Glycan Recognition

Nature-inspired engineering of an F-type lectin for increased binding strength

Sonal Mahajan and T N C Ramya¹

Institute of Microbial Technology, Sector 39-A, Chandigarh 160036, India

¹To whom correspondence should be addressed: Tel: +91-172-6665243; e-mail: ramya@imtech.res.in

Received 14 December 2017; Revised 5 September 2018; Editorial decision 5 September 2018; Accepted 7 September 2018

Abstract

Individual lectin–carbohydrate interactions are usually of low affinity. However, high avidity is frequently attained by the multivalent presentation of glycans on biological surfaces coupled with the occurrence of high order lectin oligomers or tandem repeats of lectin domains in the polypeptide. F-type lectins are L-fucose binding lectins with a typical sequence motif, HX(26)RXDX(4)R/K, whose residues participate in L-fucose binding. We previously reported the presence of a few eukaryotic F-type lectin domains with partial sequence duplication that results in the presence of two L-fucose-binding sequence motifs. We hypothesized that such partial sequence duplication would result in greater avidity of lectin–ligand interactions. Inspired by this example from Nature, we attempted to engineer a bacterial F-type lectin domain from *Streptosporangium roseum* to attain avid binding by mimicking partial duplication. The engineered lectin demonstrated 12-fold greater binding strength than the wild-type lectin to multivalent fucosylated glycoconjugates. However, the affinity to the monosaccharide L-fucose in solution was similar and partial sequence duplication did not result in an additional functional L-fucose binding site. We also cloned, expressed and purified a *Branchiostoma floridae* F-type lectin domain with naturally occurring partial sequence duplication and confirmed that the duplicated region with the F-type lectin sequence motif did not participate in L-fucose binding. We found that the greater binding strength of the engineered lectin from *S. roseum* was instead due to increased oligomerization. We believe that this Nature-inspired strategy might be useful for engineering lectins to improve binding strength in various applications.

Key words: F-type lectin domain, multivalency, oligomerization, partial duplication, tandem repeat

Introduction

Lectins are proteins that specifically recognize and bind to various carbohydrate moieties through their carbohydrate recognition domains (CRDs) (Kilpatrick 2002). Lectins recognize carbohydrates on pathogen surfaces and carry out effector functions such as immobilization, agglutination and opsonization of pathogens (Kilpatrick 2002; Sharon and Lis 2004; Vasta et al. 2004, 2007, 2017). Lectins frequently occur either as oligomers of non-covalently bound subunits or as tandem repeats within a single polypeptide. Oligomerization and/or the presence of tandem repeats enable lectins to achieve multivalent binding that results in high avidity towards the target ligand (Lis and Sharon

1998; Monsigny et al. 2000). For instance, the mannan-binding lectin polypeptide has a single CRD but individual polypeptides associate to form a bouquet-like oligomer (Drickamer and Taylor 1993; Jensen et al. 2005); similarly, the conglutinin polypeptide has a single CRD and it oligomerizes to form a multimeric “cruciform” structure that can crosslink multiple targets (Holmskov et al. 2003; Gupta and Surolia 2012). Tandem repeats of CRDs within a single polypeptide are reported in C-type immunoglobulins, in the R-type lectin, ricin, and in galectins, which have two tandem domains of different binding properties, and they serve to increase avidity and/or crosslink the ligand (Drickamer, Vasta and Ahmed 2009).

F-type lectins are L-fucose-binding lectins with a ~ 140 residue long domain with two conserved sequence motifs – a calcium-binding motif and an L-fucose binding motif, HX(26)RXDX(4)R/K, where X is any residue; the conserved basic residues, H, R and R/K of the FLD sequence motif are involved in L-fucose recognition (Vasta et al. 2008, 2017; Vasta and Ahmed 2009). The “F-type lectin fold” was first described in *Anguilla anguilla* agglutinin (AAA) as a β -barrel with jelly roll topology (Bianchet et al. 2002). The β -barrel comprises eight β -strands (Bianchet et al. 2002). On one end of the β -barrel is a shallow ligand binding pocket formed by the loops (or complementarity determining regions (CDRs)) that connect the β -strands (Bianchet et al. 2002). The ligand binding pocket contains the conserved basic residues of the FLD sequence motif, His, Arg and Arg/Lys that make hydrogen bonds with the axial 4-OH group of α -L-fucose (Bianchet et al. 2002). Mutagenesis studies have indicated that these basic residues are critically required for L-fucose recognition (Farrand et al. 2008) (Bishnoi et al. 2018). Besides hydrogen bonding with these residues, binding to L-fucose is aided by Van der Waals contacts with hydrophobic residues in the ligand-binding pocket (Bianchet et al. 2002). The structures of FLDs from *Morone saxatilis*, *Streptococcus mitis* and *Streptococcus pneumoniae* are also available (Boraston et al. 2006; Bianchet et al. 2010; Feil et al. 2012).

F-type lectins are widely distributed in organisms ranging from bacteria to vertebrates with diverse domain architectures (Vasta et al. 2004, 2017; Bishnoi et al. 2015). Eukaryotic F-type lectin domains (FLDs) usually exist singly or as one to many tandem repeats, but are sometimes also found co-occurring with other diverse domains (Bishnoi et al. 2015). Tandem FLDs are also present in mosaic combinations in some polypeptides (Vasta et al. 2004; Bishnoi et al. 2015). For example, a binary tandem domain F-type lectin from *M. saxatilis* has two CRDs of distinct specificities for fucosylated ligands, F-type lectins are composed of either triplicate or quintuplicate tandemly arrayed FLDs in *Xenopus* spp., a single FLD is found in association with a C-type lectin domain and Sushi repeats in the furrowed proteins of *Drosophila melanogaster* CG9095, and multiple tandem repeats of FLD are present in association with pentraxin-1 domain in *Xenopus laevis* (Vasta et al. 2004; Odom and Vasta 2006; Bianchet et al. 2010; Bishnoi et al. 2015). Typically, bacterial FLDs exist as a single copy and are co-associated with one or more different domains such as carbohydrate binding domains (such as CBM6 cellulase and CBM6 xylanase), carbohydrate-active enzyme domains (such as α -L-fucosidase, glycosyltransferase family domains, alginate lyase and β -N-acetylglucosaminidase), other enzyme domains (such as lipase, thiol cytolysin and methyltransferase) or various other domains (Bishnoi et al. 2015). An exception is the protein SP2159 from *S. pneumoniae*, which has three tandem FLDs associated with a glycoside hydrolase; the tandem presentation of the FLDs enable improved binding to multivalent ligands in this case (Boraston et al. 2006).

We previously identified and reported six eukaryotic FLDs (from *Branchiostoma floridae*, *Latimeria chalumnae* and *Chrysemys picta bellii*), each with a tandemly displayed partial FLD repeat, i.e., a partial repeat of its own sequence likely formed due to internal duplication (Bishnoi et al. 2015) (Supplementary data, Table SI). The region of the domain that is duplicated in these FLDs is the sequence between the β 1 and β 6 strands that includes the evolutionarily conserved FLD sequence motif (Figure 1A). These proteins, therefore, can be considered as possessing a partial FLD followed by a complete FLD, and hence two FLD sequence motifs. We found that in five of these six proteins with partial FLD duplication, the His or Arg residues of one of the two FLD sequence motifs critically

required for L-fucose binding were naturally substituted with apolar residues. In two proteins, the His or Arg residues of the FLD sequence motif in the partial FLD region were naturally substituted with apolar residues, while in three other proteins the FLD sequence motif of the partial FLD region was conserved but the His and/or Arg residues of the FLD sequence motif of the subsequent complete FLD were mutated and the His was naturally substituted with apolar amino acids. This suggested that the loss of these critical residues in one of the two FLD sequence motifs might be offset by redundancy and therefore not be deleterious. We considered that the fold adopted by the FLD is a β -barrel, that β -barrels can theoretically be expanded by the addition of more β -strands (Arnold et al. 2007), and that the partial FLD region has the FLD sequence motif in the same context of secondary structural elements (Figure 1A, B). We theorized that a protein with partial FLD duplication possessing conserved FLD sequence motifs in both the partial and complete FLD regions might display an additional functional L-fucose binding pocket. We hypothesized that the presence of two functional L-fucose binding motifs enabled by partial FLD duplication might result in avid binding to multivalent ligands.

Inspired by this example of domain expansion in Nature, we engineered a bacterial FLD (*Streptosporangium roseum* FLD; hereafter referred to as SrFLD) to contain a partial duplication of its own FLD sequence and studied the effect of the partial duplication on ligand binding. We demonstrate that the engineered lectin (hereafter referred to as SrDupFLD) has significantly improved binding strength towards multivalent fucosylated glycoconjugates and that, interestingly, this increased binding strength is due to increased oligomerization and not due to the presence of an additional functional L-fucose-binding site. We also confirm that the additional FLD sequence motif in a recombinant *B. floridae* protein with naturally occurring partial FLD duplication (B/DupFLD) does not contribute to L-fucose binding.

Results

Purification of wild type and engineered *S. roseum* FLDs

We engineered SrFLD to contain a partial duplication of its own FLD (including the CDR1, CDR2, CDR3 and CDR4 loops and the FLD sequence motif with the conserved residues implicated in L-fucose binding) at its N-terminus, mimicking the domain organization of B/DupFLD (Figure 1A). We successfully expressed and enriched SrFLD and the engineered protein, SrDupFLD to 90% purity by metal ion affinity chromatography (Figure 2A). We also successfully expressed and purified SrDup_{TM}FLD, in which the L-fucose binding residues of the N-terminal partial FLD region are mutated, and SrDupFLD_{TM}, in which the L-fucose binding residues of the complete FLD region are mutated. We performed CD spectroscopy to examine if the addition of the duplicated elements in SrDupFLD resulted in any alteration in the secondary structure profile. Our data indicated that the secondary structure of the engineered FLD (SrDupFLD) with the duplicated insert is similar to that of the wild-type FLD (Figure 2B).

Comparison of saccharide binding by wild-type and engineered lectin

Enzyme-linked lectin assay (ELLA)

We determined that SrFLD and SrDupFLD recognize L-fucose using an enzyme-linked lectin assay (ELLA) in which the lectins were

Fig. 1. F-type lectins with partial duplication. **(A)** Alignment of *S. roseum* FLD (SrFLD) with *Anguilla anguilla* agglutinin (1K12), *B. floridae* FLD (BfFLD), *B. floridae* FLD with partial duplication (BfDupFLD), and engineered *S. roseum* FLD with partial duplication (SrDupFLD). The yellow and the green arrows below the alignment represent the β -strands, and the green cylinders represent the α -helices of the typical FLDs. The blue and the purple arrows are the β -strands, and the brown cylinders are the α -helices of the partially duplicated FLD region in BfDupFLD and SrDupFLD. The asterisks indicate the residues critically involved in binding to L-fucose. The short purple downward arrows mark the first and the last residues of the region chosen for partial duplication in engineered SrDupFLD. **(B)** Topology models providing secondary structure element information and connectivity for the FLD fold and a putative DupFLD (a partial FLD followed by a complete FLD) fold assuming barrel expansion. The topology models are based on the structure of 1k12. The coloring scheme is the same as described above for the alignment.

immobilized and allowed to bind to polyacrylamide-based glycoconjugate probes provided in the solution (Figure 2C). We found that immobilized SrFLD and SrDupFLD bound to biotin-PAA- α -L-fucose and not to PAA- α -D-galactose provided in the solution; we used biotinylated PAA probe as a negative control (Figure 2C). We observed that binding of SrFLD to biotin-PAA- α -L-fucose was almost completely abolished by 300 mM free L-fucose in solution (Figure 2C). However, 300 mM free L-fucose competitively inhibited binding of SrDupFLD to biotin-PAA- α -L-fucose to a significantly less extent (Figure 2C) ($P = 0.03$, one-tailed paired Student's t -test), suggesting that the engineered lectin, SrDupFLD has increased binding strength for biotin-PAA- α -L-fucose. We calculated the IC_{50} value for competitive inhibition by free L-fucose to be significantly higher for SrDupFLD (351.72 ± 96.06 mM) than for SrFLD (5.80 ± 1.05 mM) ($P = 0.03$, one-tailed Student's t -test) (Figure 2D). This confirmed that binding strength is increased following N-terminal partial duplication of the FLD.

We hypothesized that partial duplication of the FLD might increase binding avidity because the duplicated partial FLD region contains the CDR loops, CDR1, CDR2, CDR3 and CDR4, and the FLD sequence motif with the conserved residues implicated in L-fucose binding. We reasoned that the engineered protein, SrDupFLD might have two L-fucose-binding pockets as opposed to just one in

the wild-type protein, SrFLD. To check if this was true, we mutated the conserved residues (His, Arg, Arg) implicated in L-fucose binding in the N-terminal partial FLD region to Ala. We found that this mutant, SrDup_{TM}FLD also bound specifically to biotin-PAA- α -L-fucose provided in the solution, and behaved similar to SrDupFLD in competitive ELLA, with an IC_{50} value of 231.84 ± 31.18 mM for inhibition with L-fucose (Figure 2C, D). In one-tailed Student's t -tests, this IC_{50} value was found to be significantly higher than that of SrFLD ($P = 0.009$), but not that significantly lower than that of SrDupFLD ($P = 0.089$). Further, the mutant, SrDupFLD_{TM}, which lacks the conserved residues in the complete FLD region but has the conserved L-fucose binding residues in the N-terminal partial FLD region, displayed negligible binding to PAA- α -L-fucose when compared to SrFLD, SrDupFLD and SrDup_{TM}FLD ($P = 0.006$, 0.003 and 0.002, respectively, one-tailed Student's t -test) (Figure 2C, D). Our data implied that although partial duplication of the FLD does result in higher binding strength, the L-fucose binding residues in the duplicated partial FLD region of SrDupFLD are not responsible for the increase in the binding strength.

We also used an ELLA with a modified set-up that more closely mimics the multivalent presentation of glycans on cell surfaces, i.e., FLDs in solution binding to a multivalent presentation of immobilized biotin-PAA- α -L-fucose. We found that SrFLD showed negligible

Fig. 2. Glycan binding studies of SrFLD and SrDupFLD. (A) SDS-PAGE showing SrFLD (17.11 kDa) and SrDupFLD (28.93 kDa) purified using Ni-NTA metal ion affinity chromatography; Lane 1: Protein marker from Puregene; Lane 2 and 3: purified proteins. (B) Circular dichroism spectra of wild-type lectin, SrFLD and engineered lectin, SrDupFLD. (C) Histogram showing ELLA results for binding of SrFLD, SrDupFLD, SrDup_{TM}FLD and SrDupFLD_{TM} to biotinylated PAA-α-L-Fucose and PAA-α-D-Galactose. (D) ELLA graph showing inhibition curves for L-fucose inhibition of SrFLD, SrDupFLD, SrDup_{TM}FLD and SrDupFLD_{TM} binding to biotinylated PAA-α-L-fucose. (E) Graph showing apparent binding of SrFLD, SrDupFLD, SrDup_{TM}FLD and SrDupFLD_{TM} to immobilized biotinylated PAA-α-L-fucose determined using ELLA. (F) Photographic image of agglutination of human type-O erythrocytes (in triplicate) by SrFLD, SrDupFLD, SrDup_{TM}FLD and SrDupFLD_{TM}. The buffer control is also shown.

binding to immobilized biotin-PAA-α-L-fucose in this ELLA format, whereas the engineered lectin, SrDupFLD and the mutant, SrDup_{TM}FLD bound to biotin-PAA-α-L-fucose with apparent affinities of 84.6 ± 8 nM and 267.5 ± 55 nM, respectively (Figure 2E). These results also indicated that partial duplication of the FLD results in increased binding strength. The mutant, SrDupFLD_{TM}, which lacks L-fucose binding residues in the complete FLD region but contains L-fucose binding residues in the N-terminal partial FLD region, displayed negligible binding to the biotin-PAA-α-L-fucose in this ELLA format, too (Figure 2E). We concluded that the partially duplicated FLD region does not harbor a functional L-fucose binding site, and therefore the increase in binding strength of the engineered lectin with partial FLD duplication (SrDupFLD) is not due to the presence of an additional L-fucose binding site in the duplicated region.

Hemagglutination assay

We performed hemagglutination assays and found that SrDupFLD and SrDup_{TM}FLD agglutinated human type-O erythrocytes. In contrast, SrFLD and SrDupFLD_{TM} did not agglutinate these erythrocytes (Figure 2F). Successful agglutination of erythrocytes requires an avid network of interactions, which is mediated by multiple CRDs on a lectin entity, either through tandem repeats of the CRD on a polypeptide or through oligomerization. Considering that the mutant, SrDupFLD_{TM} did not agglutinate erythrocytes, we inferred that partial duplication of the FLD might increase oligomerization of the lectin polypeptide.

Surface Plasmon Resonance (SPR)

We analyzed the binding parameters of SrFLD and SrDupFLD for immobilized biotin-PAA-α-L-fucose using SPR. We attempted to calculate the dissociation constants (K_D) of SrFLD, SrDupFLD, SrDup_{TM}FLD and SrDupFLD_{TM} using several concentrations of the respective proteins. However, we could not fit the data from different concentrations with global analysis, and due to rapid dissociation rates, also reported in the literature for other lectin-ligand interactions, we were able to use only a narrow range of the data for individual curve fitting with a simple 1:1 interaction model. Considering the narrow range of data used to fit dissociation rates, oligomeric status of the proteins (see subsequent sections), and consequential avidity effect, we note that the average dissociation constants (K_D) calculated be regarded as apparent dissociation constants.

SrFLD had an apparent dissociation constant of 6.60 ± 2.35 μM, and rapid association and dissociation rates indicative of weak binding (Figure 3A). In contrast, the engineered lectin, SrDupFLD had a ~12-fold lower apparent dissociation constant of 0.55 ± 0.16 μM for immobilized biotin-PAA-α-L-fucose (Figure 3A). SrDup_{TM}FLD displayed binding kinetics (apparent K_D : 1.03 ± 0.4 μM) similar to SrDupFLD (Figure 3A). SrDupFLD_{TM} did not display any detectable binding to immobilized biotin-PAA-α-L-fucose (data not shown). These results are in agreement with the ELLA results and suggest that the lower apparent dissociation constant observed for SrDupFLD binding to biotin-PAA-α-L-fucose might be due to either increased affinity or increased avidity that is not

Fig. 3. Binding of *SrFLD* and *SrDupFLD* to multivalently presented L-fucose and fucosylated glycoconjugates. (A) SPR sensorgrams representing binding of *SrFLD*, *SrDupFLD* and *SrDup_{TM}FLD* to biotinylated PAA- α -L-fucose coated on a CM5 sensor chip. The black lines over each sensorgram are the association and dissociation fit curves. (B) Histogram showing CFG glycan microarray binding results of *SrFLD* (in red) and *SrDupFLD* (in black) to fucosylated glycans of different glycan categories on the CFG array version 5.3.

mediated by an additional L-fucose binding site in the duplicated partial FLD region.

Glycan microarray analysis

Glycan microarray analysis of *SrFLD* and *SrDupFLD* reiterated that *SrDupFLD* in solution has higher binding strength than *SrFLD* for immobilized, “multivalently presented”, fucosylated glycoconjugates. *SrDupFLD* displayed binding to fucosylated glycans classified into different categories like Blood group H type 2, Lewis^y, Lewis^b, Blood group A, Blood group B and Fuc α 1-6, hinting at its broad glycan binding specificity (Figure 3B and Supplementary data, Table SIII). However, *SrFLD* displayed negligible binding, albeit to the same glycan groups (Figure 3B and Supplementary data, Table SII).

Isothermal titration calorimetry (ITC)

We used ITC to evaluate the stoichiometry and affinity of *SrFLD* and *SrDupFLD* binding to L-fucose and to the fucosylated oligosaccharides, type-2 blood group H trisaccharide and Lewis^y tetrasaccharide (Figure 4). The dissociation constant, number of binding sites per protein (n), and changes in enthalpy (ΔH), entropy ($-\Delta S$), entropy contribution ($-T\Delta S$), and Gibbs free energy (ΔG) of binding are presented in Table I. The number of binding sites per protein, n , was observed to be close to 1 for both *SrFLD* and *SrDupFLD* with

L-fucose and the fucosylated oligosaccharides, confirming that no additional functional L-fucose binding site was present even in the engineered lectin with partial duplication. The fractional value of n ($n < 1$) for Lewis^y tetrasaccharide [Fuc α 1-2Gal β 1-4(Fuc α 1-3)GlcNAc] might be due to cooperativity, i.e., two protein molecules binding to the two L-fucose units on one glycan molecule. As is typical of lectin–ligand interactions, binding of *SrFLD* and *SrDupFLD* to L-fucose and the fucosylated oligosaccharides was driven by a negative enthalpy change that was offset by an unfavorable entropy change, resulting in an overall negative change in Gibbs free energy. Both enthalpic and entropic contributions were higher for binding to Lewis^y than to L-fucose or H triose type-2, and the entropic contribution was insignificant for H triose type-2.

Importantly, the ITC data indicated that partial duplication of the FLD did not result in higher affinity for L-fucose or fucosylated oligosaccharides. The dissociation constant, K_D , for L-fucose and the fucosylated oligosaccharides in solution were in a similar range for *SrFLD* and *SrDupFLD*. We determined K_D values of *SrFLD* and *SrDupFLD* to be $\sim 174 \mu\text{M}$ and $219 \mu\text{M}$, respectively, for L-fucose, $\sim 46 \mu\text{M}$ and $\sim 43 \mu\text{M}$, respectively, for type-2 H trisaccharide, and $\sim 12 \mu\text{M}$ and $\sim 5 \mu\text{M}$, respectively, for Lewis^y tetrasaccharide (Table I). The K_D values indicated that *SrFLD* and *SrDupFLD* possess the highest affinity for Lewis^y tetrasaccharide, and lower affinities for type-2 H trisaccharide and L-fucose, as also determined by

Fig. 4. Binding of SrFLD (in red) and SrDupFLD (in black) to L-fucose and fucosylated glycans in solution. ITC plot obtained from the titration of SrFLD with (A) L-Fucose, (B) Blood group H antigen triose type 2, and (C) Lewis^Y antigen tetraose. ITC plot for the titration of SrDupFLD with (D) L-Fucose, (E) Blood group H antigen triose type 2, and (F) Lewis^Y antigen tetraose. The solid lines represent the best least-square fit to experimental data using a one-site model.

Table I. Thermodynamic parameters for the binding of SrFLD and SrDupFLD to different glycans mentioned in the table using a single-site model at 25°C

Ligand	Lectin	<i>n</i>	<i>K_D</i> (μM)	Δ <i>H</i> (Kcal/mol)	−Δ <i>S</i> (cal/mol/deg)	− <i>T</i> Δ <i>S</i> (Kcal/mol)	Δ <i>G</i> (Kcal/mol)
Fucose	SrFLD	1.4	174	−8.4	10.9	3.2	−5.1
	SrDupFLD	0.8	219	−7.3	7.7	2.3	−5.0
H triose type-2	SrFLD	0.8	46	−5.9	0	0	−5.9
	SrDupFLD	0.9	43	−5.5	−1.4	−0.4	−5.9
Lewis ^Y tetraose	SrFLD	0.7	12	−15.6	29.6	8.8	−6.7
	SrDupFLD	0.7	5	−12.2	16.6	4.9	−7.2

SPR assays for the wildtype, full-length *S. roseum* FLD-containing polypeptide (Bishnoi et al. 2018). The *K_D* values of SrFLD and SrDupFLD for L-fucose were higher as determined by ITC (~174 μM and ~219 μM, respectively) than SPR (~7 μM and 0.6 μM, respectively). We conjectured that this difference might be due to the

protein adopting a higher order oligomeric form, which would result in the existence of avid interactions with an immobilized multivalent L-fucose-containing probe (biotin-PAA-α-L-fucose) in the SPR assay as opposed to monovalent interactions with L-fucose in solution in the ITC experiment.

Oligomeric profile of wild type and engineered *S. roseum* FLDs

We analyzed intact masses of *SrFLD* and *SrDupFLD* using MALDI-MS as well as ESI-LC-MS. Both *SrFLD* and *SrDupFLD* were found to be mixtures of the monomer and higher oligomers (Figure 5A and B, and Supplementary data, Tables SIV and SV).

We further used analytic size exclusion chromatography to analyze the oligomeric status of *SrFLD* and *SrDupFLD*. We found that both *SrFLD* and *SrDupFLD* were heterogeneous populations, as indicated by the intact mass analysis. Using Superdex 200 gel matrix, we found four *SrFLD* species of different molecular masses. We extrapolated the molecular mass of the *SrFLD* species with the highest elution volume to be ~9 kDa (as opposed to the theoretically calculated molecular mass of 17.1 kDa for the *SrFLD* monomer) using the linear equation between log (molecular mass) and V_e/V_o (ratio of elution volume to void volume) that we obtained with standard proteins of known molecular mass [Bishnoi et al. 2018]. We confirmed by SDS gel electrophoresis that all the protein fractions indeed had intact protein of expected size [Bishnoi et al. 2018] and were not degraded protein or other protein contaminants. The results of our size exclusion chromatography therefore implied anomalous mobility.

In order to rule out the possibility of potential interaction of *SrFLD* with the carbohydrate-based Superdex 200 gel matrix via its lectin activity, we also performed size exclusion chromatography with Sephacryl S200 gel matrix (composed of allyl dextran and *N,N*-methylene bisacrylamide) that lacks any carbohydrate moiety (Figure 5C). We obtained a similar elution profile with four peaks (P1, P2, P3 and P4) indicating anomalous migration with Sephacryl S200 gel matrix too (Figure 5C). The species with the highest elution volume (P4) was calculated to be <5 kDa (the fractionation limit of the column) (Supplementary data, Table SVI). Our results confirmed that (1) the anomalous mobility was not due to any carbohydrate-mediated binding of the protein to the gel matrix but was perhaps due to the protein assuming a non-spherical shape, and (2) the protein exists as a heterogeneous population of various oligomers in solution.

With size exclusion chromatography of *SrDupFLD* on Sephacryl S200 gel matrix, we observed that the peak P4 corresponding to the fraction (that was confirmed by SDS-PAGE analysis to be the last eluting fraction containing intact *SrDupFLD*) had a mass of ~15 kDa (as opposed to the theoretically calculated molecular mass of 28.93 kDa for the monomer) (Figure 5D, Supplementary data, Figure S1 and Table SVI). Besides this species, the chromatography profile indicated the presence of three other *SrDupFLD* species (P1,

Fig. 5. Oligomeric status of *SrFLD* and *SrDupFLD*. MALDI-TOF spectra determining the intact mass of (A) *SrFLD* (in red) and (B) *SrDupFLD* (in black) using AB Sciex TOF/TOF™ Series Explorer™ 72092. Chromatograms showing size exclusion chromatography of (C) *SrFLD* (in red) and (D) *SrDupFLD* (in black) using Hiprep 16/60 Sephacryl S-200 High Resolution column. The small inset in the chromatograms shows the SDS-PAGE analysis of the respective peaks obtained in the chromatograms (P represents the protein sample prior to size exclusion chromatography, P1–P6 are the fractions corresponding to the peaks obtained during size exclusion chromatography and M is the marker). Fractions with low amounts of protein were concentrated prior to SDS-PAGE analysis, and therefore do not reflect actual peak intensities.

P2 and P3) with molecular masses of ~43 kDa, 120 kDa and 241 kDa, respectively (Figure 5D, Supplementary data, Figure S1 and Table SVI). This indicated that *SrDupFLD* also migrates anomalously on Sephacryl S200 gel matrix. Considering the ~15 kDa species to be the monomeric species, the masses of P1, P2 and P3 roughly correspond to the masses of the trimer, nonamer, and 18-mer (Figure 5D, Supplementary data, Figure S1 and Table SVI). Considering the anomalous migration of *SrDupFLD*, it is possible that these assignments are inaccurate. It is also possible that higher order oligomers exist, which are not within the fractionation range of the column and are hence not resolved. The elution profile of *SrDupFLD* with Superdex 200 gel matrix also displayed multiple oligomeric species similar to Sephacryl S200 (Supplementary data, Figure S2).

Using the size exclusion chromatography data of Sephacryl S200 gel matrix, we quantified *SrFLD* and *SrDupFLD* oligomers, P1, P2, P3 and P4 to be in the proportions (with P1 corresponding to the first eluting species and P4 to the last eluting species) P1:P2:P3:P4::8%:8%:26%:58% and P1:P2:P3:P4::15%:42%:23%:21%, respectively (Figure 5C and D, Supplementary data, Figure S1 and Table SVI).

In *SrFLD*, the lowest oligomeric form (P4) thus assumes the highest proportion and the higher order oligomers P1 and P2 are

present at <10% whereas in *SrDupFLD* the higher order oligomer (P3) is present at highest proportion and the putative monomer is present at lower proportion than in *SrFLD* (21% in *SrDupFLD* vs. 58% in *SrFLD*). This difference is expected to result in the higher glycan binding strength observed in *SrDupFLD*.

In the presence of L-fucose, both *SrFLD* and *SrDupFLD* show a shift in the size exclusion chromatography profile. The proportions of the *SrFLD* and *SrDupFLD* oligomers are P1:P2:P3:P4::16%:9%:22%:53% and P1:P2:P3:P4::6%:65%:20%:9%, respectively (Figure 5C and D, Supplementary data, Figure S1 and Table SVI).

It should be mentioned that there was also a shift in the elution times of the four peaks, P1, P2, P3 and P4 in the presence of L-fucose, with a trend of increasing molecular mass in the presence of L-fucose (Supplementary data, Figure S1 and Table SVI). The shift was especially significant for the peak P4 in both *SrFLD* and *SrDupFLD* (Supplementary data, Figure S1 and Table SVI).

Further, we determined that L-fucose effected the change in the oligomeric profiles of *SrFLD* and *SrDupFLD* via the L-fucose binding site. No effect of L-fucose on oligomeric profile was observed for *SrDupFLD_{TM}* (data not shown), and there was no similar effect of D-galactose on the oligomeric profiles of *SrFLD* and *SrDupFLD* (Supplementary data, Figure S3). Further, even 10 μ M L-fucose was

Fig. 6. Glycan binding studies of *BfFLD* and *BfDupFLD*. (A) SDS-PAGE showing purified *BfDupFLD* (30.6 kDa) and *BfFLD* (19.9 kDa) using Ni-NTA metal ion affinity chromatography; Lane 1: Protein marker from NEB; Lanes 2, 3 and 4: purified proteins. (B) Histogram showing ELLA results for binding of *BfFLD*, *BfDupFLD*, *BfDupmFLD*, *BfDupFLD_m*, and *BfFLD_m* to biotinylated PAA- α -L-fucose. (C) Graph showing inhibition curves for L-fucose inhibition of *BfFLD*, *BfDupFLD*, *BfDupmFLD* and *BfDupFLD_m* binding to biotinylated PAA- α -L-fucose calculated using ELLA. (D) Histogram showing CFG glycan microarray binding results of *BfFLD* to fucosylated glycans of different glycan categories on the CFG array version 5.3.

sufficient to bring about the change in the oligomeric profile of SrFLD and SrDupFLD (Supplementary data, Figure S4).

We also analyzed SrFLD and SrDupFLD by reducing and non-reducing SDS-PAGE for the presence of oligomeric forms. We observed monomeric and dimeric forms in both SrFLD and SrDupFLD, and higher order oligomeric forms also in SrDupFLD (Supplementary data, Figure S5).

Purification of wild type and mutant *B. floridae* FLDs and glycan binding studies

We expressed the recombinant *B. floridae* protein, BfDupFLD (30.91 kDa) which has an N-terminal partial FLD region followed by a complete FLD, and its deletion mutant, BfFLD (20.38 kDa), which lacks the N-terminal partial duplication. We purified both of them by metal ion affinity chromatography to about 95% purity (Figure 6A). Using ELLA, we found that both BfFLD and BfDupFLD bound with higher intensity to biotin-PAA- α -L-fucose than to the negative control, biotin-PAA and that the binding to biotin-PAA- α -L-fucose was abolished by the presence of free L-fucose (Figure 6B). BfFLD (inconsistently) bound with lower intensity than BfDupFLD to biotin-PAA- α -L-fucose, and its binding was inhibited to a greater extent by the presence of L-fucose in solution. We calculated the IC_{50} values for inhibition of BfFLD and BfDupFLD binding to biotin-PAA- α -L-fucose by free L-fucose to be 0.82 ± 0.13 M and 1.19 ± 0.13 M, respectively ($P = 0.22$, two-tailed Student's *t*-test) (Figure 6C). The results of this ELLA suggested that lectin activity is retained following deletion of the N-terminal partial FLD region.

To check whether BfDupFLD has two functional L-fucose binding sites, we performed site-directed mutagenesis of the conserved His residues implicated in L-fucose binding in the N-terminal partial FLD region (H386A) and in the complete FLD region (H481A). We found that the deletion mutant with the H481A mutation, BfFLD_m, bound negligibly to biotin-PAA- α -L-fucose (Figure 6B). This confirmed that His481 is critical for L-fucose binding. However, BfDup_mFLD (which has the mutation, H386A) showed similar binding as BfDupFLD to biotin-PAA- α -L-fucose with an IC_{50} value of 1.03 ± 0.06 M for inhibition with L-fucose ($P = 0.4$, two-tailed Student's *t*-test) (Figure 6B, C), suggesting that the L-fucose binding site of the N-terminal partial FLD region is superfluous. Further, BfDupFLD_m (which has the mutation H481A) showed negligible binding to biotin-PAA- α -L-fucose with an IC_{50} value of 0.04 ± 0.03 M, confirming that the N-terminal partial FLD region does not possess L-fucose binding ability and that the L-fucose binding site of the complete FLD region is the functional one (Figure 6B, C). The low level of binding to biotin-PAA- α -L-fucose demonstrated by BfDupFLD_m might be due to the incomplete abolition of lectin activity by the H481A mutation since BfFLD_m also demonstrates a similar low level of binding (Figure 6B and C).

We also characterized the glycan binding specificity of the *B. floridae* F-type lectin by glycan microarray analysis. BfFLD and BfDupFLD exhibited similar glycan binding specificity (Figure 6D and Supplementary data, Figure S6, Tables SVII and SVIII). They displayed binding to two motifs, H type-2 (Fuc α 1-2Gal β 1-4GlcNAc) and Lewis^Y (Fuc α 1-3(Fuc α 1-2Gal β 1-4)GlcNAc).

Discussion

The rationale for our study came from the observation that certain proteins in *B. floridae*, *L. chalumnae* and *C. picta bellii* have a partial FLD annexed to a complete FLD. Lectins typically display weak

affinity towards glycan ligands and this is strengthened by avidity arising out of either tandem repeats of lectin domains or multivalency (Lis and Sharon 1998). Our study was aimed at assessing if a partial lectin domain containing the sugar binding site would be able to lead to multivalency and thereby increased binding strength.

Protein domains are considered to be “evolutionarily indivisible, structurally compact units from which larger functional proteins are assembled” (Triant and Pearson 2015). Nevertheless, partial domains constitute a significant percent of all annotated protein domains. Around 5–10% of protein domains in the Pfam 27 are shorter than <50% of the characteristic domain length (Triant and Pearson 2015). Triant and Pearson examined 30961 partial domain regions from 136 domain families (Triant and Pearson 2015) in the RPD2 database (a database containing a subset of Pfam-A entries (curated entries from Pfam 27 database) with improved domain boundaries and homology relationships) (Gonzalez and Pearson 2010). They found that partial domains are mainly of three types – split domains, bounded partial domains and unbounded partial domains. Split domains (which constituted more than half of the partial domains in RPD2) are actually complete domains that have been split into multiple incomplete domain parts during the HMM alignment procedure (Triant and Pearson 2015). Bounded partial domains (~20% of partial domains in RPD2) lie in the N-terminal or C-terminal end of the polypeptide and/or are bounded by other non-homologous domains (Triant and Pearson 2015). They are frequently found in low quality protein predictions in eukaryotic genomes and are often unreviewed protein sequences lacking an ENSEMBL gene model (Zerbino et al. 2018), suggesting that their occurrence could be due to inaccuracies in gene assembly or splicing model (Triant and Pearson 2015). Unbounded partial domains are domains that lie unbounded in a region that could theoretically contain the remaining domain parts (Triant and Pearson 2015). Their occurrence could be due to inaccuracies in sequence alignment (Triant and Pearson 2015). On the whole, the study by Triant and Pearson suggests that partial domains be viewed with suspicion (Triant and Pearson 2015). However, they do find that there are bonafide structural partial domains, mainly among unbounded partial domains, where large evolutionary distances lead to incomplete HMM alignments (Triant and Pearson 2015). Triant and Pearson posit that genuine structural partial domains would be characterized by evolutionary mobility (i.e., the same partial domain would occur in different protein contexts) and structural compactness (Triant and Pearson 2015).

Genuine partial domains might also be atrophied domains (Prakash and Bateman 2015). Domain atrophy, which occurs in ~0.06% of all domains studied, involves large deletions of core secondary structural elements of domains without loss of function (Prakash and Bateman 2015). This is in contrast to domain length variations, which are caused by insertion or deletion of loops, coils or a few secondary structural elements without affecting the core (Pascarella and Argos 1992; Taylor et al. 2004; Sandhya et al. 2009). Domain atrophy is frequently aided by compensatory mechanisms such as domain–domain or subunit–subunit interactions (Prakash and Bateman 2015).

Considering these studies of partial domains, it is important to distinguish between partial domains that are annotated as such due to sequencing/assembly/gene model errors and genuine partial domains.

In this context, we would also like to make a note of another interesting study that used domain co-occurrence information together with BLAST analysis to annotate new domains. Menichelli

et al. found 2240 new domains of *Plasmodium falciparum* proteins in the Pfam database, of which 7% were partial matches to already known Pfam families (Menichelli et al. 2018). Among such domains, domains with <80% overlap but good *P* values tended to be longer than the known Pfam domain (in 89% cases), and Menichelli et al. suggest that these domains could be extensions of smaller Pfam domains rather than genuine partial domains (Menichelli et al. 2018).

From these studies, we can infer that a domain annotated as a “partial lectin domain” could be non-existent (being a consequence of sequencing/assembly/gene model errors) or an extension of a smaller Pfam domain or a genuine partial domain.

The *B. floridae* FLD sequence referred to here in the manuscript is annotated in NCBI as “hypothetical protein BRAFLDRAFT_118409 [*Branchiostoma floridae*]” (NCBI Reference Sequence gi260835699|refl XP_002612845.1|GenBank EEN68854.1|mRNA XM_002612799.1|gene 7222252; UniprotKB C3XSY6_BRAFL). The sequence entry XP_002612845.1 is a provisional Refseq not yet subject to final NCBI review, the Genbank entry EEN68854 entry includes the comment “Method: conceptual translation”, the UniprotKB entry C3XSY6 is unreviewed with annotation score 2/5 (protein predicted), and the gene lacks an annotated ENSEMBL gene model. However, our analysis of the nucleotide sequence did not point to any obvious inaccuracy in genome sequence assembly.

Further, the genome sequence of *B. floridae* as well as the transcriptomes of *B. floridae*, *Branchiostoma lanceolatum* and *Branchiostoma belcheri* are known (Yu et al. 2008; Jin et al. 2010; Oulion et al. 2012; Yang et al. 2016). The filtered gene model of “hypothetical protein BRAFLDRAFT_118409 [*Branchiostoma floridae*]” built on JGI MycoCosm server with ESTs indicates the presence of the partial FLD (Putnam et al. 2008). Moreover, ESTs reported by Yu et al. in their study (Yu et al. 2008) map to various regions of the coding sequence of “hypothetical protein BRAFLDRAFT_118409 [*Branchiostoma floridae*]” including the stretch coding for the partial FLD (Supplementary data, Figure S7). These data suggest that the partial FLD in “hypothetical protein BRAFLDRAFT_118409 [*Branchiostoma floridae*]” is existent. Interestingly, the *B. belcheri* transcript of “hypothetical protein BRAFLDRAFT_118409-like”, likely homolog of our gene of interest, “hypothetical protein BRAFLDRAFT_118409 [*Branchiostoma floridae*]” was found to be among the list of transcripts found to be differentially expressed in the various stages of amphioxus (Yang et al. 2016).

The “hypothetical protein BRAFLDRAFT_118409 [*Branchiostoma floridae*]” is a multi-domain protein (Supplementary data, Table SI), as per the online CD-Search tool in the NCBI website (Marchler-Bauer and Bryant 2004). The partial lectin domain is the third domain from the N-terminal; it follows two complete F5/8 type C domains, and is followed by nine complete F5/8 type C domains, five calcium-binding EGF-like domains and a Glycolipid transfer protein (GLTP) domain. Therefore, the partial FLD is bounded on both N- and C-terminal ends and therefore is not a structural partial domain with large evolutionary distance leading to incomplete HMM alignment. Although it is a bounded partial domain, it is not bounded by other domains but by FLDs on both sides. The partial FLD (FLD3) and the eleven complete FLDs have FLD sequence motifs (Supplementary data, Table SIX) and overall domain sequences with high sequence identity (Supplementary data, Figure S8). The sequence identity ranges from ~62% to 100% (Supplementary data, Table SX). This is highly indicative of multiple gene duplication and gene fusion events. We speculate that the partial FLD might have arisen either by an internal partial gene duplication event (i.e., extension/

expansion of a smaller domain) or by a complete gene duplication event followed by atrophy.

Attempts to recombinantly express only the partial FLD in our lab were unsuccessful, suggesting that it might not form a structural/folding unit by itself. This coupled with the absence of this partial FLD in other architectural contexts (i.e., with the partial FLD immediately bounded by domains other than a complete FLD) in our bioinformatics searches supports the idea that this partial FLD is not a partial structural domain but a consequence of either internal partial gene duplication or domain atrophy compensated for by interactions with the adjacent FLD. We would like to emphasize that although we did not find this partial FLD in other architectural contexts, the presence of similarly partially duplicated FLDs not only in another *B. floridae*, but also in other species – three in *Latimeria chalumnae* and one in *Chrysemys picta bellii* is also indicative of “evolutionary mobility” and supports our idea that these partial FLDs do indeed exist.

Inspired by the presence of F-type lectins with naturally occurring and evolutionarily conserved partial domain duplication (Bishnoi et al. 2015), we engineered a similar partial duplication at the N-terminus of an FLD from a *S. roseum* protein (which has an FLD in tandem with two other co-associated domains, an α -L-fucosidase domain and a NPCBM-associated domain; this full-length polypeptide is discussed in Bishnoi et al. 2018). CD spectroscopy indicated that the engineered lectin, *SrDupFLD* has a secondary structure profile similar to *SrFLD*. Arnold et al. have previously demonstrated that gene duplication of an 8-stranded β -barrel produces a functional 16-stranded β -barrel through barrel expansion (Arnold et al. 2007). We predict that our engineering the FLD by partial domain duplication of the FLD also results in an expanded β -barrel (with the same secondary structural elements and thus a similar CD spectrum as observed in Figure 2B).

Intact mass analysis and analytical size exclusion chromatography indicated that both *SrFLD* and *SrDupFLD* exist as a mixture of monomer and higher oligomers (Figure 5). Following size exclusion chromatography (on a HiPrep 16/60 Sephacryl S200 High Resolution column), a single peak fraction collected of *SrDupFLD* still displayed some oligomeric heterogeneity (as determined by another step of size exclusion chromatography on a Superdex 200 10/300 column or a HiPrep 16/60 Sephacryl S200 High Resolution column) (data not shown), precluding the isolation of a single oligomeric species. This might be either due to the incomplete resolution of the various peak fractions in the first chromatography run or due to equilibrium between the various oligomeric species. Further, both *SrFLD* and *SrDupFLD* migrated anomalously both in Superdex 200 gel matrix as well as in Sephacryl S200 gel matrix (that lacks any carbohydrate moiety). The oligomeric status could not be determined for *SrFLD* because the putative monomer peak had a calculated molecular mass <5 kDa (the fractionation limit of the column). For *SrDupFLD*, we predict the presence of an 18-mer, nonamer, trimer and monomer.

The presence of trimers have also been reported in AAA and *M. saxatilis* FLDs, which are trimers with three L-fucose binding sites on the same face of the molecule (Bianchet et al. 2002, 2010). However, not all FLDs are trimers. The crystal structure of the FLD containing *S. mitis* lectinolysin reveals a hydrogen-bonded head-to-head dimer, which is physiologically not relevant as the lectinolysin polypeptide as well as just the FLD of lectinolysin are a mixture of monomers and disulfide-linked dimers in solution (Feil et al. 2012). Similar disulfide-linked dimers have been reported in AAA, and Bianchet et al. have commented that these dimers are likely formed

via oxidation of the surface exposed contiguous Cys residues that are part of the L-fucose binding site, and therefore these dimers are not physiologically relevant (Bianchet et al. 2002). We also observed similar disulfide-linked dimers in *SrFLD* and *SrDupFLD*, and disulfide-linked higher order oligomers also in *SrDupFLD* by SDS-PAGE analysis (Supplementary data, Figure S5).

In *SrFLD*, the lowest oligomeric form (P4) assumes the highest proportion and the higher order oligomers P1 and P2 are present at <10% whereas in *SrDupFLD* the higher order oligomer (P3) is present at highest proportion and the putative monomer is present at lower proportion than in *SrFLD* (21% in *SrDupFLD* vs. 58% in *SrFLD*). This difference is expected to result in the higher glycan binding strength observed in *SrDupFLD*. Incubation with L-fucose shifted the equilibrium further towards higher oligomeric forms for both *SrFLD* and *SrDupFLD* with a concomitant decrease in the proportion of the putative monomeric species in the presence of L-fucose. However, the proportion of the monomeric species in *SrFLD* still remained much higher than that in *SrDupFLD* (53% in *SrFLD* vs 9% in *SrDupFLD*). Concomitantly, the proportion of the higher order oligomeric forms (P1 and P2 considered together) remained higher in *SrDupFLD* as compared to *SrFLD* (71% in *SrDupFLD* vs. 25% in *SrFLD* for P1 and P2). Thus, *SrDupFLD* is expected to display greater glycan binding strength even in the presence of L-fucose. Thus, our data indicates that the L-fucose-bound FLD is more liable to higher order oligomerization. We speculate that L-fucose binding makes the monomer interface more suitable for oligomerization.

We studied the effect of partial FLD duplication on the binding strength of *SrFLD* for its ligand, α -L-fucose. ITC experiments demonstrated no significant difference in the stoichiometry, dissociation constant, or thermodynamic parameters of binding of *SrFLD* and *SrDupFLD* for L-fucose, Blood group H type-2 trisaccharide, and Lewis^y tetrasaccharide, suggesting that partial FLD duplication does not result in the formation of an additional functional L-fucose binding site (Figure 4).

We made proteins with mutations of the L-fucose binding site (His623, Arg651 and Arg658) either at the N-terminal partial FLD region or the complete FLD region in order to dissect the contribution of each of these L-fucose binding sites to actual lectin function. Using ELLA, hemagglutination and SPR assays, we found that *SrDup_{TM}FLD* behaved similar to *SrDupFLD* whereas *SrDupFLD_{TM}* lost lectin activity, confirming that the partial FLD region does not harbor a functional L-fucose binding site. Considering that three mutations might be an overload, we also performed experiments with single mutants, *SrDup_{H623}FLD* and *SrDupFLD_{H623}* and obtained similar results (Supplementary data, Figure S9).

We also used ELLA and glycan array analysis to analyze binding of *SrFLD* and *SrDupFLD* to fucosylated glycoconjugates. Besides tandem repetition and oligomerization, surface presentation is also known to bring about multivalency and hence increased avidity in lectins (Collins and Paulson 2004). We used two different assay formats of ELLA – one wherein the lectin in solution is allowed to bind to an immobilized (multivalently presented) and multivalent α -L-fucose-containing probe, and the other wherein immobilized (multivalently presented) lectin is allowed to bind to a multivalent α -L-fucose-containing probe in solution. In the former case, i.e., when an FLD in solution binds to an immobilized multivalent L-fucose-containing probe, it follows that the binding strength can be increased by the multivalent presentation of the FLD, either through the presence of multiple L-fucose binding sites or through increased oligomerization. The latter case, i.e., the ELLA with immobilized lectin, inherently involves multivalent interactions and is therefore not

predicted to discriminate well between lectin variants differing in oligomerization and/or domain repetition. As expected, *SrFLD* and *SrDupFLD* displayed similar binding in this ELLA format. However, *SrFLD* in solution did not bind to the immobilized multivalent L-fucose-containing probe. Partial duplication of the FLD in the engineered lectin, *SrDupFLD* resulted in an increased binding strength, apparent from both the better apparent affinity in the ELLA format with the immobilized multivalent L-fucose-containing probe and the higher IC₅₀ value for L-fucose in competitive ELLA conducted with immobilized lectin.

The increased binding of *SrDupFLD* in solution with the immobilized multivalent L-fucose-containing probe, when compared with that of *SrFLD*, suggested an increase in avidity. That *SrDupFLD* and *SrDup_{TM}FLD* but not *SrFLD* and *SrDupFLD_{TM}* could agglutinate erythrocytes suggested that partial FLD duplication might result in increased oligomerization. Considered together with the similar binding affinities obtained by ITC for *SrFLD* and *SrDupFLD* and their different oligomeric profiles, we concluded that partial FLD duplication resulted in increased oligomerization and thus increased avidity of binding to the immobilized, multivalent α -L-fucose-containing probe. We believe that the higher oligomeric status (and not the presence of an additional L-fucose binding site) is key to the increased binding of *SrDupFLD* because of the similar stoichiometry of binding observed in ITC and by the fact that *SrDup_{TM}FLD*, which lacks L-fucose binding residues in the partial FLD region, displays similar apparent affinity to the multivalent L-fucose-containing probe.

We used a similar experimental approach to verify if the N-terminal partial FLD region does possess a functional L-fucose binding site in the naturally occurring *B. floridae* protein. We did not observe a significant difference in the binding strength of *B/FLD* and *B/DupFLD* in a competitive ELLA. The mutagenesis of the His residue in the L-fucose binding site confirmed that the partial FLD region does not contribute towards L-fucose binding. Considering the tandem presence of many complete FLDs in the naturally occurring *B. floridae* protein, we predict that the partial FLD will not have a significant effect on increasing glycan binding strength in the native (full-length, multi-domain) protein. The physiological role of this partial FLD duplication in Nature is therefore not clear. It is possible that it has some as yet unknown function when present in the context of the other co-associated tandem domains in this *B. floridae* polypeptide.

Structurally, it is not known whether the partial FLD would fold as an independent entity or would combine with the consecutive complete FLD to make an expanded β -barrel (as we envision in Figure 1B), and as observed previously for a β -barrel (Arnold et al. 2007). We did attempt to crystallize *SrDupFLD* with the aim of obtaining a three dimensional structural model but could not succeed, probably due to the dynamic equilibrium observed among the various oligomers in solution and the resultant heterogeneity. While we do not have any evidence that the partial FLD is correctly folded in the proteins that we studied, we think it likely because both *B/DupFLD* and *SrDupFLD* recombinant proteins are over-expressed, stable, have similar CD spectra as the respective FLDs, and display lectin activity. Nonetheless, it is also a possibility that exposed hydrophobic residues in the incorrectly folded partial FLD region facilitate monomer–monomer interactions and thereby greater oligomerization and glycan binding strength.

Lectin–carbohydrate interactions often occur in low-affinity binding sites formed by shallow indentations on protein surfaces, involving hydrogen bonding and Van der Waal packing which is

assisted with water-mediated contacts between the protein surface and the saccharides (Quiocho 1986; Weis and Drickamer 1996; Vasta and Ahmed 2009)(Quiocho 1986; Weis and Drickamer 1996). Since individual lectin-carbohydrate interactions are generally weak, multivalency is a central strategy in lectin-ligand interactions to enhance and achieve biologically relevant affinities and specificities of lectins for their cognate ligands. Multivalency of the carbohydrate ligand is frequently attained in living systems via surface (immobilized) presentation (Collins and Paulson 2004), and multivalency of the lectin are accomplished by domain duplication (or tandem repeats) and/or oligomerization of the lectin polypeptide (Lis and Sharon 1998). Alterations in lectin multivalency may have an effect on biological functions, as demonstrated by a study wherein *Ralstonia solanacearum* lectin with reduced valency displayed similar avidity towards surface fucoglycoconjugates (as compared to wildtype lectin) but was unable to effect clustering of glycolipids and induce membrane invaginations (Audfray et al. 2012). Weak affinity interactions are dramatically increased by the association of polypeptides into oligomeric structures in which single binding sites are clustered together, bind to ligand simultaneously, and help to crosslink the glycoconjugates, such as in the bouquet-like oligomer of mannose-binding lectin wherein all the CRDs of the individual subunits are directed towards a target surface (Turner 1996). Oligomerization of lectins is a critical feature that aids not only in the recognition of surface arrays of polysaccharides on potential pathogens, but also in the facilitation of various effector functions such as agglutination, crosslinking of glycoepitopes, opsonization of pathogens, and activation of the complement system (Holmskov et al. 2003; Fujita et al. 2004; Sharon and Lis 2004; Vasta and Ahmed 2009).

The FLD family includes eukaryotic members with roles in innate immunity mechanisms and immune surveillance; most members display tandem domain repeats and/or oligomerization (Saito et al. 1997; Honda et al. 2000; Vasta et al. 2007). *Anguilla anguilla* agglutinin, an F-type lectin found in European eel, is involved in the innate immunity system and known to recognize bacterial liposaccharides (Honda et al. 2000). AAA organizes as a trimer with its three CRDs oriented in the same direction and it can crosslink fucosylated glycoconjugates on the bacterial surface (Bianchet et al. 2002). The binary tandem domain fucoselectin from *M. saxatilis* is well characterized; its two CRDs possess distinct specificities for fucosylated ligands and each CRD exists as a trimer (Bianchet et al. 2010). *Streptococcus pneumoniae* F-type lectin is one of the few exceptional cases of tandem CRDs in bacterial F-type lectins; it has a glycosylhydrolase domain associated with a C-terminal triplet of FLDs (Boraston et al. 2006). Boraston et al. demonstrated that the single CRD is unable to bind immobilized L-fucose whereas a polypeptide with three CRDs binds with good apparent affinity, thus emphasizing the importance of domain repetition in multivalent binding to clustered glycoconjugates (Boraston et al. 2006).

Our study showcases an example wherein a Nature-inspired engineering approach using partial domain duplication indirectly enabled multivalent binding to clustered glycoconjugates via an increased tendency for oligomerization. Our study also indicates that the FLD with its short length, its well defined L-fucose-binding sequence motif and its β -barrel structure that is amenable to engineering via strand insertion, is a model lectin for engineering applications. We believe that strategies similar to the one we adopted here might be useful to engineer and/or exploit other lectins for achieving multivalent binding for various applications.

Materials and methods

Chemicals and reagents

Biotinylated PAA-linked sugars used in this study were obtained from Glycotech Corporation. Blood group antigens were obtained from Elicityl. Monosaccharides were purchased from Sigma.

Bioinformatics analysis

FLD boundaries were determined using the Conserved Domain Database (Marchler-Bauer and Bryant 2004; Marchler-Bauer et al. 2011). Multiple sequence alignments were performed using Clustal Omega (McWilliam et al. 2013) or T-COFFEE, Version_11.00 (Notredame et al. 2000). ESTs were mapped on coding sequence using BLAST (Altschul et al. 1990). FLD structures were retrieved from the Protein Data Bank (www.rcsb.org) (Berman et al. 2000) and visualized using The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. The topology diagram was based on 1K12 structure (Bianchet et al. 2002).

Cloning

The FLD-containing gene from *S. roseum* (Genbank accession number ACZ87343.1) was codon-optimized, custom-synthesized (Genscript), and cloned in pUC57, and the sequence corresponding to the FLD (referred to as *SrFLD*) amplified by PCR with appropriate primers (forward primer, 5'-CATGCCATGGCGCGTCCGAATCTGAGTCTGGG-3', and reverse primer, 5'-GCAGAAGTCCAAGTGCCTGCGCTCGA GCGG-3') and subcloned in the expression vector, pET-28a(+) within NcoI and XhoI sites to create *SrFLD*-pET-28a(+) such that the plasmid would encode recombinant *SrFLD* with a C-terminal hexahistidine tag. Sequences of *SrFLD*, *B/DupFLD*, *B/FLD* and *Anguilla anguilla* agglutinin were aligned in T-COFFEE, Version_11.00. The length of the *SrFLD* sequence that aligned with the duplicated portion of *B. floridae* FLD was identified by visual examination of the alignment (Figure 1A). Forward primer, 5'-CATGCCATGGCAGTCCGAATCTGAGTCTGGG-3' and reverse primer, 5'-CATGCCATGGTAACGTGCACAG CCGTCACACCCGG-3', both with NcoI sites were used to amplify the aligned portion, i.e., the partial *SrFLD* sequence that aligns with the partial duplicated FLD region of *B/DupFLD*, and the amplicon was cloned within the NcoI site of *SrFLD*-pET-28a(+) to create the clone, *SrDupFLD*-pET-28a(+), such that it would encode an FLD prefixed with a partial duplication (Dup) at the N-terminus and suffixed with a C-terminal hexahistidine tag; this protein is referred to as *SrDupFLD*.

SrFLD-pET-28a(+) with binding pocket residues of *SrFLD* mutated, H623AR651AR658A-*SrFLD*-pET-28a(+), was generated using Quikchange lightning site-directed mutagenesis kit (Stratagene). Using PCR amplification and cloning, two constructs of mutant *SrDupFLD* were made – *SrDup_{TM}FLD*-pET28a (with the three mutations, H623A, R651A and R658A, in the N-terminal partially duplicated FLD portion, and wild type residues in the C-terminal complete FLD region) and *SrDupFLD_{TM}*-pET28a (having wild type residues in the N-terminal partially duplicated FLD portion and the three mutations, H623A, R651A and R658A, in the C-terminal complete FLD region).

The gene sequence of the *B. floridae* FLD (with a partial duplication), referred to as *B/DupFLD* (Genbank accession number XP_002612845.1), was codon optimized and synthesized from Genscript, Piscataway, NJ in pUC57 and subcloned in the expression vector, pET-28a(+) within Nco I and Xho I sites to create *B/DupFLD*-pET-28a(+) such that the plasmid would encode recombinant *B/DupFLD* with a C-terminal hexahistidine tag.

Branchiostoma floridae FLD without the duplicated portion, referred to as *BfFLD*, was amplified from *BfDupFLD*-pET-28a(+) using a forward primer with NcoI site, 5'-CATGCCATGGCAAACA TTGCGCTGAACAAGAAAACC-3', and a reverse primer with XhoI site, 5'-CCCCTACGGTGGCGCGGGTCTCGAGCGG-3'. The amplicon was subcloned into Nco I and Xho I sites of pET-28a(+) to create *BfFLD*-pET-28a(+) such that the plasmid would encode recombinant *BfFLD* with a C-terminal hexahistidine tag.

BfFLD-pET-28a(+) was subjected to site-directed mutagenesis to introduce a mutation in the binding pocket, resulting in the clone, *BfFLD*_{H481A}-pET-28a(+). The protein expressed by this construct is referred to as *BfFLD*_m. Two mutant constructs of *BfDupFLD*-pET-28a(+) were made using PCR amplification and cloning, one, *BfDupFLD*_{H386A}-pET-28a(+) (with the H386 residue of the N-terminal partially duplicated FLD portion mutated to alanine) and two, *BfDupFLD*_{H481A}-pET-28a(+) (with the H481 residue of the complete FLD region mutated to alanine). The proteins expressed by these constructs, *BfDupFLD*_{H386A}-pET-28a(+) and *BfDupFLD*_{H481A}-pET-28a(+), are designated *BfDupFLD*_m and *BfDupFLD*_m, respectively. All these constructs were cloned to express proteins with C-terminal hexahistidine tags.

Expression and purification

For expression and purification of recombinant proteins, *E. coli* BL21 (DE3) cells were transformed with the respective expression vector constructs. For the wild-type constructs, *SrFLD* and *BfDupFLD*, and the engineered construct, *BfFLD*, cultures were induced with 1 mM IPTG (GoldBio) and incubated with shaking at 37°C for 4 h at 200 rpm whereas for *SrDupFLD*, *SrDupFLD*_{TM} and *SrDupFLD*_{TM}, cultures were induced with 0.1 mM and incubated with shaking at 22°C for 10 h at 160 rpm. Recombinant proteins were purified by column chromatography using Ni-NTA resin (Pierce). Briefly, the bacterial cell pellet was resuspended in lysis buffer (20 mM Tris, 150 mM NaCl, 20 mM imidazole, pH 7.5), and the cells were disrupted by a probe-type ultrasonicator (Sonics & Materials INC) for 30 min (pulse – 10 s on and 10 s off, amplitude 20%). The cleared lysate was centrifuged at 16000 × g for 40 min and loaded on a Ni-NTA column equilibrated with lysis buffer. The column was subjected to end-over-end rotation on a Rotospin (Tarsons) for 2 h at 4°C. The column was then washed with TBS (20 mM Tris, 150 mM NaCl, pH 7.5) containing 40 mM imidazole and the protein was eluted with TBS containing 300 mM imidazole and dialyzed extensively against TBS. Protein purity was assessed by SDS-PAGE. Protein concentration was estimated by OD₂₈₀ measurement and by Bradford assay.

Circular dichroism spectroscopy

Circular dichroism spectroscopic measurement was done on a Jasco J815-1505 spectropolarimeter (Jasco International Co., Ltd.). Protein samples (0.2 mg/mL) prepared in 20 mM Tris, 20 mM NaCl, pH 7.5 were placed in a rectangular quartz cuvette with 1 mm path length, and spectra were recorded at far-UV (250–190 nm) region at a scan speed of 50 nm/min and a slit width of 2 nm. The mean residue ellipticity, θ was calculated using mean residue molecular mass and number of amino acids in a respective protein and plotted against the wavelength.

Mass spectrometry

The intact mass analysis was done using AB SCIEX 5800 MALDI-TOF/TOF. Protein samples were prepared in 20 mM Tris buffer, pH 7.5.

Analytical size exclusion chromatography

Size exclusion chromatography was done using Superdex 200 10/300 GL (GE Healthcare life sciences) and HiPrep 16/60 Sephacryl S-200 High Resolution column (GE Healthcare life sciences) on an GE Healthcare Akta Purifier FPLC system.

For size exclusion chromatography using Superdex 200, the column was first equilibrated with 20 mM Tris, 100 mM NaCl, 1 mM CaCl₂, pH 7.5 at a flow rate of 0.3 mL/min. Protein samples (4 mg/mL) and reference proteins prepared in the same buffer were used with a sample volume of 0.1 mL. Reference proteins used for column calibration were lysozyme (14.3 kDa), carbonic anhydrase (29 kDa), albumin (66 kDa), alcohol dehydrogenase (150 kDa) and β -amylase (200 kDa).

For size exclusion chromatography using Sephacryl S-200, the column was first equilibrated with 20 mM Tris, 100 mM NaCl, 1 mM CaCl₂, pH 7.5 at a flow rate of 0.5 mL/min. Reference proteins used for column calibration were cytochrome C (12.4 kDa), carbonic anhydrase (29 kDa), albumin (66 kDa), alcohol dehydrogenase (150 kDa) and β -amylase (200 kDa). Protein samples (2 mg/mL) and reference proteins prepared in the same buffer were used with a sample volume of 0.5 mL.

To study the oligomeric state of proteins bound with L-fucose, the protein was incubated with 1 mM L-fucose for 1 h prior to loading on to a column equilibrated with buffer containing 1 mM L-fucose.

Hemagglutination assay

A 2% suspension of type-O human erythrocytes was made in TBS (20 mM Tris HCl, 0.15 M NaCl buffer, pH 7.5). Then, 25 μ L of purified lectin was mixed with 25 μ L of this 2% suspension of erythrocytes in a U-shaped microtitre plate, and the mixture was incubated for 1 h at 37°C. The plate was observed for hemagglutination. TBS was used instead of lectin for control.

Enzyme-linked lectin assay (ELLA)

Purified recombinant proteins (100 μ L of 2 μ g/mL stock solutions in TBS) were coated on MaxiSorp flat-bottom 96 well plate (Nunc) overnight at 4°C. Subsequently, the wells were washed twice with 200 μ L TBST (20 mM Tris, 150 mM NaCl, 0.1% Tween-20, pH 7.5) and once with TBS (20 mM Tris, 150 mM NaCl, pH 7.5). After washing, all the wells were blocked with 3% BSA and incubated at room temperature for 4 h. The wells were again washed. Biotin-PAA- α -L-fucose, biotin-PAA- α -D-galactose or biotin-PAA (Glycotek) was added to the wells. For the inhibition assays, varying concentrations of α -L-fucose (Sigma) were also added simultaneously. Plates were incubated at room temperature for half an hour and then at 4°C for one and a half h. Subsequently, plates were washed and incubated with HRP-conjugated streptavidin at room temperature for 1 h. Then, 50 μ L of TMB-ELISA substrate (Pierce) was added in all the wells. The reaction was stopped by adding 2 M H₂SO₄ and the color developed was monitored by measuring absorbance at 450 nm in Synergy H1 plate reader (Bio-Tek).

Enzyme-linked lectin assays were also performed with a different set-up using immobilized glycans and lectin in solution to more closely mimic the biological presentation of glycans on cell surfaces. For such experiments, neutravidin (Pierce) (100 μ L of 2 μ g/mL stock solution in TBS) was coated in the wells of a MaxiSorp flat-bottom 96 well plate for 12 h at 4°C. Subsequent to washing, blocking was done with 5% BSA overnight at 4°C. The wells were again washed and incubated with biotin-PAA- α -L-fucose or biotin-PAA. Different concentrations of proteins (starting from 1.38 mM with two-fold

serial dilution) were then added to the wells and incubated for 2 h at 4°C. Following washing, the plates were incubated with mouse anti-C-terminal-His antibody (NOVEX Life Sciences) for 1 h at room temperature and then with the anti-mouse secondary antibody (Invitrogen) for 1 h at room temperature. Following washes, 50 µL of TMB-ELISA substrate (Pierce) was added in all the wells, the reaction was stopped by adding 2M H₂SO₄ and absorbance measured at 450 nm in Synergy H1 plate reader (Bio-Tek).

Glycan microarray analysis

Glycan microarray analysis was done by CFG (Consortium for Functional Glycomics, Emory University, USA) using the standard procedure of the Protein–Glycan Interaction Core (H). The purified recombinant proteins, *B*/FLD, *B*/DupFLD, *Sr*FLD and *Sr*DupFLD were prepared in TBS containing 10 mM CaCl₂. The purified proteins (70 µL) at a final concentration of 50 µg/mL were allowed to bind to the surface of the CFG glycan microarrayslide version 5.3 (having 600 synthetic and natural glycans) and incubated at room temperature for one hour. The slide was then washed with TBS containing 10 mM CaCl₂ and 0.05% Tween-20 four times and with TBS containing only 10 mM CaCl₂ four times. The slide was then incubated with 70 µL of mouse anti C-terminal 6xHis antibody diluted in TBS containing 10 mM CaCl₂ at room temperature in a dark humidified chamber for one hour. The slide was again washed as described above and incubated with 70 µL of fluorescently labeled Alexa488 secondary anti-mouse antibody diluted in the same buffer and incubated for one hour at room temperature in the dark chamber. After the incubation, washes were performed as above, followed by four washes with distilled water, the slide was dried and scanned in a Perkin Elmer ScanArray scanner and data was saved for each wavelength used for detection. The saved images were opened in Imagene software and the spots on the slide were aligned by grid using biotin control spots. After the alignment of spots, the intensity of spots was measured. The highest and lowest spot of the six replicates were removed from the data and the remaining four spots were averaged along with appropriate statistics and displayed graphically. For determining the detailed specificity of the proteins, the fucosylated glycans were separated from the 600 glycans on version 5.3 of the CFG glycan microarray and categorized according to L-fucose linkages and glycan structure motif into the following different glycan categories – Fucα, Fucα1-2, blood group H type-1, blood group H type-2, blood group A, blood group B, Lewis^a, Lewis^x, Lewis^b, Lewis^y, Fucα1-3GlcNac, Fucα1-4GlcNac, Fucα1-6 and Fucβ1-3 (Supplementary data, Tables SII, III, VII, VIII). Some complex glycans contained more than one glycan motif, and they were classified into more than one glycan category. The binding of proteins to fucosylated glycans present in different glycan categories was analyzed by sorting each category according to the signal intensity in descending order and displaying graphically.

Surface Plasmon Resonance (SPR)

SPR experiments were performed on a BIAcore 3000 instrument equipped with a research-grade CM5 sensor chip. The surfaces of flow channels (FCs) 1 and 2 were activated with a 1:1 mixture of 0.1 M NHS (*N*-hydroxysuccinimide) and 0.1 M EDC (3-(*N,N*-dimethylamino)propyl-*N*-ethylcarbodiimide) at a flow rate of 5 µL/min. Neutravidin at a concentration of 40 µg/mL in 10 mM sodium acetate, pH 5.0, was immobilized on FC1 and FC2. All the surfaces were blocked with 1 M ethanolamine, pH 8.0. The biotinylated polyacrylamide probe bearing L-fucose was then captured on FC2 at a concentration of 200 µg/mL. FC1 was left blank to serve as a

reference surface. To collect kinetic binding data, the analyte (*Sr*FLD, *Sr*DupFLD, *Sr*Dup_{TM}FLD or *Sr*DupFLD_{TM}) was injected (association 240 s, dissociation 600 s) over the two FCs and binding measured as response units (RU) over time after blank subtraction (FC2–FC1). The FCs were fully regenerated with 1 M L-fucose after each injection. The data were fit to a simple 1:1 interaction model available within BIAevaluation software version 4.

Isothermal Titration Calorimetry (ITC)

ITC experiments were performed at 25°C on a VP-ITC microcalorimeter (Malvern Instruments). The purified proteins, *Sr*FLD, *Sr*DupFLD and *Sr*Dup_{TM}FLD, were dialyzed against buffer containing 20 mM HEPES and 150 mM NaCl, pH 7.5. Ligand solutions were prepared in the dialysate. Binding was studied with the ligands, L-fucose, blood group H antigen trisaccharide type 2 and Lewis^Y antigen tetrasaccharide. To study the interaction with L-fucose, protein concentration used was 0.06 mM and L-fucose concentration used was 3 mM. The protein and ligand concentrations used were 0.2 mM and 3 mM, respectively for H antigen trisaccharide type 2, and 0.04 mM and 0.4 mM, respectively for Lewis^Y antigen tetrasaccharide. The protein was placed in the sample cell (1.4 mL) and titrations were performed with fixed volume of ligand (28 injections of 10 µL each for L-fucose and H-trisaccharide type 2 and 35 injections of 8 µL each for Lewis^Y antigen) every 300 s. The first injection was performed with 3 µL of ligand to minimize the heat change due to any artifact associated with loading the syringe with the ligand. Heat changes due to dilutions were determined by titrating ligands against buffer alone and subtracted from the experimental curves prior to data analysis. Data were fitted to a single site model using MicroCal Origin 7 software and the thermodynamic parameters, *K*, *n* and ΔH were obtained. *K_D* was calculated from the value of *K*, and the parameters, ΔG and ΔS were obtained using the equations, $\Delta G = -RT \cdot \ln K$ and $\Delta S = (\Delta H - \Delta G)/T$.

Supplementary data

Supplementary data is available at *Glycobiology* online.

Funding

This work was partly supported by the Science and Engineering Research Board, Department of Science and Technology, Government of India (FAST-TRACK Grant no. SR/FT/LS-87/2012 to R.T.N.C.). S.M. acknowledges the Department of Biotechnology (DBT), Government of India for her fellowship.

Acknowledgements

The authors thank Dr. S. Karthikeyan, Dr. Krishan Gopal Thakur and Mr. Adarsh Kumar from CSIR-IMTECH for enabling our size exclusion chromatography experiments, and Dr. Yu from Academia Sinica, Taiwan for providing *B. floridae* EST sequences that map to coding sequence of “hypothetical protein BRAFLDRAFT_118409 [*Branchiostoma floridae*]”. The authors acknowledge the Protein–Glycan Interaction Resource of the CFG (supporting grant R24 GM098791) and the National Center for Functional Glycomics (NCFG) at Beth Israel Deaconess Medical Center, Harvard Medical School (supporting grant P41 GM103694) for glycan array analysis. The authors acknowledge CSIR-IMTECH (manuscript communication number – 037/2017) for the research facilities and infrastructure.

Conflict of interest statement

The authors declare that there are no conflicts of interest relevant to the subject of this manuscript.

Authors' contributions

R.T.N.C. conceived the study. S.M. performed all the biochemical experiments and wrote the first draft of the manuscript. S.M. and R.T.N.C. participated in experimental design, data analysis and editing the manuscript.

Abbreviations

CDR, complementarity determining region; CRD, carbohydrate recognition domain; CD, circular dichroism; CDD, conserved domain database; BLAST, basic local alignment search tool; ESI-LC-MS, electrospray ionization-liquid chromatography-mass spectrometry; MALDI-MS, matrix assisted laser desorption/ionization-mass spectrometry; HMM, hidden Markov Model; ELLA, enzyme linked lectin assay; FLD, F-type lectin domain; ITC, isothermal calorimetry; SPR, surface plasmon resonance.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.
- Arnold T, Poynor M, Nussberger S, Lupas AN, Linke D. 2007. Gene duplication of the eight-stranded beta-barrel OmpX produces a functional pore: A scenario for the evolution of transmembrane beta-barrels. *J Mol Biol.* 366:1174–1184.
- Audfray A, Claudinon J, Abounit S, Ruvoen-Clouet N, Larson G, Smith DF, Wimmerova M, Le Pendu J, Romer W, Varrot A et al. 2012. Fucose-binding lectin from opportunistic pathogen *Burkholderia ambifaria* binds to both plant and human oligosaccharidic epitopes. *J Biol Chem.* 287:4335–4347.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
- Bianchet MA, Odom EW, Vasta GR, Amzel LM. 2002. A novel fucose recognition fold involved in innate immunity. *Nat Struct Biol.* 9:628–634.
- Bianchet MA, Odom EW, Vasta GR, Amzel LM. 2010. Structure and specificity of a binary tandem domain F-lectin from striped bass (*Morone saxatilis*). *J Mol Biol.* 401:239–252.
- Bishnoi R, Khatri I, Subramanian S, Ramya TN. 2015. Prevalence of the F-type lectin domain. *Glycobiology.* 25:888–901.
- Bishnoi R, Mahajan S, Ramya TNC. 2018. An F-type lectin domain directs the activity of *Streptosporangium roseum* alpha-L-fucosidase. *Glycobiology.* doi:0.1093/glycob/cwy079.
- Boraston AB, Wang D, Burke RD. 2006. Blood group antigen recognition by a *Streptococcus pneumoniae* virulence factor. *J Biol Chem.* 281:35263–35271.
- Collins BE, Paulson JC. 2004. Cell surface biology mediated by low affinity multivalent protein-glycan interactions. *Curr Opin Chem Biol.* 8:617–625.
- Drickamer K. Animal Lectins. <http://www.imperial.ac.uk/research/animallectins/ctld/classes/Fucolectin1.htm> (20 September 2018, date last accessed).
- Drickamer K, Taylor ME. 1993. Biology of animal lectins. *Annu Rev Cell Biol.* 9:237–264.
- Farrand S, Hotze E, Friese P, Hollingshead SK, Smith DF, Cummings RD, Dale GL, Tweten RK. 2008. Characterization of a streptococcal cholesterol-dependent cytolysin with a Lewis y and b specific lectin domain. *Biochemistry.* 47:7097–7107.
- Feil SC, Lawrence S, Mulhern TD, Holien JK, Hotze EM, Farrand S, Tweten RK, Parker MW. 2012. Structure of the lectin regulatory domain of the cholesterol-dependent cytolysin lectinolysin reveals the basis for its lewis antigen specificity. *Structure.* 20:248–258.
- Fujita T, Matsushita M, Endo Y. 2004. The lectin-complement pathway – its role in innate immunity and evolution. *Immunol Rev.* 198:185–202.
- Gonzalez MW, Pearson WR. 2010. RefProtDom: A protein database with improved domain boundaries and homology relationships. *Bioinformatics.* 26:2361–2362.
- Gupta G, Surolia A. 2012. Glycomics: An overview of the complex glycode. *Adv Exp Med Biol.* 749:1–13.
- Holmskov U, Thiel S, Jensenius JC. 2003. Collections and ficolins: Humoral lectins of the innate immune defense. *Annu Rev Immunol.* 21:547–578.
- Honda S, Kashiwagi M, Miyamoto K, Takei Y, Hirose S. 2000. Multiplicity, structures, and endocrine and exocrine natures of eel fucose-binding lectins. *J Biol Chem.* 275:33151–33157.
- Jensen PH, Weilguny D, Matthiesen F, McGuire KA, Shi L, Hojrup P. 2005. Characterization of the oligomer structure of recombinant human mannan-binding lectin. *J Biol Chem.* 280:11043–11051.
- Jin P, Ji X, Wang H, Li-Ling J, Ma F. 2010. AmphiEST: Enabling comparative analysis of ESTs from five developmental stages of amphioxus. *Mar Genomics.* 3:151–155.
- Kilpatrick DC. 2002. Animal lectins: A historical introduction and overview. *Biochim Biophys Acta.* 1572(2–3):187–197.
- Lis H, Sharon N. 1998. Lectins: Carbohydrate-specific proteins that mediate cellular recognition. *Chem Rev.* 98:637–674.
- Marchler-Bauer A, Bryant SH. 2004. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* 32:W327–W331.
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR et al. 2011. CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 39:D225–D229.
- McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R. 2013. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.* 41:W597–W600.
- Menichelli C, Gascuel O, Brehelin L. 2018. Improving pairwise comparison of protein sequences with domain co-occurrence. *PLoS Comput Biol.* 14: e1005889.
- Monsigny M, Mayer R, Roche AC. 2000. Sugar-lectin interactions: Sugar clusters, lectin multivalency and avidity. *Carbohydr Lett.* 4:35–52.
- Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 302:205–217.
- Odom EW, Vasta GR. 2006. Characterization of a binary tandem domain F-type lectin from striped bass (*Morone saxatilis*). *J Biol Chem.* 281:1698–1713.
- Oulion S, Bertrand S, Belgacem MR, Le Petillon Y, Escriva H. 2012. Sequencing and analysis of the Mediterranean amphioxus (*Branchiostoma lanceolatum*) transcriptome. *PLoS One.* 7:e36554.
- Pascarella S, Argos P. 1992. Analysis of insertions/deletions in protein structures. *J Mol Biol.* 224:461–471.
- Prakash A, Bateman A. 2015. Domain atrophy creates rare cases of functional partial protein domains. *Genome Biol.* 16:88.
- Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature.* 453:1064–1071.
- Quiocho FA. 1986. Carbohydrate-binding proteins: Tertiary structures and protein-sugar interactions. *Annu Rev Biochem.* 55:287–315.
- Saito T, Hatada M, Iwanaga S, Kawabata S. 1997. A newly identified horse-shoe crab lectin with binding specificity to O-antigen of bacterial lipopolysaccharides. *J Biol Chem.* 272:30703–30708.
- Sandhya S, Rani SS, Pankaj B, Govind MK, Offmann B, Srinivasan N, Sowdhamini R. 2009. Length variations amongst protein domain superfamilies and consequences on structure and function. *PLoS One.* 4: e4981.
- Sharon N, Lis H. 2004. History of lectins: From hemagglutinins to biological recognition molecules. *Glycobiology.* 14:53R–62R.
- Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Res.* 14:555–566.
- Triant DA, Pearson WR. 2015. Most partial domains in proteins are alignment and annotation artifacts. *Genome Biol.* 16:99.
- Turner MW. 1996. Mannose-binding lectin: The pluripotent molecule of the innate immune system. *Immunol Today.* 17:532–540.
- Vasta GR, Ahmed H. 2009. Animal Lectins – A functional view.
- Vasta GR, Ahmed H, Odom EW. 2004. Structural and functional diversity of lectin repertoires in invertebrates, protochordates and ectothermic vertebrates. *Curr Opin Struct Biol.* 14:617–630.
- Vasta GR, Ahmed H, Tasumi S, Odom EW, Saito K. 2007. Biological roles of lectins in innate immunity: Molecular and structural basis for diversity in self/non-self recognition. *Adv Exp Med Biol.* 598:389–406.

- Vasta GR, Amzel LM, Bianchet MA, Cammarata M, Feng C, Saito K. 2017. F-type lectins: A highly diversified family of fucose-binding proteins with a unique sequence motif and structural fold, involved in self/non-self-recognition. *Front Immunol.* 8:1648.
- Vasta GR, Odom EW, Bianchet MA, Amzel LM, Saito K, Ahmed H. 2008. F-type lectins: A new family of recognition factors. In: Vasta GR, Ahmed H, editors. *Animal Lectins: A Functional View*. Boca Raton: CRC Press.
- Weis WI, Drickamer K. 1996. Structural basis of lectin-carbohydrate recognition. *Annu Rev Biochem.* 65:441–473.
- Yang KY, Chen Y, Zhang Z, Ng PK, Zhou WJ, Zhang Y, Liu M, Chen J, Mao B, Tsui SK. 2016. Transcriptome analysis of different developmental stages of amphioxus reveals dynamic changes of distinct classes of genes during development. *Sci Rep.* 6:23195.
- Yu JK, Wang MC, Shin IT, Kohara Y, Holland LZ, Satoh N, Satou Y. 2008. A cDNA resource for the cephalochordate amphioxus *Branchiostoma floridae*. *Dev Genes Evol.* 218:723–727.
- Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG et al. 2018. Ensembl 2018. *Nucleic Acids Res.* 46:D754–D761.