Glycoconjugate Journal 20, 545–550, 2004

Prediction of 3-D structures of fucose-binding proteins and structural analysis of their interaction with ligands

Sujata Majumder, Avik Roy and Chhabinath Mandal

Drug Design, Development and Molecular Modelling Division, Indian Institute of Chemical Biology Jadavpur, Kolkata-700 032, India

The importance of fucose-binding proteins stems from the presence of fucose as terminal sugars in H and Lewis (a) blood groups. Recently, the structure of a complex between *Anguilla anguilla* agglutinin (AAA) and α-L-fucose has been worked out at 1.9 A resolution. The structure of AAA characterizes the novel fold of an entire lectin family. In the present study, molecular modeling techniques have been used to identify new proteins that can provide a similar fucose binding module in the newly discovered genomic sequences using the above mentioned structural information. We modeled 3-D structures of three such proteins, namely, ebiP5322 protein of *Anopheles gambiae,* a pentraxin of *Xenopus laevis*, and the fw gene product of *Drosophila melanogaster*. α-L-fucose was docked in the binding pockets of the modeled structures followed by energy minimization and molecular dynamic runs to obtain the most probable structures of the complexes. Properties of these modeled complexes were studied to examine the nature of physicochemical forces involved in the complex formation and compared with AAA-α-L-fucose complex. It was found that ebiP5322 protein of *A. gambiae* and the pentraxin of *X. laevis* can provide a fucose-binding fold similar to AAA. We studied structures of four protein-fucose complexes to examine the electrostatic potential surfaces around the binding site and concluded that a highly positive-charged surface was not a necessary condition of fucose-binding.
Published in 2004.

Keywords: **fucose-binding proteins, molecular modeling, protein-sugar complex, lectin**

Abbreviations: **AAA: *Anguilla anguilla* agglutinin; BLAST: Basic local alignment search tool; CDR: complementarity determining regions; CRD: carbohydrate recognition domain; AA: amino acid.**

Introduction

Lectins, a class of sugar-binding proteins, play crucial roles in innate immunity and host defense not only in vertebrates, but also in invertebrates, with involvement in processes such as non-self recognition, inflammation, opsonization, cell-cell or cell-extracellular matrix interactions, fertilization, development, and regeneration [1–3]. Fucose-binding lectins (fucolectins) from eel serum have a long history of research and practical applications [4–8] as they recognize fucosylated terminals of H antigens and Lewis (a) blood groups. cDNA cloning and characterization reveal that the eel fucolectins are a group of proteins with at least seven different isoforms. It is believed that this heterogeneity or multiplicity may have evolved to provide a means for effective defense against a wide variety of pathogenic bacteria and appears to be consistent with their primary functions in innate immunity [9].

Recently, the crystal structure of the complex of *Anguilla anguilla* agglutinin (AAA) with alpha-L-fucose has been solved by x-ray crystallography at 1.9 Å resolution [10]. They have shown that the AAA folds as a β-barrel with jellyroll topology, the bulk of the fold consisting of eight major antiparallel β-strands arranged in two β-sheets of five and three strands packed against each other. Two short antiparallel strands close one end of the barrel. Five loops connect the two main β-sheets at the other end of the barrel. These loops were named as complementarity determining regions (CDRs) 1 to 5 by analogy to the immunoglobulin nomenclature. These CDR loops encircle a highly positively charged hollow that provides the complementary surface for the fucose molecule. This has been identified as a novel fold representing this entire lectin family

To whom correspondence should be addressed: Chhabinath Mandal, Drug Design, Development and Molecular Modelling Division, Indian Institute of Chemical Biology, 4, Raja S.C. Mullick Road Jadavpur, Kolkata-700 032, India. Tel.: +91 33 2473 6793; Fax: +91 33 2473 5197; E-mail: cnmandal@iicb.res.in

[10]. In the present study we searched for this novel fold in other newly discovered homologous sequences. Using BLAST search we identified three sequences, namely, from *Anopheles gambiae* (ACCESSION No: EAA04956) and from *Drosophila melanogaster* (ACCESSION No: NP_511136) which are potential candidates for providing such lectin folds. 3-D structures of these proteins have been predicted by homology modeling and the fucose molecule has been docked to arrive at the structures of the complexes using computer-aided modeling techniques. Analysis of these structures allowed us to examine the feasibility of the formation of such folds and the nature of interactions of fucose with these proteins. We also examined the binding of some other conformations of fucose in the complementary surface of the *Anopheles gambiae* protein to unfold the nature of their interactions. Another striking observation was the highly positively charged environment of the fucose-binding pocket [10]. In order to verify if this is a necessary condition of fucose binding, we calculated the electrostatic potentials of some other fucose-binding proteins and examined their nature in the vicinity of the binding pocket.

Methods

Initial structures of the fucose-binding proteins were modeled by knowledge-based homology modeling using our in-house software package of ANALYN and MODELYN [11]. The starting scaffold for homology modeling was the x-ray crystallographically determined structure of AAA (PDB ID: 1 K12). These structures were refined using the InsightII 98.0 of MSI (Biosym Technologies, San Diego, CA) equipped with DISCOVER as the energy minimization and molecular dynamics module. Structural optimization involved energy minimization (100 steps each of steepest descent and conjugate gradient methods) using cff91 force field followed by dynamics simulations. A typical dynamics run consisted of 1000 steps of one femtosecond each after 100 steps of equilibration with a conformational sampling of 1 in 10 steps at 300°K. At the end of the dynamics simulation, the conformation with lowest potential energy was picked for the next cycle of refinement. This combination of minimization and dynamics was repeated until satisfactory conformational parameters were obtained. Special care was taken in the structural zones where major insertion or deletions were made. Structures of the complexes were obtained by the superposition of the modeled protein structures with the experimental structure of a fucose complexed with AAA (1 K12) followed by optimization with repeated energy minimization and dynamics simulations. Position constraints were applied to the atoms which were more than 8 Å away during energy minimization and molecular dynamics of the complexes.

ANALYN was used for the homology analysis of pre-aligned sequences of the target and scaffold proteins and was run on an IBM compatible PC. MODELYN was used for automated prediction of the target structure and their structural analysis after refinements; it was run on both on an IBM compatible PC in the Windows environment and OCTANE workstation of Silicon Graphics, Inc. in the UNIX environment. InsightII was run on OCTANE workstation of Silicon Graphics, Inc. in the UNIX environment. CLUSTALW [12] was run through the Internet for multiple alignment of the amino acid sequences. The electrostatic potential surfaces of the proteins were determined by MOLMOL [13]. PROCHECK [14] was used for checking the structural parameters. Both MOLMOL and PROCHECK were run on OCTANE in the UNIX operating system. Hydrogen bonding patterns of the modeled and x-ray structures were obtained by adding hydrogen (x-ray structures lacked hydrogen atoms, and DISCOVER needs these atoms for minimization and dynamics) followed by optimization of the complex by energy minimization and molecular dynamics. Structural superposition was carried out using ABGEN [15] on both PC and UNIX environments. Protein BLAST [16] was used through Internet for finding homologous sequences.

Results and discussion

In search of proteins homologous to the *Anguilla anguilla* agglutinin (AAA) BLAST search was performed with its amino acid (AA) sequence as the query. Initial hits with high homology were the isoforms of this protein found in *Anguilla japonica* as expected [4]. Barring those isoforms, three proteins showed reasonable homology. First in this category was a pentraxin of *Xenopus laevis* (with an expected value of 8×10^{-26} and 56% positive score), a hypothetical protein (ebiG5322 gene product) from the newly worked out *Anopheles gambiae* genome (Expect $= 2 \times 10^{-14}$ and positive score $= 50\%$), and the fw gene product of *Drosophila melanogaster* (Expect $= 2 \times 10^{-13}$ and positive score $= 52\%$). AA sequences of these three proteins were used as target sequences for 3-D structure prediction and analysis for the novel lectin fold described recently [10]. After the initial structures were predicted by the combination of ANALYN and MODELYN, they were refined by regularization of segments with major deletion or insertion using DISCOVER module of InsightII for energy minimization and molecular dynamics. Final predicted structures were checked for main chain conformations using PROCHECK, which showed that more than 85% of the ϕ–ψ plots were in the core regions and less than 3% were in the disallowed regions of Ramachandran's plots. Root mean square deviations of bond lengths and bond angles were within 0.03 Å and 3.5 degrees from the standard values indicating reasonably good structural parameters of the predicted structures.

Superposed structures of the predicted model of the ebiG5322 gene product of *Anopheles gambiae* (yellow ribbon) and the x-ray structure starting scaffold (AAA, red ribbon) are shown in Figure 1 along with the position of the bound fucose. It may be noted that the structure superposed very well in most of the regions except for two regions marked by blue arrows. The deviation of the segment at the top was due to deletion of 4 AA in one of the complementarity determining region (CDR2) which is involved in sugar binding. Multiple alignment of all

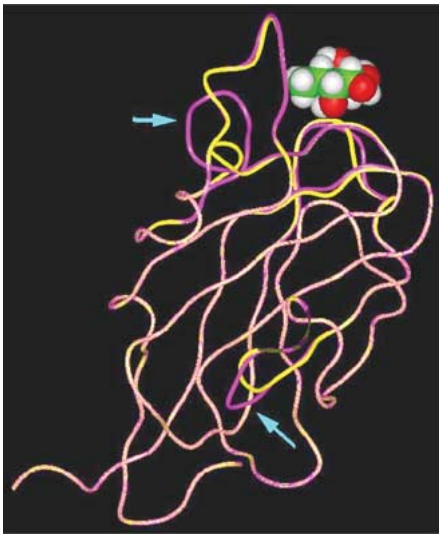

Figure 1. Ribbon representations of superposed structures of Anguilla anguilla Agglutinin (AAA, x-ray, pink) and Anopheles gambiae ebiG5322 gene product (AG, model, yellow). Blue arrows show the regions of the structures with remarkable deviations. One loop with a major deviation is due to deletion of four amino acids in CDR2 loop of AG (upper arrow). Position of bound α-L-fucose is shown by space-filling representation.

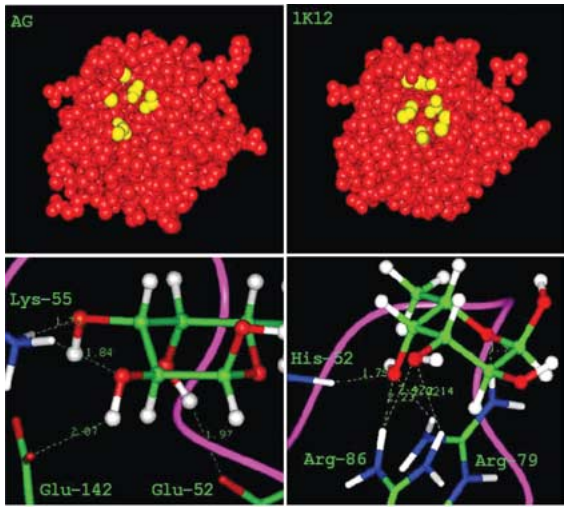

Figure 3. Space-filling representations of AG (Top left panel) and that of AAA (top right panel). Yellow spheres represent the atoms of residues within 4.00 Å around α-L fucose. Three residues of AG involved in hydrogen binding with α-L Fucose (bottom left panel) are the following: Glu-52 (OE1) —O4(H); Lys-55 —O3, O2; Glu-142 (OE2) —O2(H). The binding triad of AAA (bottom right panel) responsible for fucose binding is in the following hydrogen bonding pattern: His (52) —O4; Arg (86) —O3, O2; Arg (79) —O4, O5.

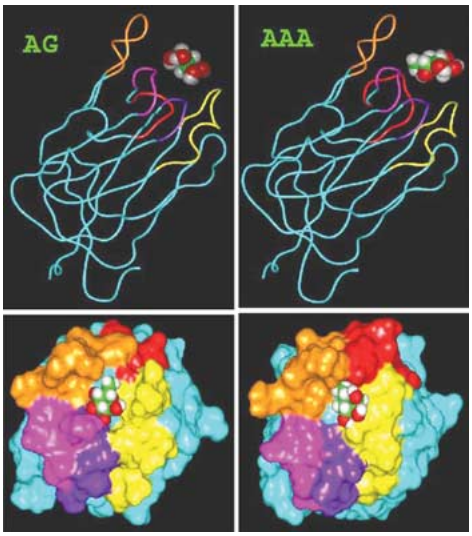

Figure 2. Ribbon representation of the structure of AG (top left panel) and AAA (top right panel). Five CDR loops that constitute the carbohydrate-binding surface as shown in different colors: mustard—CDR1, red-CDR2, pink-CDR3, yellow-CDR4, and violet-CDR5. Fucose is shown by CPK representation in both the panels. Connolly surfaces of AG (bottom left panel) and that of AAA (bottom right panel) around α-L Fucose are shown with the CDRs in the same colors. These CDRs encircle the bound α-L fucose that is shown by space-filling representation.

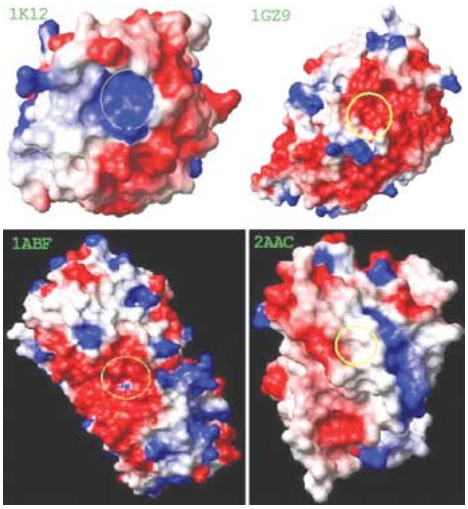

Figure 4. Electro static potential surfaces of a few fucose binding proteins. Blue region represents positively charged environment, red for negatively charged and white for hydrophobic surroundings. Top left panel: AAA (1 K12, totally positive). Top right panel: *Erythrina cristagali* fucolectin (1GZ9, mostly negative with a patch hydrophobic region). Bottom left panel: L-arabinose binding proteins of *Escherechia coli* (1ABF, mostly negative with a patch of positively charged surface). Bottom right panel: *Escherechia coli* gene regulatory protein AraC (2AAC, mostly hydrophobic and a patch negatively charged region).

three target sequences with the Anguilla anguilla agglutinin shows that the CDR2 loop is 4 AA shorter in all these proteins compared to the starting scaffold.

The complementarity determining regions of the modeled ebiP5322 protein (AG) and those of the x-ray structure (AAA) with reference to the bound fucose molecule (shown as space filling model) is shown in Figure 2. The top panels show the CDR loops in the ribbons representation and the bottom panels show those as the Connolly surface drawn in different colors. It is noteworthy that the fucose-binding site is surrounded by the CDR loops. Variation in the AA sequence or composition gives rise to different complementary surfaces leading to differences in the sugar specificity. Five CDRs for the ebiP5322 protein are found to span as follows—CDR1: Gln-22 to Ser-31 (mustard), CDR2: Arg-41 to Ser-50 (red), CDR3: Ser-53 to Asn-58 (pink), CDR4: Arg-79 to Ala-90 (yellow) and Pro-138 to Ser-142 (violet).

The contact points of the fucose molecule in the sugar-binding pockets of the modeled ebiP5322 protein (AG, left) and the x-ray structure of AAA (1K12, right) is showm in Figure 3 (upper panel). Atoms of the proteins in contact with the sugar are shown in yellow color. It may be pointed out that the patterns of fucose-binding regions are very similar in these two proteins. The bottom panel (Figure 3) shows the network of hydrogen bonding of the bound sugar with the protein molecules, which provide the binding strength between these molecules. X-ray crystallographic studies on the complex between AAA and fucose identified the binding triads of the agglutinin in analogy with the catalytic triads. The triad consists of His-52 (forming hydrogen bonding with O4 of fucose), Arg-86 (H-bonded to O2 and O3 of fucose) and Arg-79 (H-bonded to O4 and O5 of fucose) as shown in the right bottom panel of Figure 3. We have also found the binding triad of ebiP5322 protein as shown in the left bottom panel of Figure 3. In this case the sugar-binding triad comprises of Glu-52 (bound to H atom of O4 of fucose), Lys-55 (bound to O2 and O3 of fucose), and Glu-142 (bound to H atom of O2 of fucose). It may be pointed out here that in case of *Anguilla anguilla* agglutinin all the amino acids acting as the sugar-binding triad are basic in nature (*e.g.* His-52, Arg-86, and Arg-79). On the other hand the predicted binding triad of ebiP5322 protein contains one basic AA (Lys-55) and two acidic AA (Glu-52 and Glu-142). We also searched for the sugar-binding triad in the other complexes modeled by us; we found the complex between the pentraxin of *Xenopus laevis* and fucose also exhibit a binding triad containing the same set of AA at identical locations (*e.g.* His-52, Arg-79 and Arg-86) as shown in Table 1. In case of the other modeled protein from the fw gene of *Drosophila melanogaster,* we did not find any catalytic triad. It is important to note that, in case of the binding triad of ebiP5322 protein, not only the AA are of different type but also the locations of the AA are not the same as the other to triads.

Table 1. Emperical binding energies of the complexes of some proteins with fucose along with their hydrogen bonding patterns. Atoms on the right side of the residue IDs followed by the '—' sign are the atoms of the sugar molecule involved in hydrogen bonding

	Binding Energies in Kcals/mol			
Structure	*VdW**	*Elec.**	*Total*	*Residues involved in H-bonding*
1. *Anguilla anguilla* agglutinin (complexed with α-L-Fucose)	−12.91	−26.90	−39.82	His-52 —O4; Arg-79—O4, O5; Arg-86 —O3, O2;
2. *Anopheles gambiae* ebiP5322 (complexed with α-L-Fucose)	−9.13	−48.20	−57.33	Glu-52—O4 (H); Lys-55 —O3, O2; Glu-142—O2(H);
3. *Xenopus laevis* pentraxin (complexed with α-L-Fucose)	−9.53	−32.61	−42.14	His-52 —O4, O4(H); Arg-79 —O4, O5; Arg-86 —O3, O2;
4. *Drosophila melanogaster* Fw (complexed with α-L-Fucose)	−14.53	−20.26	−34.79	Glu-48 —O4;
5. *Escherichia coli* L-arabinose binding protein (complexed with β-D-Fucose, 1APB)	−15.55	−88.99	−104.5	Lys-10—O1, O2; Glu-14 (OE1)—O3(H) Glu-14 (OE2)—O3(H); Arg-151— O4, O5; Asn-205—O3; Asn-232—O4(H), O3;
6. Sugar-binding and dimerization domain of the *Escherichia coli* gene regulatory protein, AraC, (complexed with β-D-Fucose, 2AAC)	−23.40	−57.06	−80.47	Leu-9—O1(H), O2(H); Thr-24—O2; Arg-38—O5, O1; Tyr-82—O4(H); His-93—O3;

*VdW = van der Waals & Elec. = electrical energies.

The binding energies of some complexes of fucose with some proteins along with the summary of their hydrogen bonding patterns are given in Table-1. It contains empirical energies, both van der Waals and electrical components, of the complex formation and the pattern of hydrogen bonding between the protein and the sugar. Item numbers 2 to 4 are for the structures of complexes of our modeled proteins with α-L-fucose, and item number 1 is for the x-ray structure of the starting scaffold (1K12) with α-L-fucose. The sugar-binding catalytic triads could be identified in these complexes except for the fw gene product of *Drosophila melanogaster*. In this case there is only one hydrogen bond between the protein and the sugar. In terms of binding energy, this complex shows the least negative energy (-34.79 kcals/mole) indicating a most unstable complex formation. The most stable complex (-57.33 kcals/mole) in this series is formed between ebiG5322 gene product of *Anopheles gambiae* and α-L-fucose. The binding strength of the other two complexes fall in between.

Items number 5 and 6 of Table 1 are the binding energies and hydrogen bonding patterns for the x-ray crystallographically determined structures of proteins (1 APB and 2 AAC) with β-D-fucose. In these complexes the binding energies are much higher than the α-L-fucose complexes. The contributions of the van der Waals interactions are very close for all the complexes ranging from -9 to -23 kcals/mole while the electrical contributions show much variation ranging from -20 to -89 kcals/mole. The electrical energies are composed of hydrogen bonding as well as dipole-dipole interactions. In the complexes with β-D-fucose higher numbers of hydrogen bonds are formed contributing more to the electrostatic interactions. However, these complexes do not exhibit any binding triad, rather, a number of AA are involved in hydrogen bonding. Thus, the involvement of the sugar-binding triad is confined in the newly found lectin fold of the AAA including all its isoforms [10]. It may be assumed that these sequences fold in a manner similar to that of AAA with a 'fucolectin-like' carbohydrate-recognition domain (CRD), while other fucose binding proteins fold in different ways to provide different types of CRD.

Another interesting finding about the AAA fold was the distribution of a highly positive potential around the fucose binding cleft [10]. Whether it is a necessary condition was examined by calculating the electrostatic potentials of four fucose bound complexes whose structures were determined by x-ray crystallography (Figure 4) e.g., complexes of fucose with *Anguilla anguilla* agglutinin [10], *Erythrina cristagali* fucolectin [17], L-arabinose binding protein of *Escherechia coli* [18,19], and *Escherechia coli* gene regulatory protein AraC [20]. The electrostatic potential surface of *A. anguilla* agglutinin (top left, 1K12) contains a deep blue patch of a highly positive potential around the fucose binding cleft (marked by yellow ring) while that around the binding site of *E. cristagali* fucolectin (top right, 1GZ9) is mostly negatively charged (red) with a small white patch indicating a neutral region at the bottom of the cleft. The electrostatic potential of the sugar-binding site of L-arabinose-binding protein of *E. coli* (bottom left, 1ABF) is mostly negative with a small positively charged spot at the middle. The sugar binding cleft of *E. coli* gene regulatory protein AraC (bottom right, 2AAC) is characterized by a highly neutral (hydrophobic) zone with a slight pinch of red indicating a slightly negatively charged surface at the border. Thus, it is evident that the fucose (or some other sugar as well) binding sites can have very diverse types of electrostatic potential around the binding site cleft. As all these sites are for neutral sugars, there is no need for any overall charge around them.

Conclusions

We have modeled the structures of ebiG5322 gene product of *Anopheles gambiae*, fw gene product of *Drosophila melanogaste, r* and a pentraxin of *Xenopus laevis* based on the experimental structure of AAA, and fucose was docked into the sugar-binding sites of these structures to study the physicochemical forces involved in such binding. We have demonstrated that the fucose-binding triad (His-52, Arg-79, Arg-86) of AAA is conserved in the pentraxin of *Xenopus laevis* but a different set of triad (Glu-52, Lys-55, and Glu-142) has been found in ebiP5322 protein of *Anopheles gambiae*, which maintains a similar type of hydrogen bonding network with Fucose. Calculation of empirical energies of binding of the complexes of α-L-Fucose with these proteins indicates that the ebiP5322 protein of *Anopheles gambiae* would bind most strongly (-57.33 Kcals/mol) followed by pentraxin of *Xenopus laevis* (-42.14 Kcals/mol) and AAA (-39.82 Kcals/mol) (Table 1). The least binding energy -34.79 Kcals/mol) was found with the fw gene product of *Drosophila melanogaster* and only one hydrogen bond formed between the protein and fucose. Totally positive electrostatic potential environment of the fucose-binding site as found in AAA is not essential for this binding as that pattern is not observed in other known structures.

Hence, from our model building, we may conclude that the hypothetical protein obtained from the conceptual translation of the ebiG5322 gene of A*nopheles gambiae* is a fucose binding protein with a different set of triad and the pentraxin of *Xenopus laevis* provide the same set of triad as that of the novel fold of *Anguilla anguilla* agglutinin.

Acknowledgment

This work was financially supported by the Project No CMM0017 of CSIR, India.

References

1 Faye I, Acquired immunity in insects: The recognition of non-self and the subsequent onset of immune protein genes, *Res Immunol* **141**, 927–32 (1990).

2 Kubo T, Kawasaki K, Natori S, Cloning of cDNA for regenectin, a humoral C-type lectin of Periplaneta americana, and expression

of the regenectin gene during leg regeneration, *Insect Biochem* **20**, 585–91 (1990).
3 Saito T, Hatada M, Iwanaga S, Kawabata S, A newly identified horseshoe crab lectin with binding specificity to O-antigen of bacterial lipopolysaccharides, *J. Biol Chem* **272**, 30703–08 (1997).
4 Honda S, Kashiwagi M, Miyamoto K, Takei Y, Hirose S, Multiplicity, structures, and endocrine and exocrine natures of eel fucose-binding lectins, *J. Biol Chem* **275**, 33151–7 (2000).
5 Bezkorovainy A, Springer GF, Desai PR, Physicochemical properties of the eel anti-human blood-group H(O) antibody, *Biochemistry* **10**, 3761–4 (1971).
6 Matsumoto I, Osawa T, Specific purification of eel serum and *Cytisus sessilifolius* anti-H hemagglutinins by affinity chromatography and their binding to human erythrocytes, *Biochemistry* **13**, 582–8 (1974).
7 Kelly C, Physicochemical properties and N-terminal sequence of eel lectin, *Biochem J* **220**, 221–6 (1984).
8 Baldus SE, Thiele J, Park YO, Hanisch FG, Bara J, Fischer R, Characterization of the binding specificity of Anguilla anguilla agglutinin (AAA) in comparison to Ulex europaeus agglutinin I (UEA-I), *Glycoconj J* **13**, 585–90 (1996).
9 Hoffmann JA, Kafatos FC, Janeway CA, Ezekowitz RA, Phylogenetic perspectives in innate immunity, *Science* **284**, 1313–8 (1999).
10 Bianchet MA, Odom EW, Vasta GR, Amzel LM, A novel fucose recognition fold involved in innate immunity, *Nat Struct Biol* **9**, 628–34 (2002).
11 Mandal C, MODELYN–A molecular modelling program version PC-1.0 Indian Copyright No 9/98. (1998).
12 Thompson JD, Higgins DG, Gibson TJ, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**, 4673–80 (1994).
13 Koradi R, Billeter M, Wuthrich K, Related Articles, MOLMOL: A program for display and analysis of macromolecular structures. *J Mol Graph* **14**, 51–5 (1996).
14 Laskowski RA, MacArthur MW, Moss DS, Thornton JM, PROCHECK: A program to check the stereochemical quality of protein structures, *J Appl Cryst* **26**, 283–91 (1993).
15 Mandal C, Kingery BD, Anchin JM, Subramanium S, Linthicum DS, ABGEN: A knowledge based automated approach for antibody structure modeling, *Nature Biotechnol* **14**, 323–8 (1996).
16 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res* **25**, 3389–02 (1997).
17 Svensson C, Teneberg S, Nilsson C, Kjellberg A, Schwarz F, Sharon N, Krengel U, High-resolution crystal structures of *Erythrina cristagalli* lectin in complex with lactose and 2'-alpha-l-fucosyllactose and correlation with thermodynamic binding data, *J Mol Biol* **321**, 69–78 (2002).
18 Quiocho FA, Vyas NK, Novel stereospecificity of the L-arabinose-binding protein, *Nature* **310**, 381–6 (1984).
19 Quiocho FA, Wilson DK, Vyas NK, Substrate specificity and affinity of a protein modulated by bound water molecules, *Nature* **340**, 404–07 (1989).
20 Soisson SM, MacDougall-Shackleton B, Schleif R, Wolberger C, The 1.6 A crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose, *J Mol Biol* **273**, 226–37 (1997).