FEBS Letters xxx (2015) xxx–xxx

FEBS Letters

journal homepage: www.FEBSLetters.org

Evolutionary analysis of the global landscape of protein domain types and domain architectures associated with family 14 carbohydrate-binding modules

Ti-Cheng Chang, Ioannis Stergiopoulos *

Department of Plant Pathology, University of California Davis, Davis, CA, USA

ARTICLE INFO

Article history:
Received 15 April 2015
Revised 11 May 2015
Accepted 20 May 2015
Available online xxxx

Edited by Takashi Gojobori

Keywords:
Chitin
Family 14 carbohydrate-binding module
Modularity
Promiscuity
Versatility
Supra-domain

ABSTRACT

Domain promiscuity is a powerful evolutionary force that promotes functional innovation in proteins, thus increasing proteome and organismal complexity. Carbohydrate-binding modules, in particular, are known to partake in complex modular architectures that play crucial roles in numerous biochemical and molecular processes. However, the extent, functional, and evolutionary significance of promiscuity is shrouded in mystery for most CBM families. Here, we analyzed the global promiscuity of family 14 carbohydrate-binding modules (CBM14s) and show that fusion, fission, and reorganization events with numerous other domain types interplayed incessantly in a lineage-dependent manner to likely facilitate species adaptation and functional innovation in the family.

1. Introduction

Carbohydrate-binding modules (CBMs) are ubiquitous molecules in nature that are frequently found as discrete non-catalytic components of carbohydrate-active enzymes (CAZymes), in which they promote the avidity of the enzyme for the target saccharide substrate. As a result, the general architecture of modular CAZymes is frequently composed of the catalytic module linked to one or several CBMs, while additional discrete domains might be present as well, thereby increasing the complexity of these architectures [1,2]. Overall, the high propensity of CBMs to fuse with other domains suggests that these modules are part of an organism's proteome backbone that can be readily recruited for service in a milieu of enzymatic and biochemical processes [3]. However, despite the functional and evolutionary significance of acknowledging modular arrangements in CBM-containing proteins, the number and diversity of combinations in which a particular CBM can be involved, often referred to as *versatility* or *promiscuity* [4,5], remains for most CBM families largely unknown.

Recently, we reported on the molecular evolutionary analysis of family 14 carbohydrate-binding modules (CBM14s) across all domains of life [6]. Members of this family show specific affinity for chitin, a $\beta(1 \rightarrow 4)$ linked *N*-acetyl-D-glucosamine (GlcNAc) polysaccharide, and thus, not surprisingly, CBM14s are frequently connected to catalytic domains with chitinolytic activity. Our previous analysis indicated that the evolution of this family was largely shaped by horizontal gene transfer, multiple lineage-specific expansions and contractions, and positive selection that were overall suggestive of functional diversification [6]. Here, we expand these studies to further examine the dynamics of the CBM14 family evolution from the perspective of versatility or promiscuity, by surveying the diversity of domain types and complexity of domain

Abbreviations: CBM, carbohydrate-binding modules; CBM14, family 14 carbohydrate-binding module; CAZymes, carbohydrate-active enzymes; HMM, Hidden Markov Model; SP, signal peptide; GH18, glycoside hydrolase family 18 domain; GH19, glycoside hydrolase family 19 domain; LDLa, lipoprotein receptor class A domain; IGv, Immunoglobulin variable-set domain; GlcNAc, *N*-acetylglucosamine (or *N*-acetyl-D-glucosamine); Ig, immunoglobulin; VCBPs, variable region-containing chitin-binding proteins; SRCR, Scavenger receptor cysteine-rich protein domain; Sp, serine protease (e.g. as in Sp22D gene from *Anopheles gambiae*); Pdi, polysaccharide deacetylase domain

Author contributions: T.C., performed analyses; analyzed data; wrote the manuscript. I.S., conceived and supervised the study; analyzed data; wrote the manuscript.

* Corresponding author at: University of California Davis, Department of Plant Pathology, One Shield Avenue, Davis, CA 95616-8751, USA. Fax: +1 530 752 5674.
E-mail address: istergiopoulos@ucdavis.edu (I. Stergiopoulos).

http://dx.doi.org/10.1016/j.febslet.2015.05.048
0014-5793/

architectures exhibited by modular CBM14-containing proteins. Our analysis revealed an impressive repertoire of domains and consequently domain architectures associated with the CBM14, indicating that the ligand-binding properties of this module have been exploited several times in nature, possibly as a means to facilitate functional innovation in higher eukaryotes.

2. Materials and methods

2.1. Identification of domains in multimeric CBM14-containing proteins and subsequent domain–domain interaction network analysis

To precisely identify the diversity of domain types that are associated with the CBM14, we used InterProScan [7] and the Hidden Markov Model (HMM) of each domain deposited in the Pfam database [8] to search against the 3432 CBM14-containing proteins (*e*-value <1E-5) that we have previously identified [6,9]. The signal peptide (SP) was considered as one type of domain, meaning that proteins consisting only of CBM14s fused to a SP were regarded as homomultimeric, as the final polypeptide chain would consist only of CBM14s. Data on the topological arrangement of the various domains in different protein architectures were retrieved by constructing a directed network graph, using Cytoscape [10] and custom Perl scripts. The basic unit in the network analysis was a domain pair, defined as two domains located directly adjacent to each other in a polypeptide chain taking the order of the domains into account as well. For instance, in a single protein with three consecutive domains, A–B–C, two domain pairs were defined (i.e. A–B and B–C). Also, combinations such as A–B and B–A were classified as two discrete domain pairs. In the network, each node represented a domain and the edge was connected to the paired partners. The direction of the edge represented the order of each domain pair with the source of the arrow as the head domain (i.e. the domain facing the N-terminus of one protein) and the target as the tail domain (i.e. the domain facing the C-terminus of the protein). The number of domain pairs was used as the weight of the edges. The degrees in the network were defined as the number of edges connected to an individual node and were divided to in(wards) and out(wards). Finally, the frequency of each type of domain pair was counted at three taxonomic levels (i.e. Phylum, Kingdom and Domain) and the taxonomic distribution of the domain pairs as well as the overall network properties were further analyzed using the Network analyzer [11] implemented in Cytoscape. The sub-network was constructed based on the nodes with a non-zero clustering coefficient.

3. Results and discussion

3.1. Diversity of domain types co-occurring with the CBM14 in modular proteins

A total of 11633 domains, of which 3973 were domains other than the CBM14 were recovered from CBM14-containing proteins that corresponded to 94 unique Pfam domains (Table S1). This indicates that the CBM14 has the propensity to associate with a variety of other domain types, and thus can be regarded as a promiscuous domain [5]. Since chitin is mostly found in nature as an extracellular matrix polysaccharide [12–14], it is perhaps not surprising that, next to CBM14s (7660/11633), the SP is the domain most commonly recovered from CBM14-containing proteins (2236/11633). However, 1196 CBM14-containing proteins which lack a SP were also identified, and these might be involved in intermediate steps of chitin biosynthesis, metabolism, and transport to the extracellular matrix [12]. Other Pfam domains most frequently recovered were the glycoside hydrolase family 18 domain (GH18; PF00704) (866/11633), which catalyzes the enzymatic degradation of chitin [15]; the low-density lipoprotein receptor class A domain (LDLa; PF00057) (175/11633), which binds and transports ligand lipoproteins into cells, thus playing a key role in lipid metabolism [16]; and the Immunoglobulin variable-set domain (IGv; PF07686) (129/11633), which belongs to a class of Ig-like domains that are involved in a variety of functions, including cell–cell recognition, cell-surface receptors, and others [17]. The remaining 89 Pfam domain types that were recovered from CBM14-containing proteins had a frequency of occurrence less than one hundred times each, including 48 domains that were found only once (Table S1). Overall, the diversity of domains in CBM14-containing proteins indicates that these proteins can partake in a multitude of biological processes and pathways.

3.2. Diversity of domain architectures in modular CBM14-containing proteins

The large number of various domain types that can be combined with the CBM14 in modular proteins was consequently translated into an increase in protein versatility. Of the 3432 CBM14-containing proteins identified previously [6], 2646 (77.1%) are modular proteins that emerged through the combination of CBM14(s) with one or more of the other 93 Pfam domain types that were recovered from the CBM14-containing proteins. Analysis of all the domain arrangements identified a total of 224 unique domain architectures, 208 of which refer to modular proteins (Fig. 1, Tables S2 and S3). After excluding proteins with one or more CBM14s fused to a signal peptide (SP) and not to any other domain types (1572 proteins), the number of "true" modular proteins, which contain only CBM14 domains in their mature form, still remains high (1074 proteins). These modular proteins represent 185 unique domain architectures, implying that in many cases an organism's repertoire of CBM14s can be seen as a collection of CBM14-containing proteins with different domain architectures. The majority of the modular CBM14-containing proteins had only one (949 proteins) or two (74 proteins) additional domain types, however proteins with three or more extra domain types (51 proteins) were identified as well (Table S4), indicating that CBM14s can partake in the formation of proteins with complex domain architectures.

To further elucidate the complexity of the domain architectures, we performed a directed domain combination network analysis that traced the linear pattern of pairwise domain combinations in all 3432 CBM14-containing proteins [18] (Fig. 2). The analysis revealed a total of 8201 domain pairs that accounted for 172 unique pairs. We next examined the degree number of each node in the network, i.e. the number of edges connected to a single node. Overall, the degree distributions of the network followed a power law ($p(k) \propto k^{-\gamma}$, k as node degree) with an γ value of 0.934 (correlation $R = 0.957$) for in-degree distribution and 1.008 for out-degree distribution ($R = 0.991$) (Fig. S1), suggesting that it shares the properties of a scale-free network. One major feature of such a network is the presence of central hub nodes with relatively high degree numbers, which corresponds to a high number of connected partners [19]. In the constructed network, the CBM14 was placed as one of the central hubs with an in-degree of 51 and an out-degree of 38, which is considerably higher than the total degree of the rest of the domains (<21) (Fig. 2). In addition, the clustering coefficient revealed a relatively low value of 0.0067 for the CBM14 node, compared to the highest value of 1.0 in the network (Table S5). Combined, these results indicate that the CBM14 has a significantly higher number of domain partners (61 in total) as compared to any other domain in modular

Please cite this article in press as: Chang, T.-C. and Stergiopoulos, I. Evolutionary analysis of the global landscape of protein domain types and domain architectures associated with family 14 carbohydrate-binding modules. FEBS Lett. (2015), http://dx.doi.org/10.1016/j.febslet.2015.05.048