Computational Biology

Prevalence of the F-type lectin domain

Ritika Bishnoi2,†, Indu Khatri2,†, Srikrishna Subramanian2, and T N C Ramya1,2

2Institute 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; ramya.tnc@gmail.com
†These authors contributed equally to this work and should be regarded as co-first authors.

Received 2 March 2015; Revised 29 April 2015; Accepted 29 April 2015

Abstract

F-type lectins are fucolectins with characteristic fucose and calcium-binding sequence motifs and a unique lectin fold (the “F-type” fold). F-type lectins are phylogenetically widespread with selective distribution. Several eukaryotic F-type lectins have been biochemically and structurally characterized, and the F-type lectin domain (FLD) has also been studied in the bacterial proteins, Streptococcus mitis lectinolysin and Streptococcus pneumoniae SP2159. However, there is little knowledge about the extent of occurrence of FLDs and their domain organization, especially, in bacteria. We have now mined the extensive genomic sequence information available in the public databases with sensitive sequence search techniques in order to exhaustively survey prokaryotic and eukaryotic FLDs. We report 437 FLD sequence clusters (clustered at 80% sequence identity) from eukaryotic, eubacterial and viral proteins. Domain architectures are diverse but mostly conserved in closely related organisms, and domain organizations of bacterial FLD-containing proteins are very different from their eukaryotic counterparts, suggesting unique specialization of FLDs to suit different requirements. Several atypical phylogenetic associations hint at lateral transfer. Among eukaryotes, we observe an expansion of FLDs in terms of occurrence and domain organization diversity in the taxa Mollusca, Hemichordata and Branchiostomi, perhaps coinciding with greater emphasis on innate immune strategies in these organisms. The naturally occurring FLDs with diverse domain organizations that we have identified here will be useful for future studies aimed at creating designer molecular platforms for directing desired biological activities to fucosylated glycoconjugates in target niches.

Key words: domain organization, F-type lectin domain, fucose, lateral transfer, phylogenetic distribution

Introduction

The F-type lectin domain (FLD) is an ~140 residue domain with conserved 1-fucose-binding [HX(26)RXDX(4)R/K] and calcium-binding sequence motifs (where X is any residue, h is a small hydrophobic residue [Val, Ala, Ile] and x is a small hydrophilic residue [Asn, Asp, Ser]) (Vasta et al. 2008). FLDs are known to occur in proteins of numerous life forms. Among eukaryotes, FLDs have been reported in proteins from invertebrates (arthropods, echinoderms, molluscs and platyhelminthes) and vertebrates (lobe and ray finned teleosts, cartilaginous fish and amphibians) but have been thought to be absent in protozoans, nematodes, ascidians, higher vertebrates (reptiles, birds and mammals) (Vasta et al. 2004), fungi and plants (an exception being Bryohealin from the marine green alga, Bryopsis plumosa) (Kim et al. 2006; Yoon et al. 2008). Even in the former taxa where FLDs have been reported, the distribution has been reported to be selective, suggesting that this domain might have been lost even among closely related lineages (Vasta et al. 2004). Among prokaryotes, FLDs have been reported in both Gram-negative (i.e. Microbulifera degradans) and Gram-positive (i.e. Streptococcus pneumoniae, Streptococcus mitis) bacteria (Vasta et al. 2004; Boraston et al. 2006; Farrand et al. 2008), and following the discovery of Streptococcus pneumoniae SP2159, which has three FLDs as well as a carbohydrate-active
enzyme domain, FLDs have also been classified in the Carbohydrate-Active enZymes (CAZy) database (Cantarel et al. 2009) as the carbohydrate-binding module (CBM) family, CBM47 (http://www.cazy.org/CBM47).

FLDs from fish and amphibians have been implicated in innate immune response (Honda et al. 2000; Cammarata et al. 2001, 2007; Odom and Vasta 2006; Fleming et al. 2009; Parisi et al. 2010). The FLD-containing protein, Bryohealin from the marine green alga, Bryopsis plumosa has a potential role in protoplast regeneration (Kim et al. 2006; Yoon et al. 2008). The FLD-containing bacterial proteins, SP2159 and lectinolysin have been suggested to have a role in virulence (Boraston et al. 2006; Randall et al. 2008).

Four FLD structures (1K12, 2J1S, 3LEI and 3CQO) are known in the Protein Data Bank (PDB), and classified in the Structural Classification of Proteins (SCOP) database under the family “fucose-binding lectins” that is part of the only superfamily (galactose-binding domain like) of the fold “galactose-binding domain like”. There are 32 families that are classified under the galactose-binding-domain-like superfamily in SCOP, including various CBM domains, a galactose-binding domain and a discoidin domain. It is worth mentioning here that FLDs are classified in the Pfam database (Finn et al. 2014) together with these discoidin domain and galactose-binding domain members under the coagulation factor 5/8 C-terminal (FA58C/discoidin (PF00754) family. In the Simple Modular Architecture Research Tool (SMART) database (Letunic et al. 2004), eukaryotic (but not prokaryotic) FLDs are classified under SMART accession number, SM00607 [the eel-Fucolentin Tachylent-4 Pentraxin-1 domain (FTP)].

Vasta et al. have, in the past decade, meticulously analyzed the structural and functional diversity of FLDs (Vasta et al. 2004, 2008; Odom and Vasta 2006). With the advent of next generation sequencing technologies and the whole-genome sequencing of a multitude of organisms, many more FLD-like sequences have since been deposited in the online databases. We felt that a current and in-depth survey of this family would be useful to appreciate the functions and biological contexts of this domain, especially where the FLD is present as part of chimeric proteins. Given the possibility of targeting biological functions through this domain, such molecular information could also provide indications for well-targeted biochemical studies of FLD-containing proteins. In the present study, we have employed sensitive sequence search tools to generate a comprehensive inventory of prokaryotic and eukaryotic FLDs and to examine their phylogenetic distribution, sequence characteristics and domain organization. We identify 437 FLD sequence clusters (clustered at 80% sequence identity) with FLDs occurring in proteins of viruses, eubacteria and eukaryotes. Remarkably, these FLDs occur in diverse protein domain architectures, suggesting that FLDs may be involved in targeting various biological functions to distinct fucosylated niches.

Results

Definition of the FLD

The three-dimensional structure of FLD, first described in l-fucose bound Anguilla anguilla agglutinin (AAA) (PDB code 1K12), consists of an eight-stranded β-barrel with jelly roll topology, comprising two β-sheets of three (β5, β8 and β11) and five (β2, β3, β10, β6 and β7) antiparallel β-strands (Blanquet et al. 2002). The fucose-binding site is a shallow positively charged pocket at one end of the β-barrel, which is framed by five loops, termed the complementarity determining regions (CDR1 through CDR5) that arise from the two main β-sheets.

Two short antiparallel strands (β4 and β9) close the other end of the β-barrel. Binding to fucose is mediated by hydrogen bonding with basic residues and Van der Waals contacts with hydrophobic residues in the fucose-binding pocket. The amino acid residues critical for recognition of the axial 4-OH group of α-L-fucose are a His52/Arg79/Arg86 triad which is part of the sequence motif characteristic of FLDs [HX(26)RXDX(4)R/K] (Vasta et al. 2008). For the purpose of this study, we considered the sequence motif [HX(26)RXDX(4)R/K] in the structural context described for the SCOP family “fucose-binding lectins” as characteristic of a typical FLD.

We manually aligned the sequences of the four known FLD structures (PDB codes: 1K12, 2J1S, 3LEI and 3CQO) guided by structure superimposition using PyMOL in order to define domain boundaries with respect to the FLD sequence motif [HX(26)RXDX(4)R/K]. The aligned domain was ~160 residues long, with ~50 residues before and ~75 residues after the sequence motif, which itself is 35 residues long (Figure 1). These domain boundaries were used for extracting FLDs from FLD sequence motif containing polypeptides in our study. We mapped all new FLD sequences identified in our study to SCOP domain database using fold and function assignment (FFAS) (Jaroszewski et al. 2005) in order to verify structural similarity to FLD fold and we extended the domain boundaries at both ends if they aligned to AAA (SCOP entry d1k12a).

Retrieval of FLD sequences from publically available databases

We used several approaches in order to retrieve FLD sequences. First, we searched the entries in the Pfam database classified under the FA58C/discoidin family (PF00754) for the characteristic fucose-binding motif [HX(26)RXDX(4)R/K] of FLDs. Second, we subjected each of these hits to Position-Specific Iterative Basic Local Search Alignment Tool (PSI-BLAST) (Altschul et al. 1997) against the non-redundant (NR) database in order to detect FLDs that are yet to be annotated in Pfam, as well as to detect FLD homologs that may have been missed due to variations in the fucose-binding motif, but may nevertheless be capable of lectin activity. Third, we directly searched the entire NR database for the fucose-binding motif [HX (26)RXDX(4)R/K] of FLDs and then subjected the search output to a hidden Markov model scan (HMMer) (Eddy 1998, 2011) against HMM profiles generated from Pfam FLD sequences in order to remove non-FLD hits that have the sequence motif per chance. Sequence hits obtained by these search strategies were clustered at 80% sequence identity in order to maximize on computing power while retaining sequence diversity.

Retrieval of FLDs from Pfam database using Pf_scan

Pfam database classifies FLD sequences under FA58C/discoidin family of proteins (PF00754). The FA58C/discoidin family of Pfam database 27.0 (downloaded on 3 April, 2014) contained 7695 protein sequences (comprising 10,300 domains) of which 3045 belonged to eukaryotes, 4622 to prokaryotes, 23 to viruses and 2 to archaea, and 3 were unclassified sequences. We used Pf_scan (de Castro et al. 2006) and the conserved fucose-binding motif [HX(26)RXDX(4)R/K] to retrieve FLDs from these 7695 protein sequences. We obtained 448 protein sequences following Pf_scan and clustered them at 80% sequence identity using BlastClust (Altschul et al. 1990) in order to generate representative sequences (see Materials and Methods) for further searches. We then clipped the 161 protein sequence clusters so obtained with a custom Perl script in order to retrieve FLD sequence
In this list of 2212 hits, some protein sequences were represented multiple times. We obtained sequences instead of complete protein sequences in order to avoid against the NR database in order to detect homologs. We used FLD clusters since FLDs are often present as tandem repeats or associated with other domains within polypeptides. Following this step, we obtained 224 FLD sequence clusters — 28 prokaryotic, 194 eukaryotic and 2 viral, contained within 17 prokaryotic, 148 eukaryotic and 2 viral protein sequence clusters, respectively. We further clustered the 224 FLD sequence clusters using BlastClust at 80% sequence identity for the sake of consistency in using representative sequences since tandemly repeated FLDs within a polypeptide could be >80% identical. Following this, we obtained 15 prokaryotic (within 14 protein sequences), 173 eukaryotic (within 111 protein sequences) and 2 viral (within 2 protein sequences) FLD sequence clusters. Obsolete entries and partial sequences were removed. Following this, we obtained 15 prokaryotic (within 14 protein sequences), 79 eukaryotic (within 61 protein sequences) and 2 viral (within 2 protein sequences) FLD sequence clusters.

**Retrieval of FLDs from NR database using PSI-BLAST**

We subjected each of the FLD sequence clusters obtained by the Pfam search (mentioned above) to PSI-BLAST (iterations = 5, E-value 1e-5) against the NR database in order to detect homologs. We used FLD sequences instead of complete protein sequences in order to avoid hits from other domains present in the polypeptide. We obtained 2212 hits. Some protein sequences were represented multiple times in this list of 2212 hits, so we first removed redundant sequences and then scanned the remaining protein sequences for the presence of the typical fucose-binding motif [HX(26)RXDX(4)R/K]. We found that 884 protein sequences had this motif, while 1032 protein sequences did not. We clipped FLDs of the 884 protein sequences with the sequence motif using a custom Perl script and clustered the resulting domain sequences using BlastClust at 80% identity into 163 FLD sequence clusters (122 eukaryotic, 2 viral and 39 prokaryotic). All the sequences obtained from the earlier Pfam search were also present in these 163 FLD sequence clusters. In addition, we obtained 24 prokaryotic FLD sequences (within 22 protein sequences) and 43 eukaryotic FLD sequences (within 33 protein sequences) that were not identified in the earlier Pfam search, probably because Pfam had not been updated to include these sequences from the NR database. No additional viral FLD sequences were obtained. We found that 7 of the 22 prokaryotic protein sequences were obsolete in NR. We removed these from further analysis to obtain a total of 17 new prokaryotic FLD sequences (within 15 protein sequences) with the fucose-binding motif [HX(26)RXDX(4)R/K] from the PSI-BLAST approach. Obsolete eukaryotic entries were also removed, and in total, 108 eukaryotic FLD sequences (within 88 protein sequences) were obtained that contained the conserved fucose-binding motif [HX(26)RXDX(4)R/K]. However, 32 of these 88 eukaryotic protein sequences had multiple FLDs in tandem repeats of 2 or more, and so these FLDs were also extracted.

We performed domain annotation of the 1032 sequences without the typical fucose-binding motif [HX(26)RXDX(4)R/K] that were obtained following PSI-BLAST by querying Conserved Domain Database (CDD) (Marchler-Bauer et al. 2011). Only sequences that matched FTP in CDD were considered further. We obtained 61 prokaryotic FLDs (within 61 sequences), 195 eukaryotic FLDs (within 140 sequences) and 6 viral FLDs (within 6 protein sequences, later clustered to 4 viral FLD sequence clusters by BlastClust at 80% sequence identity) from this set.
Thus, using PSI-BLAST searches allowed us to retrieve not only similar sequences with the exact fucose-binding motif \([HX(26)RXDX(4)R/K]\) but also sequences with amino acid substitutions of conserved amino acids in the searched motif and variations in motif length, as will be discussed later in more detail. Sequences without the exact motif also shared significant overall sequence similarity in multiple sequence alignments and were validated to be FLDs by CDD and FFAS analysis.

Retrieval of FLDs from NR database using Pf_scan and HMMer
We performed Pf_scan search using the conserved fucose-binding motif \([HX(26)RXDX(4)R/K]\) on the NR database (downloaded on 04 September 2013), and obtained 56,188 sequences. We clipped 56,707 probable domains out of these sequences using a custom Perl script. Using HMMer, we then scanned these 56,707 sequences against an HMM profile generated from Pfam FLD sequences (see Materials and Methods), and obtained 105 hits. This included three new prokaryotic FLDs (within three protein sequences) not obtained by the previous Pfam search, but no new viral or eukaryotic FLDs not already identified in Pfam searches.

FLDs identified by searches
Following the searches, all FLD sequences identified were validated using FFAS server as mentioned above, and domains with incomplete sequence information as well as obsolete entries were removed (Supplementary data, Table SI). The validated FLD sequences shared sequence identity ranging from 5.06 to 100% (8.04–100% among eukaryotic sequences, 9.35–99.59% among prokaryotic sequences and 15.45–100% among viral sequences) (Supplementary data, Table SII). On the whole, we obtained 95 prokaryotic FLD sequence clusters (15 with the fucose-binding sequence motif \([HX(26)RXDX(4)R/K]\) from the Pf_scan search on Pfam database, 31 with and 61 without the exact motif, \([HX(26)RXDX(4)R/K]\), from the Psi-BLAST search on NR database, and 34 with the motif \([HX(26)RXDX(4)R/K]\) from the HMMer search on NR database) (Figure 1; Supplementary data, Table SIII). Among eukaryotes and viruses, we obtained a total of 342 FLD sequence clusters (109 from the Pf_scan search on Pfam database, 342 from the Psi-BLAST search on NR database and 155 from the HMMer search on NR database) (Figure 1; Supplementary data, Table SIV) and 6 viral FLD sequence clusters (2 from the Pf_scan search on Pfam database, 4 from the PSI-BLAST search on NR database and 2 from the HMMer search on NR database) (Supplementary data, Table SV).

Significantly, using our three-pronged strategy, we were able to retrieve many FLDs, hitherto not annotated as such in the CAZy or SMART databases (Figure 1).

On the whole, we obtained 437 FLD sequence clusters. The clusters varied in size, ranging from 1 to 256 sequences per cluster (Supplementary data, Figure S1). Since there is a possibility of sequences from diverse organisms being clustered together in an FLD sequence cluster by BlastClust, we manually analyzed all clusters for such occurrences. We observed that most FLD sequence clusters comprised proteins containing FLDs from the same or different strains of the same species. The exceptions were 12 prokaryotic and 7 eukaryotic FLD clusters containing FLDs from different species of the same genus, 1 prokaryotic and 2 eukaryotic clusters containing FLDs from different organisms of the same class and 1 eukaryotic FLD cluster containing FLDs from different organisms of the same subphylum (Supplementary data, Table SVI). Some eukaryotic FLD clusters included FLDs belonging to multiple proteins (with different numbers of FLD repeats) from the same organism.

Phylogenetic distribution of FLDs
We found FLDs in prokaryotic and eukaryotic cellular organisms as well as in the viruses, *Emiliana huxleyi* virus 203, *Ostreococcus tauri* virus 1, *Ostreococcus tauri* virus 2, *Ostreococcus virus OsV5*, *Ostreococcus lucimarinus* virus Olv1, *Ostreococcus lucimarinus* virus Olv3 and *Ostreococcus lucimarinus* virus Olv6, a family of dsDNA viruses with no RNA stage whose hosts are fresh water and marine algae and which are classified under Phycodnaviridae (Supplementary data, Tables SV and SIV). Among prokaryotes, we found FLD sequences in organisms from Eubacteria but not from Archaea. Within Eubacteria, we identified FLDs in organisms belonging to the phyla Acidobacteria (class Solibacteres), Actinobacteria (class Actinobacteria), Bacteroidetes (class Flavobacteria), Cyanobacteria (orders Oscillatoriales and Chroococcales of subclass Oscillatoriophycideae and orphan order Pleurocapsales), Firmicutes (class Bacilli), Planctomycetes (class Planctomycetia), Proteobacteria (classes Alphaproteobacteria, Gammaproteobacteria, Delta proteobacteria and Epipelon proteobacteria) and Verrucomicrobia (class Verrucomicrobiae) (Supplementary data, Table SIII). There was also one FLD identified from an uncultured bacterium. We could not identify any FLDs in the phyla, Aquifacia, Armatimonadetes, Chlorobi, Ignavibacteria, Caldiserica, Chlamydiae, Lentisphaerae, Chloroflexi, Chrysiogenetes, Deferribacteria, Deinococcus-Thermus, Dictyogloini, Eusimicrobi, Fibrobacteres, Marinimicrobi, Fusobacteria, Gemmatimonadetes, Nitrospirae, Nitrosopire, Spirochaetes, Synergistetes, Tenericutes, Thermodesulfobacteria and Thermotogae, although assembled genome sequence data are available for all eubacterial phyla, and all phyla other than Nitrosopire, Lentisphaerae and Marinimicrobi are represented by at least one gapless, complete genome sequence deposited in the publically available National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/).

Among eukaryotes, we could identify FLDs in a coccolithophore (orphan order Isochrysidales), diatoms (class Coscinodiscophycaceae of phylum Bacillariophyta), a cryptomonad (orphan class Cryptophyta), a brown alga (order Ectocarpales of phylum Phaeophyceae), green algae (classes Mamiellophyceae, Trebouxiophycaceae, Chlorophyceae and Ulvophyceae of phylum Chlorophyta), fungi (*Phytophthora sojae* and *Phytophthora infestans* T30-4 belonging to order Peronosporales) and metazoans (Supplementary data, Table SIV). Metazoans with FLDs belonged to phyla Cnidaria (classes Anthozoa and Hydrozoa), Arthropoda (classes Arachnida, Branchiopoda, Malacostraca, Merostomata and Insecta), Annelida (class Polychaeta), Mollusca (class Bivalvia), Hemichordata (class Enteropneusta) and Chordata (Supplementary data, Table SIV). Among chordates, we could identify FLDs in *Branchiostoma floridae* (a cephalochordate belonging to order Branchiostomidae), *Callorhinchus milii* (a cartilaginous fish of class Chondrichthyes), many bony fishes, a ray finned fish, *Larimichthys sp.* (belonging to class Actinopteri), a coelacanth fish, *Latimeria chalumnae* (belonging to order Coelacanthiformes), frogs (belonging to class Amphibia), reptiles (belonging to families Iguanidae, Cheloniidae and Crocodyliae), amphibians (belonging to order Anura, Reptilia and birds, respectively) (Supplementary data, Table SIV). Interestingly, although FLDs were previously thought to be absent in fungi and also restricted to invertebrates and cold-blooded
vertebrates, our searches identified FLDs in fungi, reptiles, birds and primitive mammals, too. We did not find any FLDs in protozoans, nematodes, echinoderms, ascidians, eutherian mammals, or higher plants. As previously observed by Vasta et al. (2008), the distribution of FLDs was very selective (Supplementary data, Figure S2). We examined the possibility that this was due to the availability of completely assembled genomic data for only certain organisms. However, even within many of the aforementioned bacterial taxa where we identified FLDs, we perceived that there were bacterial species with publically available, complete genome data which did not have FLDs, indicating loss of FLDs within even closely related lineages (Supplementary data, Table SVII). For instance, we found FLDs in Campylobacter coli, in Campylobacter curvus, Campylobacter fetus or Campylobacter concisus, all of whose genomes are completely sequenced. Similarly, while whole-genome sequences are available for many Geobacter species, i.e. Geobacter lovleyi SZ, Geobacter sulfurreducens, Geobacter bemidjiensis, Geobacter uranireducens, Geobacter metallireducens, Geobacter daltonii, Geobacter sp. M18 and Geobacter sp. M21, we could find an FLD only in Geobacter lovleyi SZ (Supplementary data, Table SVII). Moreover, we failed to detect not only FLDs but also non-FLD FA58C domains in the genomes of the other Geobacter sp. using sensitive search techniques (BLAST, PSI-BLAST and HMMer).

Ecological niches of organisms with FLDs

Our search results indicated that organisms with wide-ranging ecological niches harbor FLDs. Although FLDs were previously thought to have been lost in higher vertebrates with the appearance of cleidoic egg and colonization of land (Odom and Vasta 2006), we detected FLDs in reptiles, aves and primitive mammals, too. Among bacteria, FLDs are present in free living marine and freshwater bacteria, in soil bacteria, in a bacterium residing in glacial ice, in commensal and pathogenic bacteria of human beings and other animals, in plant endophytes, in bacteria growing on sugar-rich sites such as fermentation media and ripe fruits, and in algae-associated bacteria (Supplementary data, Figure S3). It is possible that, like SP2159 of Streptococcus pneumoniae, the FLDs of pathogenic and commensal bacteria recognize fucosylated glycans of host tissues, facilitating colonization and/or virulence. FLDs might also provide for association with algal surfaces through recognition of fucosidans or other glycoconjugates on the algae.

Domain architecture of FLD-containing proteins

We analyzed the FLD sequence clusters for the presence of co-occurring domains using CDD. We observed that bacterial FLDs generally presented as a single copy, with the exception of three sequence clusters [which included sequences from Streptococcus mitis, Streptococcus pneumoniae and Myxococcus sp. (contaminant ex sp. DSM436), respectively] where tandem repeats are present (Figure 2). We also found a few bacterial FLDs that were associated with non-FLD FA58C domains and some with no annotated co-occurring domain, although the sequence length of most of the latter proteins hints at the presence of additional, currently unannotated domains (Figure 2). On the whole, however, we found the bacterial FLD existing in tandem with one or more different domains, where the FLD was positioned in the N- or C-terminus or sandwiched between other domains (Figures 2 and 3). The co-occurring domains included carbohydrate-binding domains—Ricin-B lectin, CBM6 Agarase, CBM6 Cellulase, CBM6 Xylanase and NPCBM (novel putative CBM), and carbohydrate-active enzyme domains—α-1-fucosidase, glycosylhydrodase (GH) family domains—GH16, GH20 and GH98, alginase lyase and β-N-acetylgalcosaminidase, and glycosyltransferase (GT) family domains—GT2/GT92, GT17 and fucosyltransferase and LicD (lipopolysaccharide cholinephosphotransferase family of proteins which includes proteins involved in glycan modification). We also observed enzyme domains such as lipase, methyltransferases 21 and 23, thiol cytolysin and membrane bound dehydrogenase and various other domains and repeats associated with FLDs. We found FLDs in tandem with domains encoding carbohydrate-binding activity (lectins and CBMs) in the phyla Actinobacteria, Bacteroidetes and Planctomycetes, domains encoding hydrolytic activity (GH, peptidase and lipase) in the phyla Actinobacteria, Bacteroidetes, Planctomycetes, Firmicutes and Verrucomicrobia and domains encoding transferase activity (glycosyltransferase and methyltransferase) in the phyla Proteobacteria and Cyanobacteria (Figure 2).

On the whole, we noticed that diverse domain architectures exist in different groups of bacteria and domain architectures are conserved in related organisms, suggesting that FLDs have evolved to suit specific requirements (Figure 3). Bacteria from classes Verrucomicrobiae (phylum Verrucomicrobia) and Planctomycetaceae (phylum Planctomycetes) which belong to the same superphylum, Planctomycetes-Verrucomicrobiae-Chlamydiae (PVC), have FLDs with high sequence similarity in similar arrangements (Figure 3). However, taxonomically distant members of phylum Cyanobacteria and class Alphaproteobacteria of phylum Proteobacteria, and Zobellia galactanivorans (phylum Bacteriodes, class Flavobacteria) and Rhodospirillia baltica SH28 (phylum Planctomycetes, class Planctomycetaceae) also have proteins where FLDs with high sequence similarity co-occur with methyltransferase domains and with α-1-fucosidase domains, respectively, perhaps suggestive of recent lateral transfers (Figure 3). We also observed the converse scenario in some proteins (belonging to the same organism, to closely related organisms or to taxonomically distant organisms) that have FLDs with high sequence similarity but different domain architectures, probably resulting from gene duplication and/or transfer of a portion of the polypeptide that includes the FLD into different gene neighborhoods. Instances of this include proteins from Flavobacteriaceae bacterium ALC-1 and Streptomyces violaceusniger Tu 4113, proteins from Saccharomonospora cyanea NA-134 and Candidatus Solibacter usitatus Ellin6076, proteins from Pirellula staleyi DSM6068 and Cellulophaga algicola DSM 14237, proteins from Pirellula staleyi DSM6068 and Singulisphaera acidiphila DSM18658, proteins from Planctomyces brasiliensis DSM 5305 and Gemmatobacter urinaeformans, proteins from Arthrobacter platensis and Xenococcus sp. PCC 730, proteins from Arthrobacter platensis and Methylobacterium mesophilicum, proteins from Cyanothece sp. PCC7001 and Methylobacterium extorquens DM4 and proteins from Pseudolalteromonas atlantica T6c and Gluconobacter moribifer G707 (Figure 3).

In dsDNA viruses, we noticed that FLDs were present singly (in Ostreococcus viruses) or associated with a pentraxin domain (in Emiliania huxleyi virus 203), and the domain architectures were different from those in their eukaryotic algal hosts [Emiliania huxleyi FLD found in tandem with an MAM (meprin, A5 protein and protein tyrosine phosphatase Mu) domain, Ostreococcus lucimarinus FLD found in tandem with a TECPR (β propeller repeats in Physarum polycel- lum teconins) domain] (Table 1; Supplementary data, Table SV and Figure S4). Though the FLDs from the algal hosts, Emiliania huxleyi and Ostreococcus lucimarinus share high-sequence similarity, the Ostreococcus virus FLDs and the Emiliania huxleyi virus 203 FLD do not share as much sequence similarity, perhaps indicating greater
divergence by virtue of higher mutation rates. Nevertheless, all virus
and host FLDs are found in the same region of the phylogenetic
tree, and the virus FLDs are closest to FLDs from the diatom,
Thalassiosira oceanica, the green algae Volvox carteri f. nagariensis
and Chlorella variabilis, the bivalve molluscs—Crassostrea gigas,
Crassostrea nippona, Crassostrea sikamea, Crassostrea angulata and

Fig. 2. Taxon-wise distribution of domains that co-occur with FLDs in prokaryotes. Domain abbreviations are according to CDD. The taxa appearing in the figure are classes of Eubacteria. A single FLD identified from an uncultured bacterium is not included in this figure. (*No class designation is available. Oscillatoriophyceae is a subclass and Pleurocapsales is an order; both belong to Cyanobacterial. This figure is available in black and white in print and in color at Glycobiology online.

Fig. 3. Phylogenetic tree of prokaryotic FLDs depicting domain organization of FLD-containing polypeptides. Branches are colored according to taxonomic status (as indicated in the legend). This figure is available in black and white in print and in color at Glycobiology online.
Table I. Taxon-wise distribution of domains that co-occur with FLDs in eukaryotes and viruses

<table>
<thead>
<tr>
<th>Taxon*</th>
<th>Co-occurring domain/sb</th>
</tr>
</thead>
<tbody>
<tr>
<td>dsDNA viruses-Phycodnaviridae*</td>
<td>None [5], PTX [1]</td>
</tr>
<tr>
<td>Haptoptalgeae-biochrysidae*</td>
<td>MAM [1]</td>
</tr>
<tr>
<td>Coscinodiscophyceae</td>
<td>Abf[B] [1], DUF1501[1], DUF1800 [3], FA58C [1], FLD [9], GH_16 [1], None [1], PT [2]</td>
</tr>
<tr>
<td>Cryptophyta</td>
<td>None [1]</td>
</tr>
<tr>
<td>Phaeophyceae-Ectocarpales*</td>
<td>RCC1 [4]</td>
</tr>
<tr>
<td>Mamiellophyceae</td>
<td>TECPR [1]</td>
</tr>
<tr>
<td>Trebouxiophyceae</td>
<td>Gal_Lectin [1]</td>
</tr>
<tr>
<td>Chlorophyceae</td>
<td>FLD [1], None [1], PAN_APPLE [3]</td>
</tr>
<tr>
<td>Ulvophyceae</td>
<td>None [1]</td>
</tr>
<tr>
<td>Oomycetes-Peronosporales*</td>
<td>Omph [1], RCC1 [7]</td>
</tr>
<tr>
<td>Anthozoa</td>
<td>FA58C [1], None [1]</td>
</tr>
<tr>
<td>Hydrozoa</td>
<td>RICIN [1], EGF_CA [3], CCP [1], HYR [1], GCC2_GCC3 [1]</td>
</tr>
<tr>
<td>Arachnida</td>
<td>CLECT [1], Commd8 [1], CUB [2], FLD [1], None [2]</td>
</tr>
<tr>
<td>Branchiopoda</td>
<td>CCP [10], CLECT [1]</td>
</tr>
<tr>
<td>Malacostraca</td>
<td>None [1]</td>
</tr>
<tr>
<td>Insecta</td>
<td>CCP [125], CLECT [17]</td>
</tr>
<tr>
<td>Merostomata</td>
<td>None [1]</td>
</tr>
<tr>
<td>Polychaeta</td>
<td>CLECT [1], None [1]</td>
</tr>
<tr>
<td>Bivalvia</td>
<td>CCP [2], DUF1768 [1], EGF_CA [1], FLD [4], None [9], PAN_APPLE [3]</td>
</tr>
<tr>
<td>Enteropneusta</td>
<td>ASC [2], DUF3246 [1], FA58C [1], FLD [20], LDLa [3], PAN_APPLE [1], PTX [36]</td>
</tr>
<tr>
<td>Cephalochordata-Branchiostomidae*</td>
<td>7tm_2 [1], An_peroxidase-like [3], ARM [1], ArsB_NhaD_permease [1], CTP [37], Collagen [2], CPW_WPC [1], CRD_F [4], CUB [7], DUF294 [1], DUF4600 [1], EGF_CA [10], FA58C [29], FLD [54], Fz [1], Fz-like [2], Gb [1], Gal_Lectin [7], GLTP [1], GMP_PDE_delta [1], GON [4], GPS [2], GT1_GTP_like [1], Hyd_WA [1], Ig_like [1], Ig_like [1], KR [14], LDLa [2], LNK [2], LRR_RI [3], MACPF [4], Methyltransfer_FA [2], MUC [3], Mucin2_WxxW [1], Neur_chan_memb [1], NHL [2], None [7], PAN_APPLE [7], Pander_like [2], PLAT_polycystin [1], PTX [10], Sox_C_T [1], SR [20], Sulfotransfer_2 [1], Sulfotransfer_3 [1], TECPR [2], TM2 [1], TPR [4], Trefoil [1], TSP1 [14], WSC [13], XLINK [1], Ygr210_GTpase [1]</td>
</tr>
<tr>
<td>Chondrichthyes</td>
<td>None [1]</td>
</tr>
<tr>
<td>Actinopterii</td>
<td>CLECT [5], FLD [40], None [8]</td>
</tr>
<tr>
<td>Coelacanthomorpha-Coelacanthiformes*</td>
<td>FA58C [1], FLD [2], None [1]</td>
</tr>
<tr>
<td>Amphibia</td>
<td>FA58C [1], FLD [26], None [8], PTX [2]</td>
</tr>
<tr>
<td>Sauria-Lepidosauria*</td>
<td>None [1]</td>
</tr>
<tr>
<td>Sauria-Testudines*</td>
<td>None [1]</td>
</tr>
<tr>
<td>Aves</td>
<td>FLD [3]</td>
</tr>
<tr>
<td>Mammalia</td>
<td>FLD [24], PTX [1]</td>
</tr>
</tbody>
</table>

*The taxa mentioned in the table are classes, except for those marked with an asterisk (in cases where the class designation is not available).

Domain abbreviations are according to CDD and are listed in Supplementary data, Table S-IX. Number of occurrences of domain appears within brackets.

Mytilus edulis, the crabs—Macrob Sacculata and Tachypleus tridentatus and the mites, Ixodes scapularis and Metaseiulus occidentalis (Table I; Supplementary data, Figure S5).

In diatoms, we found FLDs existing singly or in multiple copies and associated with glycosyl hydrolase family 16, α-L-arabinofuranosidase, non-FLD FA58C and pentraxin domains, and with domains of unknown function (DUFs)—DUF1800 and DUF1501 (Table I; Supplementary data, Table SIV and Figure S4). We observed that while a standalone FLD was present in the cryptomonad Guillardia theta, FLDS were part of huge polypeptides containing a C-type lectin domain and several to many CCP modules. In hemichordates, too, we found several FLD repeats within the Nectostyla vectensis and DUF1768 (Table I; Supplementary data, Table SIV and Figure S4). In molluscs, we found FLDs existing singly and standalone, in tandem repeats of 5 proteins of Arachnida, and FLDs were also found together with complement C1r/C1s, Uegf, Bmp1 (CUB) and COMMD8 (cooper metabolism gene MURR1 domain-containing protein 8) domains in the Arachnid, Ixodes scapularis (Table I; Supplementary data, Table SIV and Figure S4). As reported earlier (Odom and Vasta 2006), we observed that in Arthropods (belonging to classes Insecta, Branchiopoda and Arachnida), FLDs were part of huge polypeptides containing a C-type lectin domain and several to many CCP modules. In Macrob Sacculata and Tachypleus tridentatus (belonging to class Malacostraca) we found standalone FLDs. Such standalone FLDs were also found in a few proteins of Arachnids, and FLDs were also found together with complement C1r/C1s, Egf, Bmp1 (CUB) and COMMD8 (cooper metabolism gene MURR1 domain-containing protein 8) domains in the Arachnid, Ixodes scapularis (Table I; Supplementary data, Table SIV and Figure S4).
Supplementary data, Table SIV and Figure S4). We observed an increasing trend of size of FLD-containing polypeptide, number of FLD-containing polypeptides within a given organism, domain complexity and diversity of co-associated domains in FLD-containing polypeptides in the taxa Mollusca and Hemichordata, culminating in 73 different FLD sequence clusters (with >80% sequence identity) with very diverse domain architectures in Branchiostoma floridae, a cephalochordate belonging to order Branchiostomidae (Table I; Supplementary data, Table SIV and Figures S4, S6). It is interesting that the expansion of FLDs in these aforementioned taxa coincide with greater emphasis and use of innate immunity strategies.

We found FLDs existing singly or in tandem repeats of 2–6 in fishes, amphibians, reptiles, birds and mammals. Among these chordates, the only co-occurring domains (other than FLD itself) were a C-type lectin domain in proteins from the bony fishes, a non-FLD co-agglutination factor FA5SC domain in proteins from Coelachanhta and Amphibia, and a pentraxin domain in Amphibia and Mammalia (Table I; Supplementary data, Table SIV and Figure S4).

On the whole, therefore, in eukaryotes, we found FLDs majorly existing as a single copy or in tandem repeats, and, in some phyla, associated with diverse domains (Supplementary data, Figure S6). FLDs occurred towards the C- or N-terminus or were sandwiched between other domains of the polypeptide. As with prokaryotes, we observed similar domain architectures within closely related organisms (Supplementary data, Figure S4).

As suggested by Vasta et al., the relative paucity of bacterial FLDs indicate that perhaps the origin of evolution of FLDs was eukaryotic and FLDs were laterally transferred to bacteria, or less likely that the FLD originated early in evolution and was secondarily lost in most prokaryotic lineages (Vasta et al. 2004). While bacterial and eukaryotic FLDs mostly cluster separately in the phylogenetic tree, the bacterial FLDs are interspersed with several eukaryotic FLDs, hinting at more than one lateral transfer event (Supplementary data, Figure S5).

Sequence features of FLDs identified in this study
We analyzed the representative sequences (see Materials and Methods) of the FLD sequence clusters obtained in this study. Within the prokaryotic and viral FLD sequence clusters, sequences clustered together at 80% sequence identity shared the same motif and CDR1 characteristics. However, among sequences clustered together at 80% sequence identity in the eukaryotic FLD sequence clusters, we did notice variations in CDR1 and occasionally in the FLD motif, too. We would like to note that the analysis below is based on only the representative FLD sequences of the FLD clusters. Accession IDs of all FLDs identified in this study are available in Supplementary data, Table SVI.

FLD sequence motif
Binding to fucose by FLDs is mediated by hydrogen bonding between the basic side chains of a His/Arg/Arg triad (which is also part of the sequence motif characteristic of hitherto known FLDs, [HX(26)RXDX(4)R/K] (Vasta et al. 2008) and the axial 4-OH, equatorial 3-OH and ring O5 atoms of α-L-fucose. Our search strategy allowed for the identification of FLDs with variations in this sequence motif. Of the 95 prokaryotic sequence clusters identified in this study, there were 34 representative sequences with the typical FLD motif [HX(26)RXDX(4)R/K] and 30 representative FLD sequences with the conserved residues of the FLD motif but with variations in motif length (Supplementary data, Figure S7). Considering these variations in motif length found in this study, the FLD motif for prokaryotes can be regarded as HX(26–31)RX(0–2)DX(1–10)R/K. Further, in 8 of the 95 representative FLD sequences, the motif length seemed to be conserved but the motif presented variations in key amino acid residues. There were also 23 representative sequences with variations in key amino acids as well as motif length. In FLD sequences that lacked the critical His residue of the FLD motif, we observed the polar residues, Ser, Cys, Thr and Gln, and the apolar residues, Trp, Ala and Leu in the equivalent position. It is likely that the side chains of the polar residues make contact with the sugar, and it is also possible that where apolar residues substitute the His residue, a neighboring polar residue (such as Ser) may participate in sugar binding. Biochemical characterization would be required to verify this and note any effect on carbohydrate-binding specificity. The two (key) Arg residues could be located in the region of the motif in almost all sequences. In a few sequences, however, the amino acid residues, Ile (with a consecutively placed Lys), Ser and Cys replaced the first Arg of the motif. In place of the key Asp residue of the motif, the physicochemically similar residues, Glu or Asn occurred in some sequences, but there were also a few sequences where we could not pinpoint the Asp equivalent residue in this region (Supplementary data, Figure S7).

Among the six viral FLD sequence clusters, the typical FLD motif [HX(26)RXDX(4)R/K] was observed in two FLDs. Two viral FLDs had the motif HX(26)RXNX(4)R, one had the motif HX(29)RXXS(4)R and one had a motif with the Asp and Arg residues and conserved motif length but lacked the His residue (Supplementary data, Figure S8).

Of the 336 eukaryotic FLD sequence clusters, we found 118 representative sequences (from sequence clusters) with the FLD motif [HX(26)RXDX(4)R/K] and 49 representative FLD sequences with the conserved residues of the FLD motif, but with variations in motif length (Supplementary data, Figure S9). Considering the variations in motif length of these 49 representative FLD sequences, the FLD motif for eukaryotes can be regarded as HX(26–36)RX(0–2)DX(4–14)R/K. There were 34 representative sequences (from sequence clusters) with conserved motif length but without the conserved amino acids of the FLD motif. All of these 34 FLDs had both the conserved Arg/Lys residues. In two of these FLDs, an Asn and a Ser residue replaced the conserved Asp residue, and in all but one of these FLDs, various amino acids (Gln, Leu, Ser, Lys, Met, Arg, Cys and Ala) replaced the key His residue (Supplementary data, Figure S9). We also found 135 representative sequences (from the FLD sequence clusters) that had FLD motifs with variations in motif length and that lacked conserved motif residues. Among these 135 FLDs, we observed polar as well as apolar residues in place of the key His (Gln, Leu, Glu, Arg, Ala, Lys, Asn, Ser, Ile, Val, Trp, Met, Gly, Cys and Tyr), the first Arg (Lys, Thr, Ile, Trp, Val, Asn, Met, Gln, Pro and Ser) and the second Arg/Lys residues (Ile, Val, Gly, Trp, Pro, Met, Thr, His, Ser, Asp, Glu, Leu, Gln, Ala, Tyr and Phe) of the motif (Supplementary data, Figure S9). Many of these 135 sequences had the conserved Asp residue of the FLD motif. In the remaining sequences, Glu, Asn, Ser and Gly were frequently observed in place of the conserved Asp residue, though there were some occurrences of other amino acids, too (Supplementary data, Figure S9).

The His/Arg/Arg residues are known to be critical for fucose binding and mutagenesis of the first Arg in the FLD motif of Streptococcus mitis lectinolysin (R112A) resulted in abrogation of fucose binding (Farrand et al. 2008). However, there is not sufficient experimental evidence in the literature to predict if relaxation of the FLD motif in terms of length or substitution of conserved residues to residues other than Ala (perhaps polar residues that can still make contact with the fucose ring) will lead to changes in specificity or complete
loss of lectin activity. As mentioned above, FLDs in eukaryotes are frequently found in tandem repeats. We reasoned that loss-of-function substitutions of His/Arg/Arg in the FLD motif might be better tolerated in proteins with multiple FLDs due to the redundancy offered. We therefore examined substitutions of these positions in the FLD motif in proteins with multiple repeats, when compared with those with single repeats (Figure 4). Of 142 proteins with single repeats (with sequence alignment quality good enough to observe His substitutions), we observed that His was present in 82 FLDs. His was substituted by polar residues (Arg, Ser, Lys, Gln, Asn, Glu, Tyr and Cys) in 31 FLDs and by apolar residues (in Ala, Leu, Ile, Gly and Val) in 29 FLDs (Figure 4). The proteins with apolar residues substituting His belonged to diverse organisms; insects were prominent among them. Future biochemical investigations will be required to verify if these proteins are indeed capable of fucose binding.

Among proteins with multiple FLD repeats, 37 proteins had a conserved His in all repeats, 34 proteins had a conserved His in at least one FLD and 9 proteins had no conserved His. In proteins with at least one FLD with a conserved His, residues that substituted the His in the FLDs were polar residues such as Cys, Asp, Gln, Arg, Ser, Asn, Tyr, Met and Lys as well as apolar residues, Leu, Ala, Val, Gly and Trp. Of these, Leu, Gln and Ser occurred most frequently, and apolar residues occurred in ~35% cases. In multi-FLD repeat proteins where no FLDs contained the conserved His, polar residues Gln, Arg, Met, Ser and Lys, as well as apolar residues Ala, Ile and Leu occurred in place of His. The polar residue Gln was the most abundant substituting residue in this set of proteins. Apolar residues occurred in only ~10% cases. It is important to note that the apparent abundance of substituting residues might be biased by the number of proteins from closely related organisms with overall high-sequence similarity. Despite this, it is noteworthy that in each of these multi-repeat FLDs, even where apolar residues did occur in place of His, at least one FLD had a polar residue substituting His, suggesting that all these FLDs might still be capable of binding fucose.

Disulfide bridges and subcellular localization

The three-dimensional structure of AAA presents an unusual disulfide bridge between consecutive residues Cys82 and Cys83, which contacts the ring atoms C1 and C2 of α-1-fucose (Bianchet et al. 2002). This disulfide bridge is however absent in the structures of the prokaryotic FLDs—3LEI and 3CQO. We analyzed all the FLD sequence clusters identified in this study for the conservation of Cys residues involved in this unusual disulfide bridge. We observed that 18 of the 95 prokaryotic FLD sequence clusters had these Cys residues. These Cys residues occurred in the solitary FLD from Solibacteres, in all three Deltaproteobacteria FLDs, in a majority of Flavobacteria FLDs, and in a few Gammaproteobacteria, Actinobacteria and Bacilli FLDs. They were absent in the FLDs belonging to the classes Alphaproteobacteria, Epsilonproteobacteria, Planctomycetia and Verrucomicrobiae, and phylum Cyanobacteria (Supplementary data, Figure S7).

Considering that protein disulfide bonds would only be formed in a non-reducing environment, and conversely, that the conserved Cys

Fig. 4. Sequence characteristics of FLDs identified in the study. (A) Residues that substitute the critical His in FLDs occurring singly in eukaryotic proteins. (B) Residues that substitute the critical His in FLDs occurring in multiple repeats in eukaryotic proteins. (C) Multiple sequence alignments of eukaryotic FLDs showing insertions with duplicated FLD motifs. The FLD sequences with duplicated motifs have been aligned along with AAA. Residues of the FLD motifs are highlighted. This figure is available in black and white in print and in color at Glycobiology online.
residues implicated in the unusual disulfide bond might be replaced by other residues in proteins localized in a reducing environment, we next looked at the subcellular localization of the prokaryotic FLD sequences. Of the 95 prokaryotic FLD sequence clusters, 34 representative FLDs belonged to proteins predicted using the SignalP 4.0 webservice (Petersen et al. 2011) to have a canonical signal peptide, and 18 of these FLDs have the two consecutive Cys residues implicated in the unusual disulfide bond. A considerable number of FLDs predicted to have a signal peptide thus do not have this unusual disulfide bridge, suggesting that the FLD structure tolerates its absence well. Of the remaining 61 representative FLD sequences predicted to have cytoplasmic localization, only 4 have these Cys residues. It is possible that these four polypeptides are also located in a non-reducing environment, perhaps secreted through a non-canonical mode.

The three-dimensional structure of AAA also presents two other disulfide bridges, one between β-11 and CDR2 and another between β-7 and β-8 which help stabilize the entire structure. None of the 95 prokaryotic FLD sequence clusters has these Cys residues.

All of the six viral FLDs have the contiguous Cys residues implicated in the unusual disulfide bond, but do not have the Cys required for the two other conserved disulfide bridges (Supplementary data, Figure S8).

Among the 336 eukaryotic representative FLD sequences, the consecutive Cys residues involved in the unusual disulfide bond were observed in 228 FLDs. The Cys required for the two other disulfide bridges were conserved in proteins from most eukaryotic FLDs, but were absent in FLDs from diatoms and green algae (Supplementary data, Figure S9). Further, as discovered by Vasta et al. (Odom and Vasta 2006; Vasta et al. 2008), teleost proteins with binary FLD repeats showed a trend wherein the Cys residues involved in both these disulfide bridges were present in one FLD repeat that lacked the contiguous Cys residues involved in the unusual bond formation, and the other FLD repeat contained the contiguous Cys residues but instead lacked the two Cys residues involved in the internal of the two conserved disulfide bridges. This trend was however not absolute. Moreover, there were also proteins with the contiguous Cys residues that had these nested Cys residues. The occurrence of nested Cys residues in the FLD repeat lacking the contiguous Cys residues thus seems to be peculiar to certain teleosts as suggested by Vasta et al. (Odom and Vasta 2006; Vasta et al. 2008); we did not observe this trend in FLDs from other organisms. We also found two Branchiostoma floridae proteins (Branchiostoma floridae_8 gi|229295846 and Branchiostoma floridae_11 gi|260831065) with four consecutive Cys residues instead of just two.

Ca²⁺-binding motif
Available FLD structural data indicate that on one side of the β-barrel, a Ca²⁺ cation is bound by a 3_10 helix-rich substructure which includes the loops CDR1 and CDR2, and plays a role in structural stabilization. In AAA, the Ca²⁺ is coordinated by six residues, Asn35 (O), Asp38 (Oβ1), Asn40 (O), Ser49 (O, Oγ1), Cys146 (O) and Glu147 through seven oxygen atoms. Most of these interactions are provided by main chain atoms and hence the amino acid residues need not be conserved. The amino acids that interact through side chains are Ser49 and Glu147. We observed that the Glu was conserved in ~90% eukaryotic FLDs and ~50% prokaryotic FLDs (Supplementary data, Figures S7 and S9). Those that did not have Glu frequently had an Asp, Gln, Asn, Ser or Lys in this position. The position corresponding to Ser49 was conserved in ~50% eukaryotic and prokaryotic FLDs. In cases where the Ser is substituted with similar amino acids such as Thr, Ca²⁺ might still be coordinated by the same geometry. Alternatively, where other residues substitute this position, it is possible that a water molecule may substitute in coordination or the coordination geometry is modified as Vasta et al. previously suggested (Vasta et al. 2008). Both Ser and Glu were absent in the viral FLDs (Supplementary data, Figure S8).

Length variations in CDR1
As mentioned earlier, five loops named the complementarity determining regions (CDR1 through CDR5) connect the main β-sheets of the FLD structure proximal to the α-L-fucose-binding site. Of these, the length and sequence of CDR1 in particular has been suggested to alter interactions with subterminal saccharides and thereby determine fine oligosaccharide specificity (Bianchet et al. 2002). We designated the start and end residues of the CDR1 by performing multiple sequence alignments of the FLDs along with AAA, roughly following the convention in Bianchet et al. (2002) (the amino acids between XXX/XXX and XXG were considered as comprising CDR1). Of the 95 representative prokaryotic FLD sequences obtained in this study, we could assign CDR1 residues in 83 sequences. The CDR1 was 9 residues long in 44 of these sequences. The CDR1 length in the remaining sequences ranged from 5 to 15 residues (Supplementary data, Figure S10). The CDR1 varied from 8 to 12 residues, but was typically 9 residues long in most members of Bacilli, Deltaproteobacteria, Planctomycetia, Verrucomicrobiae, Actinobacteria, Gammaproteobacteria, Flavobacteria and Solibactera. FLDs belonging to Alphaproteobacteria members had longer CDR1s. In the 303 representative eukaryotic FLD sequences in which we could assign CDR1 residues, the CDR1 length ranged from 7 to 23 residues (Supplementary data, Figure S10). We could assign the CDR1 in three viral FLDs, and the CDR1 length in these FLDs was 9, 10 and 11 residues long.

Insertions in FLD sequences
We found five FLDs in the eukaryotes, Branchiostoma floridae and Latimeria chalumnae with large insertions in the β2 element between the CDR1 and CDR2 regions. These insertions seem to be consequences of internal duplication events. The region duplicated is that between n3 and the β-strand preceding CDR3 and it includes the FLD sequence motif (Figure 4), so these polypeptides have two fucose-binding modules each. In three of these FLDs, the His and the two Arg residues that are part of the additional fucose-binding sequence motif are conserved; the His is replaced by Leu in one FLD, and the second Arg of the motif is replaced by Val in another FLD. Experimental characterization will be required in order to verify if these large inserts with FLD sequence motifs do participate in fucose binding, and perhaps increase binding avidity.

FLD repeats within polypeptides
As mentioned earlier, many eukaryotic FLDs occur in tandem repeats. The sequence identity of these repeats ranges from 13 to 100% (Supplementary data, Figure S11), and hints at gene duplication and divergence. In polypeptides with more than two repeats, gene duplication seems to have occurred on more than one occasion, far apart in time. For instance, we found a Branchiostoma floridae protein with three FLDs sharing 74, 13 and 14% sequence identity with each other (Supplementary data, Figure S11). The difference in percent sequence identity between the domains suggests the occurrence of independent gene duplication events. In another instance, we found a protein from Ornithorhynchus anatinus that has five FLD repeats, while the third and fourth repeats share 100% sequence identity, the remaining domains share only 48–58% sequence identity (Supplementary data, Figure S11).
There are only three prokaryotic FLD sequence clusters with more than one FLD. They include proteins from *Mycococcus* sp. (contaminant ex DSM436) and *Streptococcus pneumoniae* GA41301, each having two FLDs and a three-FLD protein from *Streptococcus pneumoniae* CDC0288-04. The percent sequence identities of the FLD repeats in these proteins are 61 and 45% and 45 and 49%, respectively (Supplementary data, Figure S11).

No FLD repeats were observed in viruses.

**Discussion**

Previously published biochemical studies indicate that FLDs do not have absolute specificity for fucose (Vasta et al. 2004). While AAA binds to α-L-fucose and fucosylated oligosaccharides, it can also recognize 3-O-methyl-D-galactose and 3-O-methyl-D-fucose (Springer and Desai 1970). Tachylectin-4 from horseshoe crab exhibits Ca²⁺- dependent binding to terminal fucose residues and recognizes bacterial lipopolysaccharide, probably by binding to fucose-like colitose (3-deoxy fucose) and abequose (D-colitose) sugars (Saito et al. 1997). Bryohelmin, the sole algal F-type lectin identified so far, does not bind fucose, but instead recognizes N-acetyl glucosamine and N-acetyl galactosamine residues (Kim et al. 2006; Yoon et al. 2008), and Ranaspumatin-4, a component of the foam nest of the tungara frog, has specificity for galactose and lactose, not α-L-fucose (Fleming et al. 2009). The relatively small size of an FLD, the natural occurrence of FLDs that bind diverse glycoconjugates and its existence in diverse domain architectures make FLDs interesting additions for the glycobologist’s tool box. Studies aimed at creating molecular platforms for directing biological activities to specific glycosylated niches will benefit from analysis of naturally occurring FLDs.

Considering our aim of expanding the FLD sequence space, the search strategies we adopted were primarily sequence based, but reinforced with methods for checking the structural context of newly identified putative FLD sequences. With 437 FLD sequence clusters (clustered at 80% sequence identity) identified from proteins in eukaryotes, eubacteria and viruses by our approaches, our study reiterates that FLDs are found in proteins from various forms of life. Remarkably, we have identified FLDs for the first time in viruses, fungi, reptiles, birds and primitive mammals, and we notice a burst of FLD expansion in terms of occurrence and domain organization diversity in the taxa *Mollusca*, *Hemichordata* and *Branchiostoma*, perhaps coinciding with greater emphasis of innate immune strategies in these organisms.

Notwithstanding the wide phylogenetic distribution, FLD distribution remains selective (Supplementary data, Figure S2), being absent in protozoans, nematodes, echinoderms, ascidians, eutherian mammals and higher plants. Moreover, even among taxa/lineages with FLDs, loss of FLDs is discernible in closely related organisms. There are three likely explanations for this discontinuous distribution of the FLDs. First, FLDs have been selectively lost from some taxa, due to the lack of fitness value. This is a likely explanation for the aforementioned absence of FLDs in closely related organisms for which gapless, complete genome sequence data are available (Supplementary data, Table SVI). An example of this is the presence of an FLD in *Geobacter loyayi* SZ and the apparent absence of FLDs or any domain of the FA58C Pfam family (PF00754) in *Geobacter sulfurreducens*, *Geobacter bemidijensis*, *Geobacter uranireducens*, *Geobacter metallireducens*, *Geobacter daltonii*, *Geobacter* sp. M18 and *Geobacter* sp. M21. Secondly, the presence of FLDs in certain taxa is due to horizontal transfer. Several atypical phylogenetic associations can be observed in the phylogenetic tree of all FLDs, which hint at possible lateral transfer events (Supplementary data, Figure S5). However, individual cases need to be interpreted carefully as most of the eukaryotic genomes are not complete (i.e. they are draft genomes) and may have bacterial contamination. Thirdly, FLDs of some taxa may have diverged to the extent that they are no longer picked up by sequence homology searches but would still share the same structural fold. We performed searches in the Pfam database for polypeptides with domain architectures similar to the ones obtained in our study, but with a non-FLD FA58C domain (Pfam family PF00754) in place of the FLD in an attempt to retrieve some such diverged sequences. We obtained only three hits (Supplementary data, Figure S12), but two of them in the organisms, *Leeuwenhoekiella blandensis* and *Mycoplasma alligatoris*, where we did not find any FLDs with our three-pronged search strategy. This suggests that gene divergence might indeed account for the selective distribution in certain taxa. Interestingly, many FLDs identified in our study are actually associated with a non-FLD FA58C domain, a likely consequence of (gene duplication and) divergence, and the distribution of the FA58C Pfam family PF00754 is discontinuous, too (Supplementary data, Figure S2). As mentioned earlier, FLDs are classified in SCOP as the family “fucose-binding lectins”, which together with 31 other families (that include galactose-binding and discoidin domains (which in Pfam are classified together with FLDs in the FA58C domain family PF00754); various CBM domains and other domains) are classified under the “galactose-binding domain like” superfamily, the sole superfamily of the “galactose-binding domain like” fold. Sequences of these different families share negligible sequence similarity though many of them also share a similar placement of binding site (Vasta et al. 2004), and it is possible that FLDs in certain taxa have diverged to the sequence space of these families. In the absence of significant sequence similarity, it is difficult to know if this is indeed the case. However, in this context, we would like to note that many of the hits of our PSI-BLAST searches were indeed proteins without FLD domains, as judged by querying CDD and FFAS (and were not included in our FLD inventory).

Despite the identification of many eubacterial FLDs in this study, eukaryotic metazoan FLDs are more widespread and greater in number (Figure 1; Supplementary data, Figure S2) and hence, it seems more likely that this domain originated in eukaryotic metazoans and was later horizontally transferred to bacteria as suggested by Vasta et al. (2004). Though bacterial and eukaryotic FLDs, for the large part, cluster separately in the phylogenetic tree, several eukaryotic FLDs are interspersed among the bacterial sequences, suggestive of more than one lateral transfer event (Supplementary data, Figure S5).

The available FLD structures establish that the N- and C-termini of the polypeptides are placed close together at the end of the β-barrel opposite to the binding site (Bianchet et al. 2002, 2010; Farrand et al. 2008; Feil et al. 2012), and this perhaps facilitates the evolution of multiple tandem FLDs within a polypeptide (Vasta et al. 2008). Even prior to this study, FLDs have been known to exist singly (as in AAA) (Saito et al. 1997) or as multiple repeats, either arrayed in tandem (as in bony fishes) (Camparata et al. 2001; Vasta et al. 2004; Odom and Vasta 2006) or in mosaic combinations with other domains (as in Xenopus laevis, Drosophila melanogaster, Streptococcus mitis and Streptococcus pneumoniae) (Boraston et al. 2006; Farrand et al. 2008; Parisi et al. 2010). Our study has revealed new FLDs in diverse domain architectures, both in eukaryotes and in prokaryotes. While (prokaryotic) bacterial FLDs typically exist singly in combination with diverse domains, eukaryotic FLDs occur more frequently in multiple repeats, and sometimes also in tandem with other domains. The nature of domains that co-occur with FLDs varies widely. Carbohydrate-binding domains, carbohydrate-active enzyme
domains and other enzyme domains make up a large fraction of the co-occurring domains in bacteria. Among eukaryotes, C-type lectins, CCP modules and (non-FLD) FA58C domains are the common co-occurring domains. Diverse domain organizations occur in the cephalochordate, Branchiostoma floridae, which, intriguingly, also has an explosion of FLDS, perhaps in keeping with the emphasis on innate immunity in this organism. Domain architectures and FLD sequences are more similar among proteins of closely related organisms, though there are instances of atypical associations and atypical sequence similarities among FLD-containing proteins that hint at lateral transfer events. The enormous diversity of domain organization of the FLDS attests to the considerable functional diversification and biological specialization of this domain and suggests that FLDS have evolved to suit-specific requirements in organisms belonging to different lineages and ecological niches.

Fucosylated glycoconjugates have been reported in eukaryotes and prokaryotes (Ma et al. 2006). In eukaryotes, fucosylated carbohydrates play important roles in physiological processes such as fertilization, development, cell adhesion, inflammation, angiogenesis and cancer metastasis, while in prokaryotes, adhesion, colonization, molecular mimicry and modulation of the host immune response are themes that involve fucosylated glycoconjugates (Ma et al. 2006). 1-Fucose (6-deoxy-1-galactose) is a very common mammalian terminal monosaccharide (and accounts for ~7.1% abundance of the mammalian glycome represented in the carbohydrates database, GLYCOSCIENCES.de) (Lutter et al. 2006; Adibekian et al. 2011), and fucosyltransferases are expressed in other vertebrates and invertebrates, too, attesting to the presence of fucosylated glycoconjugates in diverse eukaryotic metazoa (Ma et al. 2006). The content of 1-fucose is, however, very low in bacteria [and accounts for just ~0.9% abundance of all the bacterial glycomes represented in the Bacterial Carbohydrate Structure Data Base (BCSDB) (Herget et al. 2008)] – [the values of percent abundance of 1-fucose in the glycomes of various taxa represented in BCSDB are 0.4% (Actinobacteria), 0.8% (Alphaproteobacteria), 0.1% (Bacilli) 0.1% (Betaproteobacteria), 0.3% (Gammaproteobacteria) and 0.8% (Enterobacteriales)] with the exception of Deltaproteobacteria and Epsilonproteobacteria, where 1-fucose is more abundant, constituting ~5.6% of the bacterial glycome (Herget et al. 2008; Adibekian et al. 2011); for instance, Helicobacter pylori, which is a member of Epsilonproteobacteria, expresses a relatively large proportion of Lewis A glycan, a fucosylated O-glycan otherwise commonly found in mammalian glycines, thus using mimicry to avoid detection by the host immune system (Nilsson et al. 2008). Streptococcus mitis and Streptococcus pneumoniae are both host pathogens and their FLDS recognize fucose on host tissues. Similarly, we might expect other FLDS from commensals or pathogens to preserve fucose binding. FLDS in algal-associated and plant endophytic bacteria could mediate association/colonization through recognition of fusocids or other glycoconjugates. In bacteria, 1-fucose is a component of the cell wall, polymers such as carotenoids, antibiotics and signaling molecules such as nodulation factors (Takaike et al. 2001; Samuel and Reeves 2003; Maki and Renkonen 2004; Riely et al. 2004), and it is likely that FLDS are involved in processes related to their function. Given the low content of 1-fucose in certain bacteria, as mentioned above, the contribution of FLDS in enabling diverse proteins such as methyltransferases, glycosyltransferases and glycosylhydrolases to effectively access and act on target fucosylated substrates could be very significant. Alternatively, it is possible that, akin to Rasnuspamin and Bryohealin, FLDS in some bacteria may have ‘custom’ evolved to other sugar specificities in keeping with individual factors such as ecological niche, metabolic constitution and cellular abundance of monosaccharides. In particular, biochemical analysis of site-directed mutants of FLDS co-occurring with glycosyltransferases/glycosylhydrolases (of different saccharide specificities) considered together with sequence analysis, might provide information about amino acid residues in the FLD evolved to accommodate different sugar specificities.

Fucose binding in the FLDs is mediated by key interactions between atoms of the fucose ring and amino acids in the CDRs—primarily, a His, an Asp and two Arg residues, which also make up the typical FLD motif (Vasta et al. 2008). Motif variations indicating altered carbohydrate binding or even loss have been noted earlier in the second domain of the four-FLD protein from Oncorhynchus mykiss and in the Furrowed and CG9095 proteins from Drosophila melanogaster and their homologs from other insects wherein the His/Arg/Arg triad are completely missing (Vasta et al. 2008). Our study has uncovered FLDS with variations in length as well as in the key amino acids of the characteristic FLD motif—His, Asp and Arg residues. We identified many bacterial proteins lacking one of the critical His/Arg/Arg residues. However, considering that polar amino acids are frequently located in place of the key His and Arg residues, it is likely that many of these variant bacterial FLDS retain the function of sugar binding. It will be interesting to see if these sequences do indeed code for functional lectins, and whether these natural FLD variants have altered specificity. Among eukaryotes also, there are FLDS sequences where apolar residues replace the critical His/Arg/Arg residues. In proteins with more than one FLD, the deleterious effect of such mutations might be offset by FLD redundancy. Indeed, we find that in each of the eukaryotic multi-FLD repeat proteins, a His or a charged residue substituting His is present in at least one of the FLD repeats. There are, however, proteins with single FLDS that do have apolar residues. It will be interesting to know what functions such FLDS have evolved to assume.

The available FLD structures indicate similar key interactions with fucose, the main differences being in the CDRs. The length and sequence of CDR1, in particular, is thought to alter interactions with subterminal saccharides, thus changing fine oligosaccharide specificities (Bianchet et al. 2002). The crystal structures of Streptococcus pneumoniae SP2159 FLD complexed with the H trisaccharide, blood group A tetrasaccharide and Lewis y antigens, and Streptococcus mitis FLD complexed with Lewis y and Lewis b antigens indicate that while the primary determinant of binding to these oligosaccharides is clearly through specific interactions with the key terminal fucose, there are several direct or water-mediated hydrogen bonds and van der Waals contacts made with the subterminal saccharide residues that shape complementarity between the binding pocket and the oligosaccharide to which fucose is attached (Boraston et al. 2006; Feil et al. 2012). Our study also reveals FLDS with widely varying CDR1 length in prokaryotes and eukaryotes. Interestingly, FLDS belonging to Alphaproteobacteria members (in proteins with co-occurring glycosyltransferase domains) had longer CDR1s, perhaps an adaptation tailored to accommodate these proteins’ substrates.

The bound L-fucose ring in AAA also makes contact with an unusual disulfide bond between two contiguous Cys residues (Bianchet et al. 2002). The FLDS from Streptococcus pneumoniae SP2159 and Streptococcus mitis lectinolysin, and the C-terminal FLD of Morone saxatilis MsAFB32 lack these adjacent Cys residues that are involved in the inter-strand disulfide bridge in AAA but their 3D structures indicate that a hydrophobic residue in CDR4 provides van der Waals contact with α-L-fucose (Boraston et al. 2006; Feil et al. 2012). Similarly, we find that a majority of bacterial FLDS do not have the contiguous Cys residues that form the unusual disulfide bond providing
hydrophobic contact with the ligand. Among eukaryotes, too, the contiguous Cys residues are absent in 108 FLDs. It seems likely that the FLD structure tolerates the loss of this disulfide bond fairly well. Interestingly, the bacterial FLDs also lack Cys residues that form two other conserved disulfide bonds in eukaryotes. It is tempting to speculate that the FLD fold originally evolved in a non-reducing environment; loss of signal peptide and subsequent loss of disulfide bridges (and Cys residues in the absence of the selective pressure of a non-reducing environment) occurred upon lateral transfer to certain prokaryotes, likely countered by other stabilizing non-covalent interactions during the process of natural selection.

In all, our study has uncovered a wealth of FLDs in diverse domain organizations from eukaryotes, prokaryotes and viruses. This newly annotated sequence space is a starting point for biochemical and molecular biology studies to determine the saccharide specificity of naturally occurring FLD variants. We hope that ultimately these FLDs will be exploited for targeting desired biological activities to distinct glycosylated niches.

Materials and methods

Databases and search tools

Pfam database 27.0 (Finn et al. 2014) (sequences of PF00754 downloaded on 3 April 2014), NR database [downloaded from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov) on 4 September 2013], Pf_scan (de Castro et al. 2006), PSI-BLAST (Altschul et al. 1997), HMMer (Eddy 1998, 2011), CDD (Marchler-Bauer et al. 2011) and BlastClust (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) from the NCBI-BLAST package (Altschul et al. 1990) were employed for this study. BlastClust clusters sequences according to a user-defined percent identity and the sequences within a cluster are arranged in decreasing order of length. The first sequence, which is the longest sequence, was taken as the representative member of each cluster.

Venn diagrams

Venn diagrams were plotted using VennMaster (Kestler et al. 2008). Sequences from SMART (Letunic et al. 2004) and CAZy (Cantarel et al. 2009) databases used in generating the Venn diagrams were downloaded on 14 April 2014.

Generation of HMM profile

We aligned FLD sequence clusters obtained by the Pfam search using profile consistency multiple sequence alignment (PCMA) (Pei et al. 2003) for analysis of conserved motif and flanking regions in the domain. We aligned FLD sequence clusters from eukaryotes, prokaryotes and viruses separately and generated hidden Markov model (HMM) profiles using HMMer (Eddy 2011).

Multiple sequence alignments

We generated multiple sequence alignments by tree-based consistency objective function for alignment evaluation (T-Coffee) (Notredame et al. 2000) using 1k12 PDB structure (AAA) as structural template and manually adjusted, and analyzed sequences for extent of variation in critical residues and length between residues in the motif.

Generation of phylogenetic tree

We constructed a phylogenetic tree of FLDs using Molecular Evolutionary Genetic Analysis (MEGA v6.0) (Tamura et al. 2013) to draw evolutionary inferences and to visualize co-occurrence of domains. We generated a maximum likelihood (ML) tree using all sites with 100 bootstraps [WAG (Whelan and Goldman) model of substitution; rates among sites: gamma distributed with invariant sites; ML heuristic method: subtree-pruning-regrafting, initial tree for ML: maximum parsimony (MP) tree]. We generated domain architecture figures and mapped them on to the tree using the Interactive Tree Of Life (iTOL) server (Letunic and Bork 2011).

Supplementary data

Supplementary data for this article are available online at http://glycob.oxfordjournals.org/.

Authors contributions

R.T.N.C. conceived the study, S.S. designed and coordinated the bioinformatics analysis, R.B. and I.K. performed the bioinformatics analysis. All authors participated in interpretation of data analysis, and in writing and editing the manuscript.

Funding

The authors acknowledge Department of Science and Technology, Government of India (FAST-TRACK grant no. SR/FT/LS-87/2012 to R.T.N.C.) and CSIR-IMTECH (manuscript communication number 017/2015) for funding this study.

Acknowledgements

R.B. and I.K. acknowledge University Grants Commission, New Delhi, India for their fellowships.

Conflict of interest statement

None declared.

Abbreviations

AAA, Anguilla anguilla agglutinin; BCSDB, Bacterial Carbohydrate Structure Database; CAZy, Carbohydrate-Active enZymes; CRM, carbohydrate-binding module; CCP, complement control protein module; CDD, conserved domain database; CDR, complementarity determining region; COMMDs, copper metabolism gene MURR1 domain-containing protein 8; CUB, complement C1r/ C1s, Uegf, Bmp1; DUF, domain of unknown function; EGF-like, epidermal growth factor-like domain; FASCG, coagulation factor 5/8 C-terminal domain; FFAS, fold and function assignment system; FLD, F-type lectin domain; FTP, cel-fucolactin tachylectin-4 pentraxin-1 domain; GH, glycosylhydrolase; GT, glycosyltransferase; GCC2, GCC3, GRIP and coiled-coil domain-containing 2 and 3; HMM, hidden Markov model; HYR, hyalin repeat domain; iTOL, interactive tree of life; LiC(D), Lipopolysaccharide cholinephosphotransferase protein family; MAM, meprin, A5 protein and protein tyrosine phosphatase Mu domain; MEGA, Molecular Evolutionary Genetic Analysis; ML, maximum likelihood; MP, maximum parsimony; NCBI, National Centre for Biotechnology Information; NPCBM, novel putative carbohydrate-binding module; NR, non-redundant; OmpNH, outer membrane protein (OmpNH-like) domain; PAA_N_APPLE, APPLE-like domains present in Plasminogen, C. elegans hypothetic- al ORFs and the extracellular portion of plant receptor-like protein kinases; PCMA, profile consistency multiple sequence alignment; PDB, Protein Data Bank; PSI-BLAST, Position-Specific Iterative Basic Local Search Alignment Tool; PVC, plantacytomes-verrucomicrobia-chlamydiae; RCC1, regulator of chromosome condensation repeat; SCOP, structural classification of proteins;
SMART, Simple Modular Architecture Research Tool; T-Coffee, tree-based consistency objective function for alignment evaluation; TECPR, β-propeller repeats in Phytophthora polycyclic tectonin domains; WAG, Whelan and Goldman. Abbreviations for co-associated domains are listed in Supplementary data, Tables VIII and IX.

References


Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local align-


Biancheta MA, Odom EW, Vasta GR, Amzel LM. 2002. A novel fucose recogni-


tures and PRORule-associated functional and structural res-


Fleming RI, Mackeren CD, Cooper A, Kennedy MW. 2009. Foam nest compo-
nents of the tungara frog: A cocktail of proteins conferring physical and bio-


son with mammalian glycans. BMC Struct Biol. 8:35.


