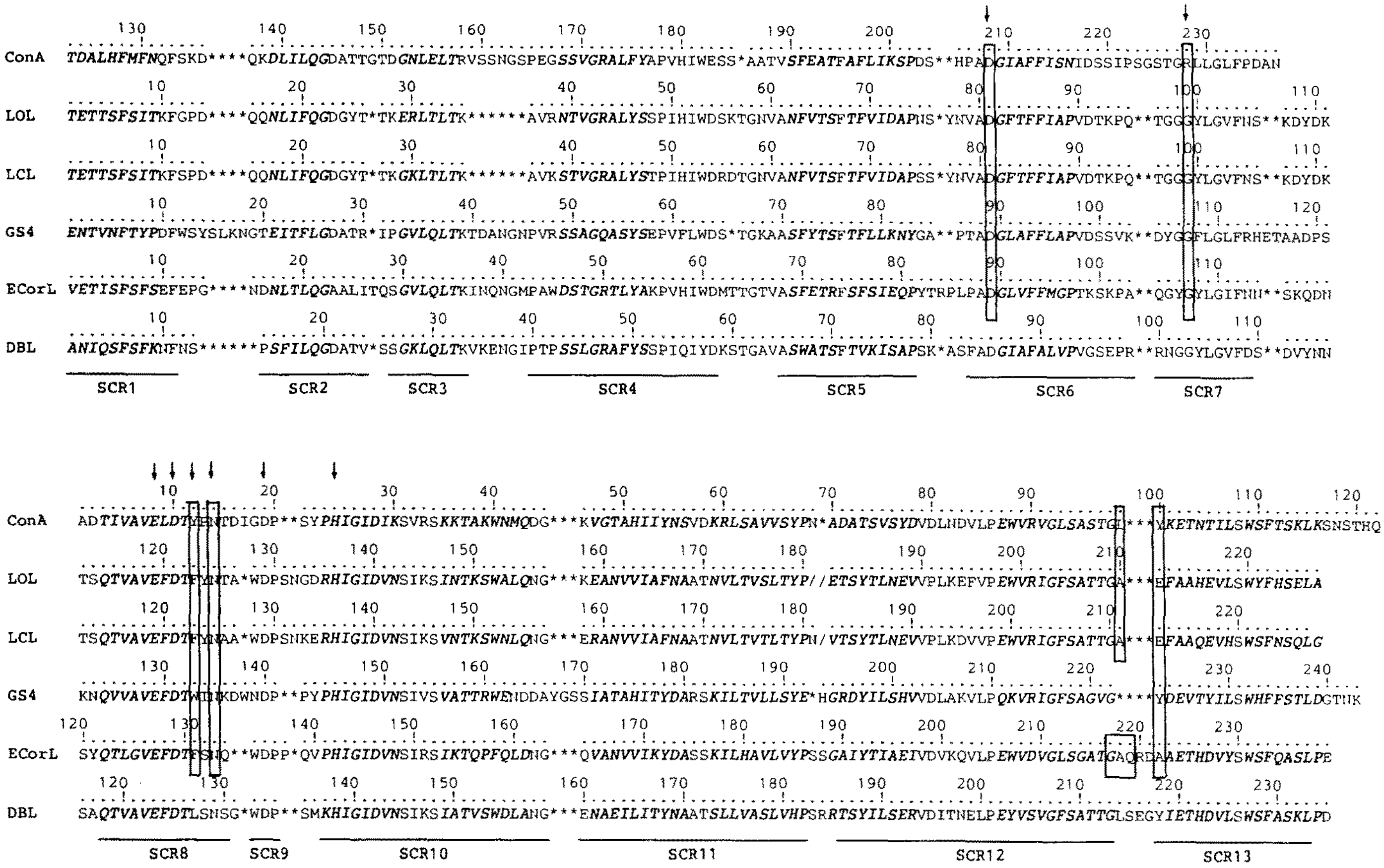


Figure 2. Sequence alignment of six lectins as deduced from the HCA plots. Deletions have been indicated by *. Amino acids belonging to *β*-strands have been represented by bold letters. The sequences defining the structurally conserved regions used for model building have been underlined under the plot. Amino acids directly involved in the binding of calcium and manganese have been indicated by arrows. The amino acids that bind carbohydrate in the known three-dimensional structures have been contoured.

residues are used to align the known three-dimensional structures and then define the structurally conserved regions from the alignment. At this stage, the alignment resulting from the HCA plot was used to check and correct the seed residues. The SCRs determined by the program are displayed in Fig. 2. Due to the very conservative folding of leguminous lectins, the SCRs cover about 85% of the sequences. From these SCRs, the DBL 'framework' was built by the program (Fig. 3). This framework, consisting only of the α carbons, is built from SCRs of the lectins library. For each SCR of the model, the most homologous one is selected from the known structures. The origin of SCRs is as follows for the DBL model: SCR2 was selected from ConA, SCR1, SCR12 and SCR13 were selected from ECorL and the other nine SCRs were either from LOL or from LCL. None of the GS4 fragments were used to build the DBL model.

Building the loops of DBL

After generating the side-chains, the most tenuous part of the model building is to complete the protein model by constructing the loop regions and special care has to be taken since several loops are in the perimeter of the carbohydrate binding site and play a role in the specificity. The program selected most loops from homologous regions of LOL and LCL. In three loops, ECorL was selected because of deletion in the other sequences (loops Val33-Ile39, Ser183-Arg185 and Leu214-Gly218). Only loop Asn12-Ser13 could not be modelled from the lectins since it corresponds to a deletion which occurs only in DBL. In this case, the search was performed in the Protein Data Bank available in the SYBYL program. Several loops were proposed and the one having the best energy was selected. It has to be noted that this loop occurs in the contacting region between monomers and does not play any role in the binding of carbohydrate. Each loop, or each attachment region between SCRs (i.e. zero length loop), was roughly optimized. A 'hot' region of a few amino acids and an 'interesting' neighbouring region were defined for further energy minimization of this subset.

Refining the DBL model

Prior to any further refinement, the location of the manganese and calcium ions were determined from the

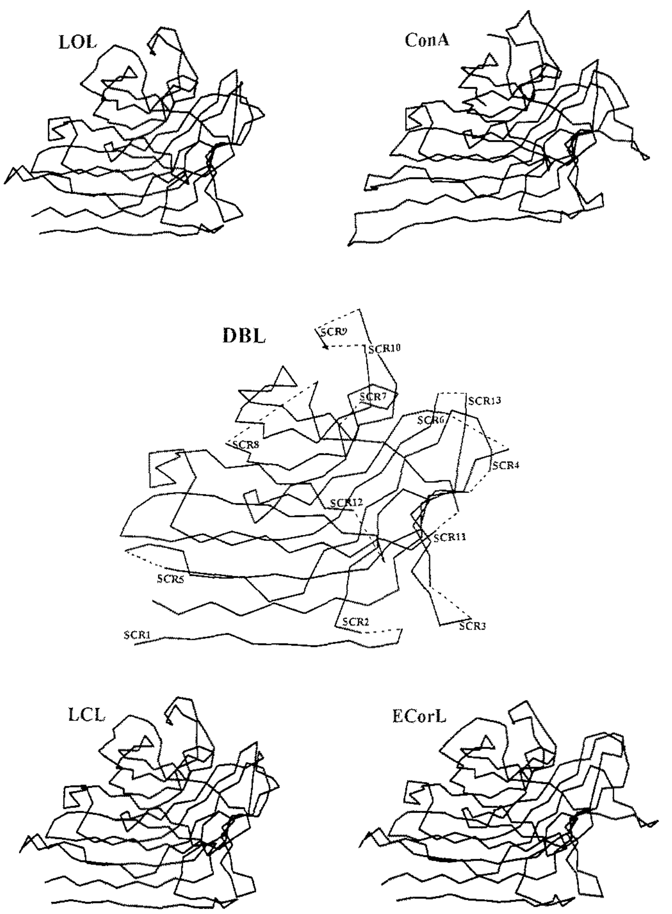


Figure 3. Framework of the backbone of the DBL model as built from the Structurally Conserved Regions of four crystal structures of legume lectins. At this stage of building, variable loops are not yet determined and are indicated by dotted lines in the framework.

ECorL crystalline structure [30], They were carefully located in our model, with respect to the same alignment of the backbone. Van der Waals energy parameters, for 6–12 potential, were set as follows: the values proposed in the TRIPOS force field for calcium atom were kept unchanged ($r = 1.2$ Å and $\varepsilon = 0.6$ kcal mol^{-1}) whereas manganese was given values of $r = 1.8$ Å and $\varepsilon = 0.4$ kcal mol^{-1} by analogy with its size and with equivalent ions in the TRIPOS force-field. To avoid collapsing of the binding site due to the absence of the water molecules, a GalNAc residue was modelled in the DBL binding site in the same position as the galactose in the corresponding amino acids of ECorL.

In this complex, all the hydrogen atoms were added and their positions were optimized. The second step was the geometry optimization of all the side chains. In the last ystages of the optimization procedure, the whole molecule was optimized in several cycles of energy minimization with decreasing constraints on the torsion angles of the backbone. Constraint forces were slowly decreased from 2 kcal mol^{-1} $^{\circ 2}$ down to 0.05 kcal mol^{-1} $^{\circ 2}$. The overall procedure of minimization took about 1 week of calculations on a Silicon Graphics with a R4000 processor.

Table 2. Structural variations between the three-dimensional model of DBL and the five known three-dimensional structures of leguminous lectins. Root mean squares (Å) have been calculated from all the backbone atoms of the structurally conserved regions (SCR) as indicated in Fig. 2.

	LOL	*LCL*	*GS4*	*ECorL*	*DBL*
ConA	0.91	0.96	1.53	1.00	0.93
LOL		0.35	1.34	0.69	0.61
LCL			1.36	0.72	0.62
GS4				1.42	1.38
ECorL					0.69

Validation and description of the DBL model

Stereochemical quality of the model

The model of DBL can be compared to other lectin structures to assess its validity. First, the variations in the backbone geometry can be estimated between the different structures (Table 2). Root mean squares (rms) have been calculated using all the non-hydrogen atoms of the backbone from the structurally conserved regions as defined in Fig. 2. The five three-dimensional structures from x-ray crystallography display rms of less than 1 Å, with the exception of the GS4 lectin which differs greatly from the others and will not be considered further. The rms's between the DBL model and the known structures are in the same range: less than 0.7 Å when compared with LOL, LCL and ECorL, and less than 1 Å when compared with ConA. In our model of DBL, one peptide bond is in the cis conformation (Ala84), a feature which is common and homologous to all the known three-dimensional structures of legume lectins (Ala207 in ConA, Ala80 in LOL and LCL, Ala88 in ECorL).

Stereochemical quality of the model can be evaluated by the use of the PROCHECK program [49]. The quality of the backbone geometry can be assessed by the scattering of the Φ and Ψ torsion angles of the peptide linkages on a Ramachandran map. With the exception of glycine, the values of the Φ and Ψ torsion angles should belong to a well defined zone in the diagram. For the four crystal structures, more than 85% of the peptide linkages are located in the core region of the Ramachandran diagram and none of them (with the exception of two amino acids of ConA) are in the disallowed region. The model structure of DBL also displays an excellent quality of the backbone since 89% of the residues are in the core region of the Ramachandran diagram, 21% in the allowed region and none in the 'generously allowed' or disallowed areas.

For further analysis of the backbone, three parameters,

Table 3. Stereochemical characteristics of the three-dimensional model of DBL and of the four legume lectins which were used to build the model.

	ConA	*LOL*	*LCL*	*ECorL*	*DBL*
Code PDB			2LAL	1LTE	
Resolution (Å)	1.75	2.0	1.8	2.0	
R-factor	0.167	0.182	0.184	0.190	
Refinement method	FFLS[a]	XPLOR	RESTRAIN	PROLSQ XPLOR	This work
Main chain					
Φ and Ψ %[b]	85.9	91.0	86.1	93.1	89.3
ω SD[c]	10.1	5.7	9.3	2.8	5.6
ζ SD[d]	9.0	0.9	2.8	0.8	2.1
Side chain					
χ_1 SD[e]	19.6	18.2	19.7	16.1	12.5
χ_2 SD[f]	17.5	17.3	18.4	17.6	11.0

[a] Fast Fourier least squares procedure. Geometric constraints were also applied to the refined coordinates [31].
[b] Percentage of Φ and Ψ values in the best region of the Ramachandran diagram.
[c] Deviation of peptide linkage from planarity.
[d] Deviation of Cα chirality.
[e] Deviation from staggered orientation.
[f] Deviation from trans orientation.

which have been found to be good indicators of stereochemical quality [50], are listed in Table 3. These indicators of the main chain goodness are the percentage of residues belonging to the best region of the Ramachandran diagram the deviation of planarity for the peptide linkage and the deviation of chirality for the Cα. Even though the four lectin crystal structures were solved at high resolution (1.75–2.0 Å), they display quite different parameter values. This is not due to the accuracy of the structure but rather to the refinement procedure [50]. Energy-based refinement methods, such as XPLOR [53], would lead to better stereochemical features than the least-squared methods, such as Fast Fourier least-squares procedure [54] or the more recent programs RESTRAIN [55] and PROLSQ [56]. In fact, it is quite common to use both types of methods in a refinement procedure. The DBL model refined with Tripos force-field with very weak constraints on the backbone torsion angles, displays good quality for the three parameters calculated here.

Quality of side chains geometry has been evaluated by the standard deviation of the χ_1 torsion angles about the three staggered orientations *gauche+*, *trans* and *gauche−*, and also of the χ_2 torsion about the *trans* orientation. The crystal structure displayed good geometries, in the usual range, whereas the DBL model exhibits even higher quality, due to the unrestrained energy minimization of the side-chains.

Physicochemical and biochemical quality of the model

Some physicochemical and biochemical features of the DBL model are summarised on Fig. 4. The folding displays large β-sheets with extensive inter-strands hydrogen bonding. The amino acids around the two cations are very well conserved, when compared to other lectins. They display the same coordination shell around the Ca^{2+} and Mn^{2+} ions. Since DBL is an *N*-glycoprotein [17], another validation of the model is the location of the glycosylation consensus sequence Asn114-Asn115-Ser116. As indicated by an arrow in Fig. 4, Asn114 is located on the top of a loop, allowing therefore for the presence of an *N*-glycan. This feature has not been induced by the modelling procedure, since none of the 3D-structures used for building

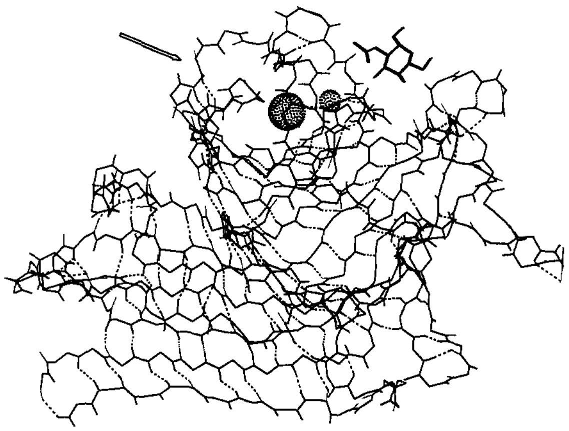


Figure 4. Optimised three-dimensional model of DBL along with the GalNAc residue docked in the binding site. The side chains of the protein are not displayed nor the hydrogen atoms of the glycan moiety. Hydrogen bonds are represented as dotted lines. Cations are drawn with their van der Waals surfaces. The arrow indicates the location of the consensus sequence for glycosylation.

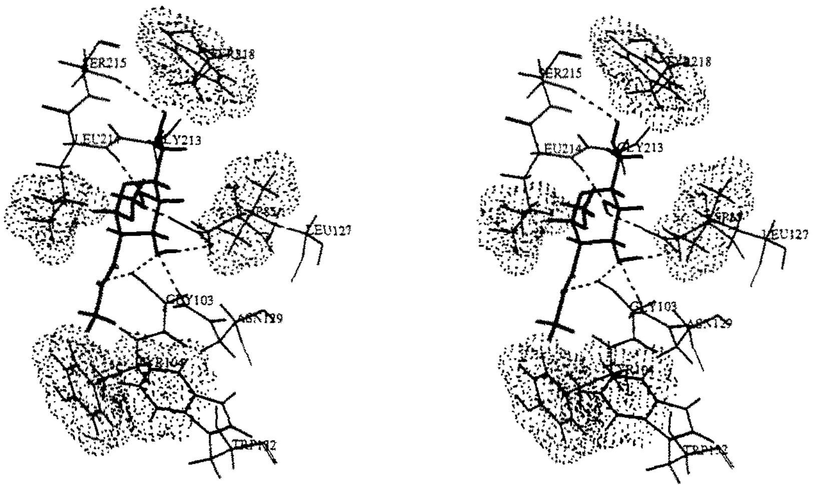


Figure 5. Stereodrawing of α-D-GalNAc-*O*-Me (bold lines) in the modelled binding site of DBL. Hydrogen bonds are indicated by dotted lines. In order to display apolar interaction, aromatic groups and methyl groups of amino acids are displayed surrounded by dotted suraces calculated using 85% of the van der Waals radius.

DBL contains this particular *N*-glycoslation site. The ability of the lectin to associate as a dimer or tetramer would also be a validation of the model. From Figs 3 and 4, it appears that the six strands β-sheet, which is arranged back to back with another like β-sheet in the crystal structure of ConA and LCL, would be conserved in the DBL model. However, it should be pointed out that our model lacks 18 amino acids of the carboxy terminal peptide which exists only in the *Dolichos biflorus* lectin family. This peptide could extend into the domain of the canonical ConA interface and prevent such intermolecular association. Another biochemical validation of the model is the ability to bind carbohydrates. The detailed description of the GalNAc binding site is given in the following section.

Modelling the interactions between DBL and carbohydrates

DBL/GalNAc interaction

Prior to any refinement, the GalNAc residue has been given an orientation similar to that of the galactose moiety in the crystal structure of ECorL/lactose complex [30]. The docking of α-D-GalNAc-*O*-Me in the binding site of the DBL model has been performed by full optimization without any constraint in a 6 Å sphere around the glycan (21 amino acids) while the atoms in a 15 Å sphere (91 amino acids) were kept fixed, but taken into account for energy calculations.

The overall orientation of the GalNAc residue remains unchanged during optimization and many features can be readily compared with those displayed by galactose in ECorL. As shown in Fig. 5, seven hydrogen bonds occur between the glycan and the amino acids of the binding site. These hydrogen bonds, along with their geometrical characteristics are listed in Table 4. Six out of the seven hydrogen bonds are homologous to those formed by the binding of galactose by ECorL. The amino acids Asp85, Gly103 and Asn129 which occupy similar positions in the binding site of other legume lectins, establish hydrogen bonds with the O3 and O4 atoms of GalNAc. Leu214 and Ser215 form hydrogen bonds with O4 and O6, respectively. These two amino-acids occupy the analogous positions to Ala218 and Gln219 in ECorL. The length and the shape of this loop are particular to lectins binding glycans with a *galacto* configuration. The seventh hydrogen bond involves the carboxyl oxygen of GalNAc *N*-acetyl group and the NH of Gly103 backbone. This atom would therefore be creating a bifurcated or three-centred hydrogen bond since it is already bonded to O3 of GalNAc. Such three-centred hydrogen bonds are very common in amino acid structures due to proton deficiency [57].

Van der Waals forces also contribute significantly to the stability of protein-carbohydrate interactions. Among these forces, hydrophobic interactions which involve stacking with aromatic residues have been shown to be of great importance [58]. All the known structures of legume lectins display such a stacking between the apolar face of the glycan and an aromatic residue in a conserved position (Tyr12 of ConA, Phe131 of ECorL). In DBL, this aromatic residue is substituted by a leucine. As seen in Fig. 5, the apolar interaction occurs between the methyl group of Leu127 and the apolar face of GalNAc created by the methine hydrogens H3, H4 and H5. Other van der Waals interactions involving the apolar hydrogen of the carbohydrates, are listed in Table 5. This is clearly visible in Fig. 5 and also in Plate 1 where the Conolly surface of the binding site [59] has been calculated and represented with colour coding of the hydrogen bonding potential using the MOLCAD option of the SYBYL software [44]. A second apolar region of the carbohydrate, involving H1 and H2, interacts with Leu214. An important stabilization of the complex is due to the

Table 4. Description of the hydrogen bonds between the DBL binding site and αGalNAc monosaccharide, Forssman disaccharide and blood group A trisaccharide.

Name	*Donor*	*Acceptor*	*Dis D..A*[a] (*Å*)	*Dist H..A*[b] (*Å*)	*Angle D-H..A*[c] (°)
αGalNAc	Gly103 NH	O3	3.3	2.4	148
	Asn129 NH	O3	3.2	2.2	160
	O3H	Asp85 OD2	3.0	2.1	168
	O4H	Asp85 OD1	2.6	1.8	141
	Leu214 NH	O4	2.9	2.0	146
	O6H	Ser215 O	3.2	2.7	111
	Gly103 NH	O7	3.3	2.4	148
Forssman disaccharide	Gly103 NH	O3	2.8	2.2	116
	Asn129 NH	O3	3.1	2.2	162
	O3H	Asp85 OD2	3.0	2.1	165
	O4H	Asp85 OD1	2.6	1.8	142
	Leu214 NH	O4	2.9	2.0	147
	O6H	Ser215 O	3.2	2.7	113
	Gly103 NH	O7	3.1	2.3	148
Blood group A trisaccharide TriA_conf1	Gly103 NH	O3	2.9	2.3	116
	Asn129 NH	O3	3.1	2.2	159
	O3H	Asp85 OD2	3.0	2.1	163
	O4H	Asp85 OD1	2.7	1.8	144
	Leu214 NH	O4	3.0	2.1	147
	O6H	Ser215 O	3.3	2.8	115
	Gly103 N	O7	3.1	2.2	149
	O2″H	Ser128 O	2.9	2.1	147
TriA_conf2	Gly103 NH	O3	2.9	2.3	118
	Asn129 NH	O3	3.0	2.1	160
	O3H	Asp85 OD2	2.9	2.0	161
	O4H	Asp85 OD1	2.6	1.8	146
	Leu214 NH	O4	3.0	2.1	147
	O6H	Ser215 O	3.2	2.7	118
	Gly103 NH	O7	3.2	2.4	146
	O4″H	Ser128 O	2.7	1.9	137

[a] Dist D..A, distance between donor and the acceptor.
[b] Dist H..A, distance between the hydrogen and the acceptor.
[c] Angle D-H..A, angle between donor, hydrogen and acceptor.

interaction of the methyl of the GalNAc *N*-acetyl group with both Tyr104 and Trp132 aromatic amino acids.

DBL/Forssman disaccharide interaction

The conformational behaviour of the α-D-GalNAc(1-3)βD-GalNAc-*O*-Me disaccharide was studied using the 'relaxed map' approach (see [60] for a review of methods) coupled with the molecular mechanics program MM3 [42, 43]. As indicated in Scheme 1, orientation of the glycosidic linkage can be defined by the values of the torsion angles $\Phi = \Theta$(O-5_C-1_O-1_C-3′) and $\Psi = \Theta$(C-1_O-1_C-3′_C-4′). Three different low energy conformations were determined by the MM3 molecular mechanics program: F1 ($\Phi = 80°$, $\Psi = 80°$), F2 ($\Phi = 100°$, $\Psi = 160°$) and F3 ($\Phi = 80°$, $\Psi = -60°$); the three were tested for interactions with DBL. Each conformation was located on the surface of the protein with its terminal α-D-GalNAc in the binding site with the same position and orientation for the monosaccharide in the refined complex. Optimization was performed in the same way as above, using a 'hot' region and an 'interesting' region.

None of the three conformations generate steric conflicts, but the F1 conformation ($\Phi = 80°$, $\Psi = 77°$ after optimization) yielded a better stabilization of the complex in terms of energy of interaction. This conformation, which corresponds to the lowest energy for the disaccharide in the isolated state, displays a weak hydrogen bond between the O4 oxygen of the reducing residue and the NH hydrogen of the other one. This complex is depected in Fig. 6 and Plate 1, whereas the hydrogen bonds and apolar contacts

Table 5. Hydrophobic contacts between DBL and αGalNAc, Forssman disaccharide and blood group A trisaccharide.

GalNAc	*DBL*	*Forssman*	*DBL*	*TriA_conf1*	*DBL*	*TriA_conf2*	*DBL*
C1H	MeCD1 Leu214	C1H	MeCD1 Leu214	C1H	MeCD1 Leu214	C1H	MeCD1 Leu214
C2H	MeCD1 Leu214 HCG	C2H	MeCD1 Leu214 HCG	C2H	MeCD1 Leu214 HCG	C2H	MeCD1 Leu214 HCG
AcMe	HCE2 Tyr104	AcMe	HCE2 Tyr104	AcMe	HCE2 Tyr104	AcMe	HCE2 Tyr104
AcMe	HCH2 Trp132	AcMe	HCH2 Trp132	AcMe	HCH2 Trp132 HCZ3	AcMe HCZ3	HCH2 Trp132
C3H	MeCD1 Leu127	C3H	MeCD1 Leu127	C3H	MeCD1 Leu127	C3H	MeCD1 Leu127
C4H	MeCD1 Leu127	C4H	MeCD1 Leu127 HCG	C4H	MeCD1 Leu127 HCG	C4H	MeCD1 Leu127 HCG
C5H	MeCD1 Leu127	C5H	MeCD1 Leu127	C5H	MeCD1 Leu127	C5H	MeCD1 Leu127
C6H	HCB Tyr218	C6H	HCB Tyr218	C6H	HCB Tyr218	C6H	HCB Tyr218
		C6H	MeCD2 Leu128	C6H	MeCD2 Leu127	C1″H	MeCD1 Leu127
		Ac″Me	MeCD1 Leu127	C3″H	MeCD2 Leu127	C2″H	MeCD1 Leu127
			MeCD2	C4‴H	HCE2 Tyr218		MeCD2 Leu127
						C6″Me	HCB Ser130

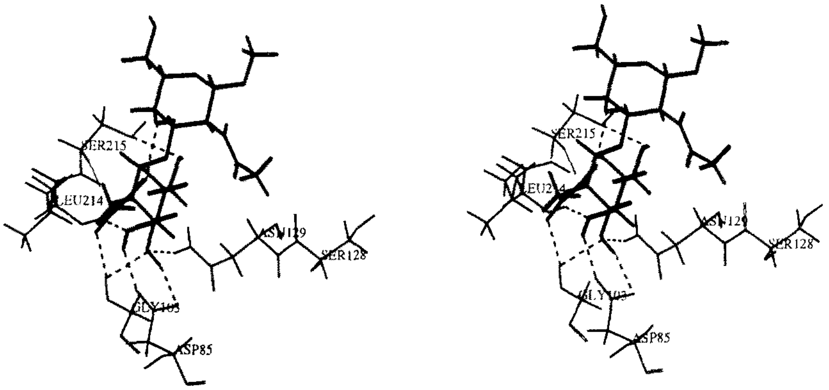


Figure 6. Stereodrawing of the Forssman disaccharide in the modelled binding site of DBL. Only amino acids hydrogen bonded to the carbohydrate are displayed. Hydrogen bonds are indicated by dotted lines.

are listed in Tables 4 and 5. The terminal α-D-GalNAc interacts with DBL with the same polar and apolar contacts as described above. The reducing β-D-GalNAc creates a new region of interaction on the other side of Leu127. Only the methyl hydrogens of its *N*-acetyl group interact with the methyl group of Leu127 creating therefore a strong apolar contact between patches of hydrogen atoms. This type of interactions, favourable in terms of energy but limited to the terminal atoms of the *N*-acetyl group, is in agreement with affinity experiments. Only the Forssman disaccharide has higher affinity when compared to the α-GalNAc monosaccharide. The α-GalNAc(1-3)Gal terminal disaccharide of blood group A, which lacks this *N*-acetyl group, does not show any increase in affinity.

It appears that the Forssman pentasaccharide, αGalNAc-(1-3)βGalNAc(1-3)αGal(1-4)βGal(1-4)α,βGlc, is the glycan with highest affinity for DBL [14]. From the model proposed here, it seems that galactose residues added on the reducing end of the second GalNAc would extend in the solvent and therefore would not directly interact with the protein surface. Such a proposal is not in contradiction with the observed affinity enhancement since several factors could play a role. First, water molecules can be arranged between the protein surface and distant glycan residues, creating therefore a favourable interaction as it has been observed in several protein/carbohydrate complexes [61]. Second, the longer oligosaccharides could have less conformational flexibility at the terminal αGalNAc(1-3)GalNAc linkage. In this case, the entropy term will be reduced in the thermodynamics of binding [62] and the affinity will be enhanced.

DBL/blood group A trisaccharide interaction

For blood group A trisaccharide, low energy conformations were taken from a recent molecular mechanics study [63]

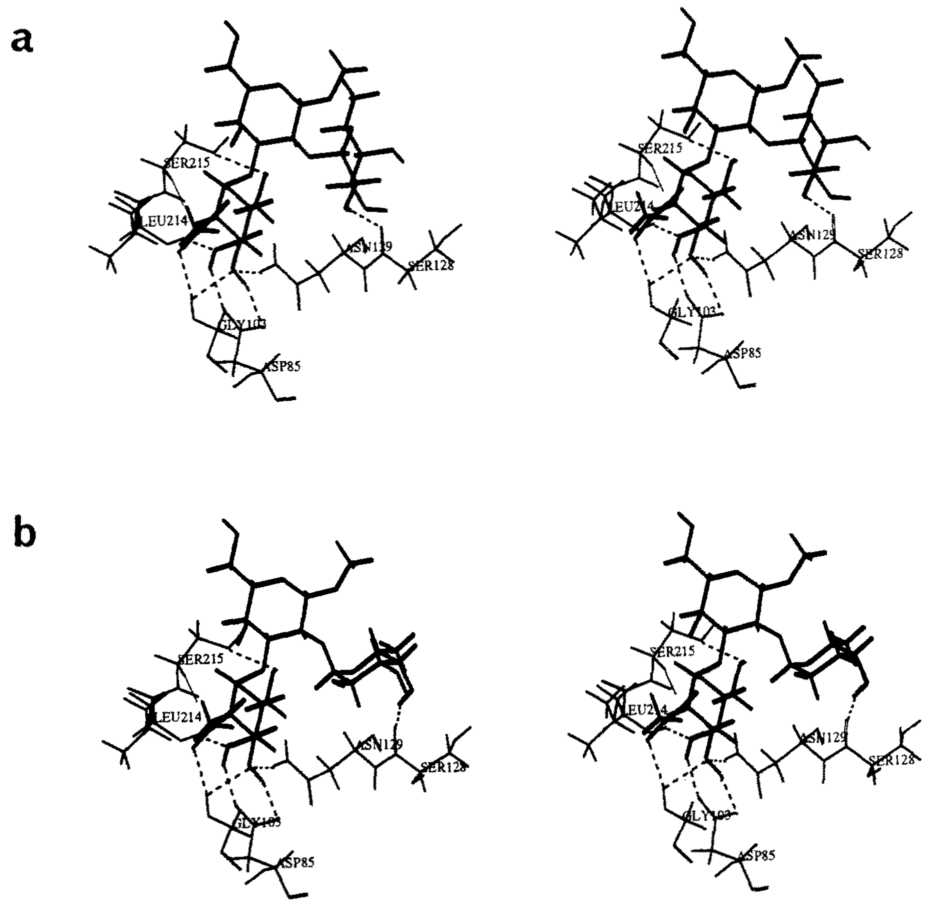


Figure 7. Stereodrawing of the blood group A trisaccharide in the modelled binding site of DBL. Two conformations of the trisaccharide are displayed: (a) TriA_conf1; and (b) TriA_conf2. Only amino acids hydrogen bonded to the carbohydrate are displayed. Hydrogen bonds are indicated by dotted lines.

which included an exploration of the whole conformational space defined by the four torsion angles: $\Phi = \Theta(\text{O-5\_C-1\_O-1\_C-3}')$ and $\Psi = \Theta(\text{C-1\_O-1\_C-3}'\text{\_C-4}')$ at the α-D-GalNAc(1-3)β-D-Gal-*O*-Me glycosidic linkage and $\Phi'' = \Theta(\text{O-5}''\text{\_C-1}''\text{\_O-1}''\text{\_C-2}'')$ and $\Psi'' = \Theta(\text{C-1}''\text{\_O-1}''\text{\_C-2}'\text{\_C-3}')$ at the α-L-Fuc(1-2)β-D-Gal-*O*-Me glycosidic linkage (see Scheme 1). Several low energy conformations were described and the two best ones are considered here: TriA_conf1 ($\Phi = 71°$, $\Psi = 76°$, $\Phi'' = -95°$, $\Psi'' = -165°$) and TriA_conf2 ($\Phi = 70°$, $\Psi = 68°$, $\Phi'' = -69°$, $\Psi'' = -114°$). These two conformations were located on the DBL surface with the terminal GalNAc in the binding site and these complexes were optimized.

These two conformations differ strongly for the orientation of the Fuc(1-2)Gal linkage. Both give very stable complexes with DBL and it is not possible to favour one over the other in terms of energy. Therefore, both solutions are presented in Fig. 7 and Plate 1 and the characteristics of their interactions are listed in Tables 4 and 5. In both cases, the trisaccharide has a tendency to wrap around Leu127. This conformational behaviour of an oligosaccharide bonded to a legume lectin has been observed in the complex LOL/octasaccharide [64] where the glycan folds around Phe123 which has the same location as Leu127 in DBL. For both conformations, this folding around Leu127 allows the fucose residue to come in close contact with the lectin, whereas the galactose has has no contact at all with any amino acid.

The first conformation (TriA_conf1: $\Phi = 66°$, $\Psi = 77°$, $\Phi'' = -79°$, $\Psi'' = 175°$) brings the fucose residue in close contact to Tyr218. Two regions of hydrophobic contact, involving methine hydrogens of the fucose residue, are created: one with Leu127 and one with Tyr218. In addition, one strong hydrogen bond is established between the backbone of Ser128 and the O2 oxygen of the fucose. For the second low energy conformation (TriA_conf2: $\Phi = 68°$, $\Psi = 77°$, $\Phi'' = -68°$, $\Psi'' = -90°$), the fucose residue has a different orientation and the hydrogen bond with Ser128 now involves the O4 oxygen of fucose. Hydrophobic contacts are rather more extended than in the other conformation since they involve the apolar face of the glycan (H1 and H2) which interact with methyl groups of Leu127, but also the methyl group of fucose which creates van der Walls contacts with Ser130 methine hydrogens.

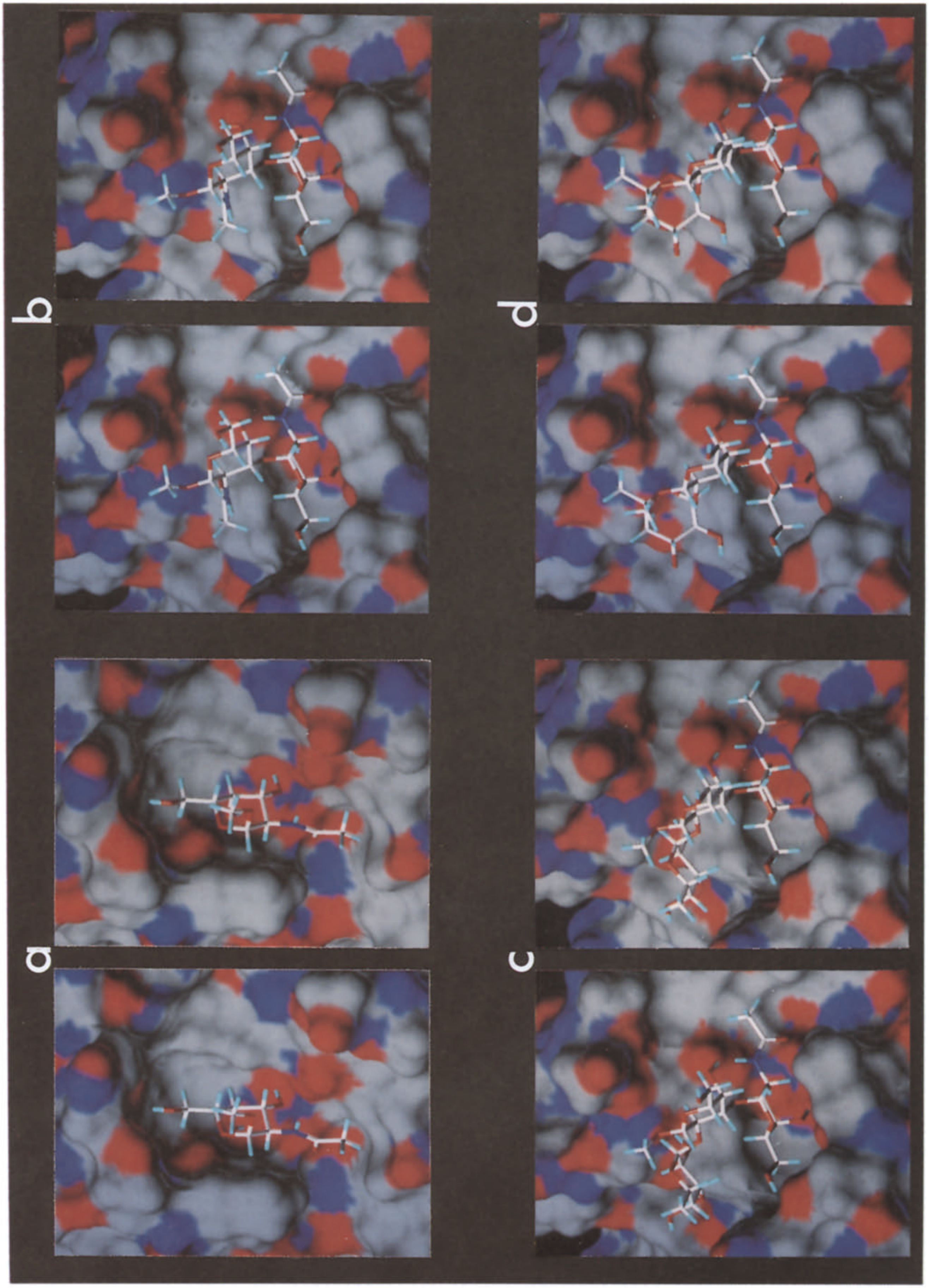


Plate 1. Stereorepresentation of the modelled binding site of DBL complexed with different carbohydrates. (a) GalNAc-*O*-Me; (b) Forssman disaccharide; (c) blood group A trisaccharide in the Conf1 conformation; and d/blood group A trisaccharide in the Conf2 conformation. The binding site is represented by its Connolly surface, with colour coding for hydrogen bonding potential: blue for hydrogen bond donors, red for hydrogen bond acceptors and grey for apolar atoms. The carbohydrate moiety is colour coded by atom type.

This second conformation, which is less populated than the first one in the isolated state [63], could be the one favoured in the interaction with DBL.

From the present study, it is proposed that the fucose on position 2 of Gal interacts with a secondary binding site, located in the vicinity of Ser128. On the contrary, position 1 of Gal points towards the solvent, and it seems that longer oligosaccharides would not create further contacting regions, at least for the first one or two residues. It would be of interest to know what are the affinities of DBL for larger oligosaccharides of blood goup A substance. Except for the data on the pentasaccharide αGalNAc(1-3)[α-L-Fuc(1-2)]βGal(1-4)βGlcNAc(1-6)R, where R is a hexenetetrol [1], there are no quantitative studies with A oligosaccharides. More data are available on the agglutination power of DBL with red cells of various blood groups. DBL is known to agglutinate A_1 better than A_2 blood group erythrocytes [6]. However; it has been shown that such a difference is due to a quantitative difference in the expression of the A antigen at the cell surface [65]. This has been recently confirmed by the fine structure of the glycosyltransferase genes, that show different activities of the A glycosyltransferases present in A_1 and A_2 blood group erythrocytes [66]. DBL has therefore been confirmed to be a general A type lectin, which is in agreement with the present calculations.

Conclusions

The present paper reports the first attempt to construct a three-dimensional model of a lectin from the sole knowledge of its primary structure as derived from its genomic sequence. This was accomplished despite a sequence identity of only 40% with the known crystal structures used in the knowledge-based building procedure. The quality of our present model was assessed using several stereochemical and biological criteria.

It is the first time that a GalNAc specific lectin is described and our model provides unique insights into the atomic bases of the interaction. As with other lectins, complementary forces emanate from hydrogen bonding, and van der Waals forces including hydrophobic interactions. In our model, the *N*-acetyl group stabilizes the complex through the formation of a hydrogen bond involving its carboxyl oxygen whereas the methyl group is involved in a hydrophobic cluster with two aromatic amino-acids.

We have also extended our docking investigations to di- and trisaccharides with the hope of understanding further the fine specificity of DBL. Satisfactory arrangement of these carbohydrates in the combining site of lectin could be proposed. They lead to the proposal of a secondary binding site, with hydrophobic character, which can accommodate, on the galactose moiety linked to the terminal GalNAc, either a fucose residue, or the *N*-acetyl group of a second GalNAc. However, fine specificity for higher oligosaccharides, or expected agglutinating activities towards red cells of different blood group A subgroups could not be inferred from the present study.

Legume lectins are now essential tools in the glycotechnology field, for analysis, separation or localization of complex glycans. We believe that the modelling technique exemplified on DBL can be used to engineer other lectins. This may be quite useful since most of these proteins are difficult to crystallise. Even when crystals are obtained, the three-dimensional model provides a useful tool for helping the structural elucidation. The availability of such three-dimensional models is also a great asset in the field of molecular biology. The joint use of computer-aided model building and molecular biology will lead to 'lectin engineering', a methodology which should allow the determination of new functions for such naturally abundant molecules.

Acknowledgements

The authors are much endebted to the following crystallographers, Yves Bourne, Louis Delbaere, Remy Loris, Karl Hardman and Boaz Shaanan who gave access to the coordinates of their crystal structures, either directly or by deposition in the Brookhaven database. Careful reading and helpful suggestions are acknowledged from Dr Rafael Oriol (INSERM U.178, Villejuif) and Dr Jean-Paul Mornon (CNRS, Université Paris VI). This work was partially supported by a National Institutes of Health Grant (M.E.E.).

References

1. Etzler ME, Kabat EA (1970) *Biochemistry* **9**:869–77.
2. Roberts DM, Etzler ME (1984) *Plant Physiol* **76**:879–84.
3. Etzler ME, Quinn JM, Schnell DJ, Spadoro JP (1986) In *Molecular Biology of Seed Storage Proteins and Lectins* (Shannon LM, Chrispeels MJ, eds) pp. 65–72. Rockville, MD: American Society of Plant Physiology.
4. Quinn JM, Etzler ME (1987) *Arch Biochem Biophys* **258**:535–44.
5. Etzler ME (1992) In *Glycoconjugates, Composition, Structure and Functions* (Allen HJ, Kisailus EC, eds) pp. 521–39. New York: Marcel Dekker Inc.
6. Bird GWG (1952) *Nature* **170**:674.
7. Boyd WC, Shapleigh E (1954) *J Immunol* **73**:226–31.
8. Carter WG, Etzler ME (1975) *Biochemistry* **14**:2685–9.
9. Quinn JM, Etzler ME (1989) *Plant Physiol* **91**:1282–6.
10. Etzler ME, Gupta S, Borrebaeck C (1981) *J Biol Chem* **256**:2367–70.
11. Schnell DJ, Etzler ME (1987) *J Biol Chem* **262**:7220–5.
12. Harada JJ, Spadoro-Tank J, Maxwell JC, Schnell DJ, Etzler ME (1990) *J Biol Chem* **265**:4997–5001.
13. Hammarström S, Murphy LA, Goldstein IJ, Etzler ME (1977) *Biochemistry* **16**:2750–5.
14. Baker DA, Sugii S, Kabat EA, Ratcliffe RM, Hermentin P, Lemieux RU (1983) *Biochemistry* **22**:2741–50.

15. Piller V, Piller F, Cartron J-P (1990) *Eur J Biochem* **191**:461–6.
16. Wu AM, Sugii S (1991) *Carbohydr Res* **213**:127–43.
17. Carter WG, Etzler ME (1975) *Biochemistry* **14**:5118–22.
18. Padlan EA, Kabat EA (1988) *Proc Natl Acad Sci USA* **85**:6885–9.
19. Nashed EM, Perdomo GR, Padlan EA, Kovac P, Matsuda T, Kabat EA, Glaudemans PJ (1990) *J Biol Chem* **265**:20699–707.
20. Oomen RP, Young NM, Bundle DR (1991) *Prot Eng* **4**:427–33.
21. Rao VSR, Reddy BVS, Mukhopadhyay C, Biswas M (1990) In *Computer Modeling of Carbohydrate Molecules* ACS Symposium Series Vol 340 (French AD, Brady JW, eds) pp. 361–76. Washington DC: American Chemical Society.
22. Imberty A, Hardman KD, Carver JPC, Pérez S (1991) *Glycobiology* **1**:631–42.
23. Reddy VS, Rao VSR (1992) *Int J Biol Macromol* **14**:185–92.
24. Imberty A, Pérez S (1994) *Glycobiology*, **4**:351–66.
25. Young NM, Oomen RP (1992) *J Mol Biol* **228**:924–34.
26. Imberty A, Bourne Y, Cambillau C, Rougé P, Pérez S (1993) *Adv Biophys Chem* **3**:61–117.
27. Berstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) *J Mol Biol* **112**:535–42.
28. Loris R, Steyaert J, Maes D, Lisgarten J, Pickersgill R, Wyns L (1993) *Biochemistry* **32**:8772–81.
29. Delbaere LTJ, Vandonselaar M, Prasad L, Quail JW, Wilson KS, Dauter Z (1993) *J Mol Biol* **230**:950–65.
30. Shaanan B, Lis H, Sharon N (1991) *Science* **254**:862–66.
31. Hardman KD, Agarwal RC, Freiser MJ (1982) *J Mol Biol* **157**:69–86.
32. Bourne Y, Roussel A, Frey M, Rougé P, Fontecilla-Camps J-C, Cambillau C (1990) *Proteins* **8**:365–76.
33. Gaboriaud C, Bissery V, Benchetrit T, Mornon J-P (1987) *FEBS Lett* **224**:149–55.
34. Lim VI (1974) *J Mol Biol* **88**:857–72.
35. HCA-PLOT V2, DORIANE Scientific Softwares, 44ter bvd Saint Antoine, 78150 Le Chesnay, France.
36. Blundell TL, Carney DP, Gardner S, Hayes FRF, Howlin TJP, Overington JP, Singh DA, Sibanda DL, Sutcliffe MJ (1988) *Eur J Biochem* **172**:513–20.
37. Sali A, Overington JP, Johnson MS, Blundell TL (1990) *TIBS* **15**:235–40.
38. Sutcliffe MJ, Haneef I, Carney D, Blundell TL (1987a) *Prot Eng* **1**:377–84.
39. Sutcliffe MJ, Hayes FRF, Blundell TL (1987b) *Prot Eng* **1**:385–92.
40. Jones TA, Thirup S (1986) *EMBO J* **5**:819–22.
41. Pérez S, Delage M-M (1991) *Carbohydr Res* **212**:253–9.
42. Allinger NL, Rahman M, Lii J-H (1990) *J Am Chem Soc* **112**:8293–307.
43. Allinger, NL, Yuh YH, Lii J-H (1989) *J Am Chem Soc* **111**:8551–66.
44. SYBYL V6.0, Tripos Associates, 1699 S. Hanley Road, Suite 303, St Louis MO 63144, USA.
45. White DNJ, Guy MHP (1975) *J Chem Soc Perkin Trans II* 43–6.
46. Clark M, Cramer RD III, van Opdenbosch N (1989) *J Comp Chem* **10**:982–1012.
47. Berthod H, Pullman A (1965) *J Chem Phys* **62**:942–6.
48. Powell MJD (1977) *Math Program* **12**:241–52.
49. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) *J Appl Cryst* **26**:283–91.
50. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) *Proteins* **12**:345–64.
51. Van Driessche E (1988) In *Advances in Lectin Research* (Franz H, ed.) pp. 73–134. Berlin: Springer-Verlag.
52. Sharon N, Lis H (1990) *FASEB J* **4**:3198–208.
53. Brünger AT (1988) In *Crystallographic Computing 4: Techniques and New Technologies* (Issacs NW, Taylor MR, eds) pp. 126–40. Oxford: Clarendon Press.
54. Agarwal RC (1978) *Acta Crystallogr* **A34**:791–809.
55. Haneef I, Moss DS, Stanford MJ, Borkakoti N (1985) *Acta Crystallogr* **A41**:426–33.
56. Hendrickson WA, Konnert JH (1980) In *Computing in Crystallography* (Diamond R, Ramaseshan D, Venkatesan K, eds) pp. 13.01–13.23. Bangalore: Indian Academy of Sciences.
57. Jeffrey GA, Saenger W (1991) In *Hydrogen bonding in Biological Structures* (Jeffrey GA, Saenger W, eds) pp. 136–146. Berlin: Springer-Verlag.
58. Vyas NK (1991) *Curr Opinion Struct Biol* **1**:732–40.
59. Connolly ML (1983) *Science* **221**:709–13.
60. Pérez S, Imberty A, Carver JP (1994) *Adv Comput Biol* **1**:147–202.
61. Bourne Y, Cambillau C (1993) In *Water and Biological Macromolecules* (Westhof E, ed.) *Topics in Molecular and Structural Biology*, vol. 17, pp. 321–37. Houndmills: Macmillan Press.
62. Carver JP (1993) *Pure Appl Chem* **65**:763–70.
63. Koca, J, Pérez S, Imberty A (1994) *J Comp Chem*, in press.
64. Bourne, Y, Rougé P, Cambillau C (1992) *J Biol Chem* **267**:197–203.
65. Watkins WM (1980) In *Advances in Human Genetics* (Harris H, Hirschhorn K, eds.) pp. 1–136, 379–85, New York: Plenum Publishing.
66. Yamamoto FI, McNeil PD, Hakomori FI (1992) *Biochem Biophys Res Com* **187**:366–74.