FEBS Letters 584 (2010) 2421–2426

FEBS Letters

journal homepage: www.FEBSLetters.org

Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module

Luciane V. Mello, Xin Chen, Daniel J. Rigden *

School of Biological Sciences, University of Liverpool, UK

ARTICLE INFO

Article history:
Received 4 March 2010
Revised 13 April 2010
Accepted 16 April 2010
Available online 21 April 2010

Edited by Gianni Cesareni

Keywords:
Protein domain
Metagenomics
Carbohydrate-binding
Gut bacteria
Bacteriodetes

ABSTRACT

Third-generation sequencing has given new impetus to protein sequence database growth, revealing new domains. Description and analysis of these is required to further improve the coverage and utility of domain databases. A novel domain, here named BACON, was discovered from analysis of metagenomic data obtained from gut bacteria. Domain architectures unambiguously link its function to carbohydrate metabolism but a further strong connection to protease domains suggests that many BACON domains bind glycoproteins. Conserved residues in the BACON domain are also characteristic of carbohydrate binding while its biased phyletic distribution and other data suggest mucin as a potential specific target.

1. Introduction

The arrival of third-generation sequencing technology has allowed the continuation in the exponential rate of growth of databases of protein sequences [1]. Notably, the technology has facilitated the unbiased and cloning-independent exploration of the populations occupying particular environments – metagenomics. The resulting volume of data converts ORFans into small families and enhances sequence diversity in existing families [2]. Importantly, such projects also reveal new families [2], although different accounting methods lead to different estimates of how quickly the number of known families is increasing [1,2].

A proper description of this fast-expanding protein universe requires that domain databases cover as much protein sequence as possible and, where possible, are annotated with functions [3]. Such annotations should ideally be made experimentally but careful bioinformatics predictions can provide a helpful stopgap and direct the laboratory experiments. As recently suggested, unless concerted efforts at reliable function annotation are made, the ever-increasing volume of sequence data combined with a more slowly expanded knowledgebase of function could lead to serious problems of erroneous annotation [4]. Superfamilies harbouring diverse catalytic activities are particularly prone to such problems [5].

In this work, we identify and characterize a novel domain, BACON (Bacteriodetes-Associated Carbohydrate-binding Often N-terminal), discovered by comparison between sequences of metagenomic origin. The BACON domain bears no detectable relationship to any other but a variety of data – domain architectures, sequence conservation and phyletic distribution – argue for a carbohydrate-binding function. Mucin binding may be a common theme of many BACON domains.

2. Methods

Database searching was carried out using BLAST and PSI-BLAST [6] and an e-value threshold of 0.01. The protein query was the N-terminal region of a cellulase sequence deriving from a buffalo rumen metagenomics project (accession ACA61145; [7]) and the database nr [8]. The resulting sequences were aligned using MUSCLE [9]. The domain alignment of this domain (named BACON, as explained elsewhere) was manipulated and corrected using Jalview [10], which was also used to obtain smaller maximally non-redundant sets for presentation and *ab initio* modelling. Fold recognition at the META server [11] and profile-profile matching with HHpred [12] were used to verify that the BACON domain bore no statistically significant similarity to any previously described domain. Domain architectures were first analysed using RPS-BLAST and the CDD database [13]. More distant, though still reliable,

* Corresponding author. Address: School of Biological Sciences, University of Liverpool, Crown St., Liverpool L69 7ZB, UK. Fax: +44 151 795 4414.
E-mail address: drigden@liv.ac.uk (D.J. Rigden).

0014-5793/$36.00
doi:10.1016/j.febslet.2010.04.045

relationships between BACON-containing sequences and Pfam domains [14] were sought using HHpred. Signal peptides were sought in sequences containing BACON domains using LipoP [15] and SignalP 3.0 [16] and twin arginine export motifs with TatP [17]. Searches for transmembrane helices were done with TMHMM 2.0 [18]. Phylogenetic analysis was done using the Neighbour Joining and Minimum Evolution algorithms of MEGA 4 [19].

3. Results and discussion

3.1. Domain definition

Metagenomics results provide promising raw material for discovery of novel domains [2]. We initially noted a BLAST match between the N-terminal portion of a putative cellulase (NCBI accession CAJ19151; [8]) and two distinct regions of an annotated 1,3–1,4-beta-glucanase (accession AAX16429). The former sequence derived from a bovine rumen metagenomics project [20] while the latter was found in an uncultured murine bowel bacterium [21]. Our PSI-BLAST database searching for related sequences reached convergence with a set of 143 proteins containing one or more instances of what we subsequently named as the BACON (for Bacteriodetes-Associated Carbohydrate-binding Often N-terminal) domain. By applying fold recognition and profile-profile matching to diverse representatives we verified that the BACON domain is novel: no significant relationships to any other domain in CDD could be demonstrated.

Determination of the sequence limits of the BACON domain was significantly simplified by its occurrence in arrays of up to six consecutive domains and, most helpfully, by its terminal location in some proteins, either immediately post signal peptide or at the very C-terminus (Fig. 1). The alignment of 13 representative BACON sequences in Fig. 2 reveals only four strongly conserved residues. In the representative sequence, the second BACON domain from a probable peptidase of *Bacteriodes plebius* (accession ZP_03208788; Figs. 1 and 2) these are residues Trp154, Asn176, Arg182 and Gln206. Of these amino acids, only Arg is statistically common at catalytic sites [22] suggesting that the BACON domain is unlikely to have a catalytic function.

3.2. Domain architectures

Most of the sequences containing the BACON domain contain signal peptides. All the 137 sequences from identified organisms come from Gram negative bacteria, the remaining six being of metagenomic origin. With LipoP [15], 103 of the 137 Gram negative sequences are predicted to contain type II signal peptides leading to lipoprotein cell membrane localization. A further 14 are predicted to code for extracellular proteins since they have cleavable type I signal peptides. By analysis with SignalP 3.0, five of the six sequences from unidentified organisms were predicted to contain signal peptides, whether subject to Gram positive- or Gram negative-specific predictors. For just 21 of 143 proteins was neither signal peptide predicted. In some of these cases, other domains within the protein are characteristically extracellular e.g. CARDB (Cell Adhesion related Domain found in Bacteria; PF07705). These sequences may result suggesting from errors in annotating open reading frame starts and, consequently, the N-termini of some proteins. No sequences were predicted to contain transmembrane helices.

The set of BACON-containing domain architectures exhibits some intriguing themes. Most strikingly, BACON is found in combination with eight different domains that are catalytically active on carbohydrates [23], seven glycoside hydrolases (GHs) [24] and one polysaccharide lyase. Other, non-catalytic domains present in Fig. 1 are also associated with carbohydrate binding – CBM_48 (PF02922; [25]), PA14 (PF07691; [26]) and F5_F8_C (also known as the discoidin domain; PF00754 [27]) However, the link to protease activity is almost as strong with five different metallo- and serine protease catalytic domains found fused with BACON domains. Examination of domain architectures for other carbohydrate-binding modules (not shown) shows that this degree of fusion with proteases is unusual and possibly therefore significant. Most simply, this could be an indication that BACON binds to ligands that contain both carbohydrate and protein moieties.

3.3. Domain distribution

The species containing BACON domains are largely those of the Bacteroidetes phylum and in particular *Bacteroides* species (Table 1). Of the 27 species, only seven are not in the Bacteroidetes phylum. Although, this tendency is exaggerated by current efforts to sequence multiple *Bacteroides* species as part of the Human Microbiome Project [28], other considerations confirm the association. Thus, both the most BACON-containing proteins (17) and the largest number of BACON domains (25) are seen in *Bacteroides thetaiotamicron*. Outside the Bacteroidetes phylum the largest number of proteins is four in *Solibacter usitatus*. Three BACON-containing proteins are presently found in the incomplete genome of *Verrucomicrobiae bacterium* DG1235. The distribution of BACON-containing species outside the Bacteroidetes phylum is notably sporadic, suggestive of an evolutionary scenario in which BACON evolved in the Bacteroidetes and spread by horizontal gene transfer to other species. It is also notable that tandem BACON arrangements of more than three domains are confined to the Bacteroidetes phylum. By bootstrapped phylogenetic analysis (not shown), rather few nodes in consensus trees were strongly supported by high confidence values. However, some nodes that were strongly supported in both Neighbour Joining and Minimum Evolution trees offers hints that duplication of BACON domains is an ongoing process. Thus, the first BACON domains of the two *B. coprocola* proteins encoded by BACCOP_03538 and BACCOP_00860 loci, are consistently linked. Furthermore, the second and third BACON domains in the *Verrucomicrobiae bacterium* protein encoded by the VDG1235_3084 locus are neighbours, suggesting a recent intragenic duplication.

The most BACON-rich species, *B. thetaiotamicron* is a well-characterized human gut symbiont [29], living on and in the mucus layers of the distal intestine [30]. Its genome sequence [31] revealed a rich 'glycobiome' containing, for example, 226 glycoside hydrolases, compared to its human host's tally of around 98 [32]. In common with other gut bacteria, *B. thetaiotamicron* produces enzymes that degrade carbohydrates that are indigestible to the host [32]. Intriguingly, experiments have shown that at times when the diet of the human changes to reduce the availability of dietary carbohydrates the bacterium can turn to digesting host mucin. *B. thetaiotamicron* then produce enzymes specific for mucin breakdown, and thereby uses this glycoprotein as a source of carbohydrate and energy [30]. *B. thetaiotamicron* and other *Bacteroides* with similar capability therefore require mucin-recognising proteins for two reasons: first, as mentioned, for times when mucin becomes an energy source. The second need is that of cell adherence: without the means to attach to the mucin layer, symbiotic bacterial cells would simply be swept away [32].

3.4. Domain function

Taken together with the unusual twin association with carbohydrate- and protein-active enzymes (Fig. 1), it is tempting to propose from the species distribution that BACON may bind to mucin and thereby fulfil the needs of *Bacteroides* species for adhesion and

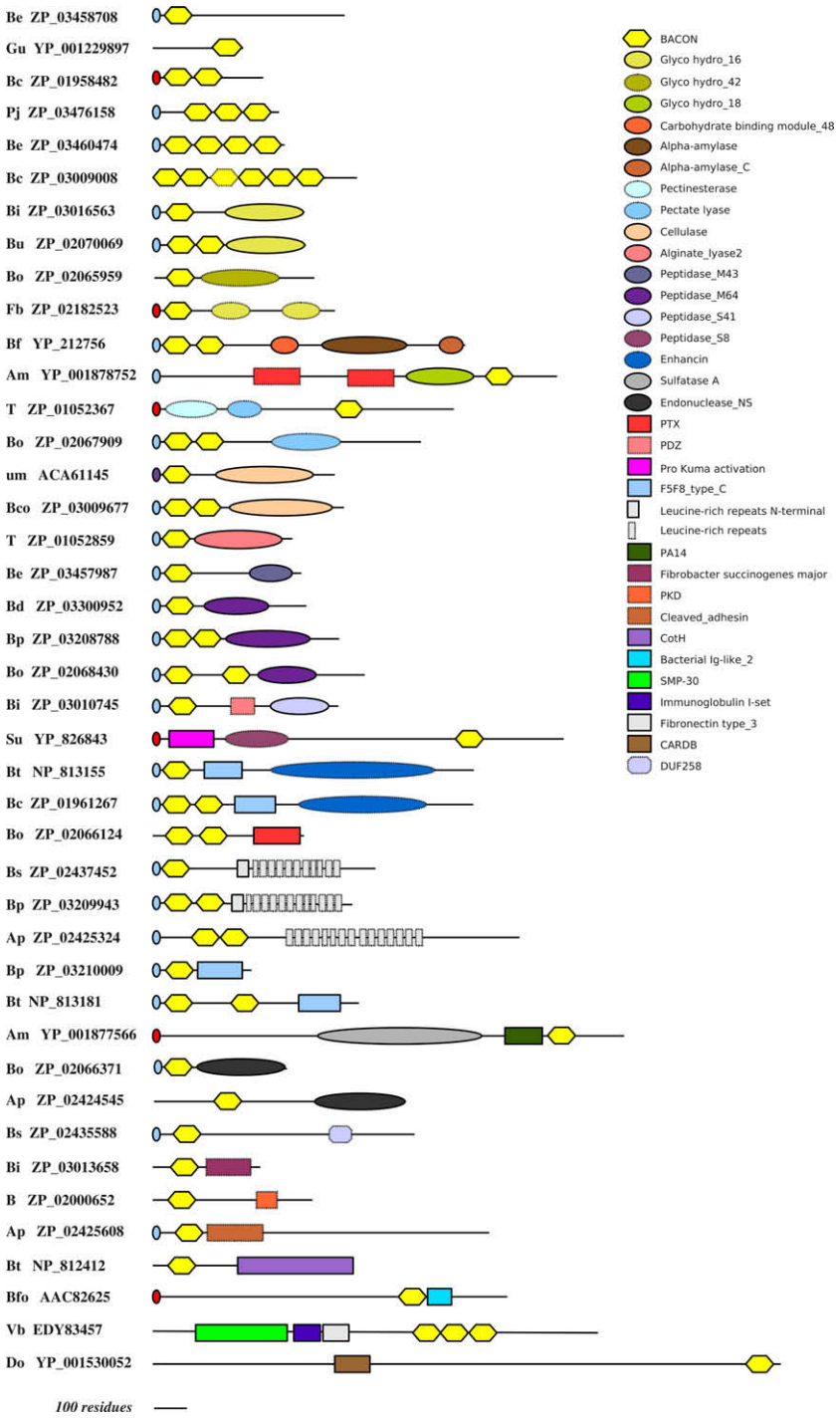

Fig. 1. Pfam [39] domain architectures involving BACON. Domains were located using RPS-BLAST [13] with the exception of those delineated by dashed lines which were only detected with HHPRED [12]. Small blue N-terminal circles indicate type II signal peptides from LipoP predictions [15] that would lead to lipoprotein modification; red circles are LipoP predicted type I cleavable signal peptides. Purple circles indicate signal peptides predicted with SignalP [16] for sequences of unknown origin. Proteins are labelled with abbreviated species name (Be = *Bacteroides eggerthii*; Gu = *Geobacter uraniireducens*; Bc = *B. caccae*; Pj = *Parabacteroides johnsonii*; Bd = *B. dorei*; Bp = *B. plebeius*; Bo = *B. ovatus*; Bt = *B. thetaiotaomicron*; Su = *Solibacter usitatus*; Bi = *B. intestinalis*; Bu = *B. uniformis*; Fb = Flavobacteriales bacterium; Am = *Akkersmansia muciniphila*; T = *Tenacibaculum*; Bf = *B. fragilis*; um = uncultured microorganism; Bco = *B. coprocola*; Bs = *B. stercoris*; Ap = *Alistipes putredinis*; B = Beggiatoa sp; Vb = Verrucomicrobiae bacterium; Bfo = *B. forsythus*; Do = *Desulfococcus oleovorans*) and NCBI accession.

occasional recognition as energy source. Other data are consistent with this notion. First, one of the non-Bacteroidetes species containing BACON domains is *Akkermansia muciniphila*, a bacterium isolated for its property of degrading human mucin [33]. This organism contains two BACON domain architectures not otherwise found in the databases. The first combines BACON with a Glycoside Hydrolase (GH) 18 domain and two domains distantly related to pentraxin (PF00354). Catalytic activities exhibited by members of GH18 are chitinase and endo-β-N-acetylglucosaminidase. The latter activity would be appropriate for mucin breakdown and is

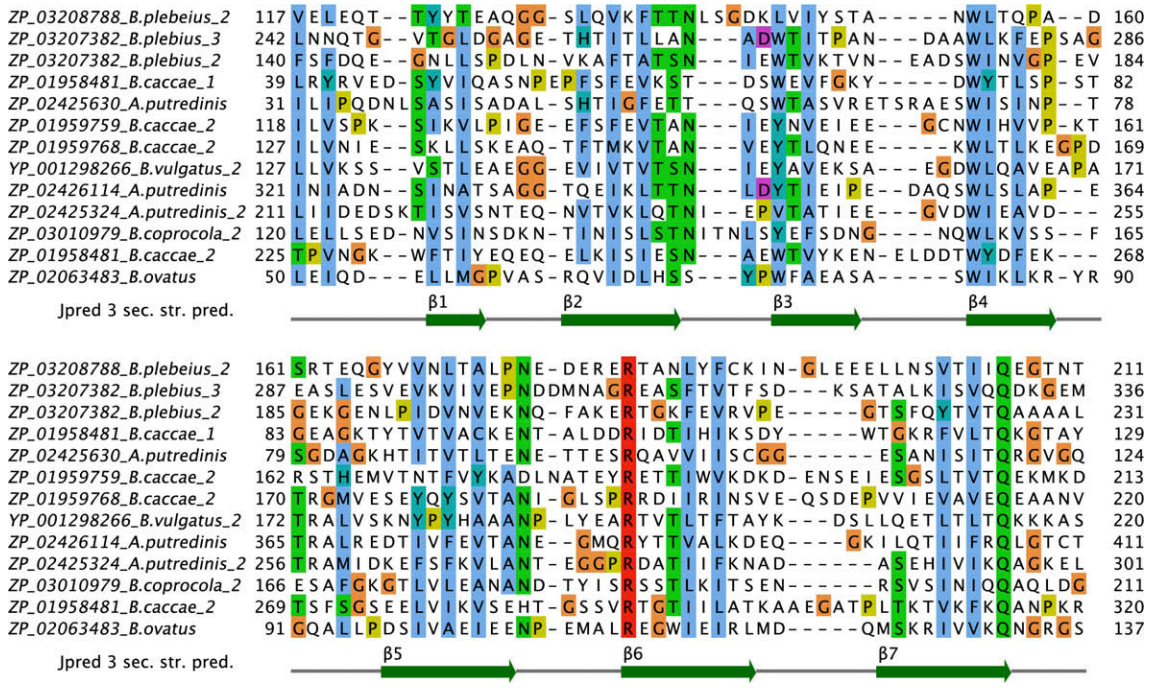

Fig. 2. Alignment of 13 minimally redundant BACON sequences. Each is labelled with NCBI accession code and abbreviated species name. Where a number follows an underscore it is that numbered BACON instance in a protein containing multiple examples. Jpred 3 [49] secondary structure prediction is shown beneath the alignment. The figure was made with Jalview 2 [10].

Table 1
Species distribution of the BACON domain.

Group	Species name	Number of proteins containing BACON domain	Number of BACON domains
Bacteroides species	*B. caccae*	13	18
	B. coprocola	6	15
	B. dorei	1	1
	B. eggerthii	9	11
	B. finegoldii	8	8
	B. forsythus	1	1
	B. fragilis[a]	3	5
	B. intestinalis	4	5
	B. ovatus	16	18
	B. plebeius	13	17
	B. stercoris	7	7
	B. thetaiotaomicron[a]	17	25
	B. uniformis	6	9
	B. vulgatus[a]	3	7
Others in the Bacteriodetes phylum	*Alistipes putredinis*	11	12
	Flavobacteriales bacterium	1	1
	Flavobacterium johnsoniae[a]	1	1
	Geobacter uraniireducens[a]	1	1
	Parabacteroides johnsonii	1	3
	Tenacibaculum sp. MED152	2	2
Others	*Akkermansia muciniphila*[a]	2	2
	Beggiatoa	2	2
	Chloroflexus aggregans[a]	1	1
	Desulfococcus oleovorans	1	1
	Geobacter uraniireducens	1	1
	Solibacter usitatus[a]	4	9
	Verrucomicrobiae bacterium DG1235	3	11
Uncultured/ unidentified		6	8

[a] Genome completed at the time of analysis.

one of the enzyme activities up-regulated in *B. thetaiotamicron* when mucin degradation is signalled [30]. Pentraxin domains are part of a Pfam clan of distantly homologous domains whose predominant common molecular activity is carbohydrate recognition. The second *A. muciniphila*-specific domain architecture sees BACON combined with presumed carbohydrate-binding PA14 and catalytic sulphatase (PF00884) domains. The large homologous group of sulphatases are frequently found in enzymes that degrade glycosaminoglycans [34]. Within this carbohydrate class, mucins are often sulphated [35] and this sulphation is thought to offer protection against mucin-degrading bacteria [36] so combination of a sulphatase enzyme with a mucin-recognising domain could easily be advantageous for the bacterium. Again, *B. thetaiotamicron* produces a sulphatase when in mucin-degrading mode [30].

Few domains have been experimentally characterized as mucin binding, but it is worth noting that one strong candidate, MucBP (PF06458), like BACON exhibits tandem domain duplications. This presumably represents a general strategy for enhancing affinity for structurally repetitive ligands like carbohydrates.

Finally, it is noteworthy that two domain architectures found in *Bacteriodes* species combine one or two BACON domains with a F5_F8_type_C domain and a domain distantly related to Enhancin (Fig. 1). Although these latter domains bear only around 15% sequence identity to the true enhancin consensus, this resemblance is stronger than any similarity to other protease families. It is therefore interesting to note that viral enhancins are insect mucin-degrading enzymes [37]. It may be that their distant relatives in *Bacteroides* species have the same role towards human host mucin.

Although several lines of evidence point to mucin as a ligand of BACON domains, it is clear that not all BACON domains would share that proposed specificity. Thus, the starting point for this study, the metagenomics-derived protein (NCBI accession CAJ19151; [20]), combines BACON with a cellulase domain. Similarly, in the marine bacterium *Tenacibaculum* sp. MED152, for example, BACON is found in combination with an alginate lyase

domain (Fig. 1) suggesting that it would act as an accessory domain to bind substrate alginate.

Supporting the link to carbohydrate-binding, there is a remarkable correspondence between conserved BACON residues and amino-acids known to be over-represented at carbohydrate-binding sites [38]. Trp and Arg are known to be the two amino acids with the highest propensity for occurrence at these sites: BACON conserved positions contain one of each. Furthermore, the remaining conserved BACON residues Asn and Gln fall into a second category of strongly over-represented residues, polar amino acids with planar side chain groups [38].

The lack of detectable similarity suggestive of homology between BACON and any known domain or protein structure obviously precluded any comparative modelling. The size and all-β structure of BACON would be consistent with its having an immunoglobulin-like topology, as seen for many bacteria protein families ([39]; http://pfam.sanger.ac.uk/clan/CL0159), yet extensive testing failed to demonstrate any significant similarity between BACON sequences and any known family or structure. *Ab initio* modelling using ROSETTA [40,41] or I-TASSER [42,43] (data not shown) failed to produce strongly reliable predictions. However, it was interesting to note that the avidin-like β-barrel architecture commonly found in the models places the set of conserved residues mentioned above were found in close proximity, consistent with their constituting a binding site.

In conclusion, we have discovered and described a novel protein domain which appears to be particularly associated with the Bacteriodetes phylum. Various lines of evidence suggest that the BACON domain binds carbohydrate-containing molecules and in many cases perhaps, more specifically, mucin. While this predicted function remains to be confirmed, medical interest in mucin-binding proteins [44,45] and in *Bacteriodes* species as an important component (20–40%; [46]) of a variable intestinal microbiome [47,48] makes it an attractive target for future study.

References

[1] Levitt, M. (2009) Nature of the protein universe. Proc. Natl. Acad. Sci. USA 106, 11079–11084.

[2] Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C.S., Li, H., Mashiyama, S.T., Joachimiak, M.P., van Belle, C., Chandonia, J.M., Soergel, D.A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B.J., Bafna, V., Friedman, R., Brenner, S.E., Godzik, A., Eisenberg, D., Dixon, J.E., Taylor, S.S., Strausberg, R.L., Frazier, M. and Venter, J.C. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 5, e16.

[3] Sammut, S.J., Finn, R.D. and Bateman, A. (2008) Pfam 10 years on: 10,000 families and still growing. Brief Bioinform. 9, 210–219.

[4] Furnham, N., Garavelli, J.S., Apweiler, R. and Thornton, J.M. (2009) Missing in action: enzyme functional annotations in biological databases. Nat. Chem. Biol. 5, 521–525.

[5] Rigden, D.J. (2008) The histidine phosphatase superfamily: structure and function. Biochem. J. 409, 333–348.

[6] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

[7] Duan, C.J., Xian, L., Zhao, G.C., Feng, Y., Pang, H., Bai, X.L., Tang, J.L., Ma, Q.S. and Feng, J.X. (2009) Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. J. Appl. Microbiol. 107, 245–256.

[8] Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 35, D5–D12.

[9] Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

[10] Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. and Barton, G.J. (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191.

[11] Bujnicki, J.M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001) Structure prediction meta server. Bioinformatics 17, 750–751.

[12] Soding, J., Biegert, A. and Lupas, A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244–W248.

[13] Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Krylov, D., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Thanki, N., Yamashita, R.A., Yin, J.J., Zhang, D. and Bryant, S.H. (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 35, D237–D240.

[14] Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. and Bateman, A. (2008) The Pfam protein families database. Nucleic Acids Res. 36, D281–D288.

[15] Juncker, A.S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. and Krogh, A. (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. 12, 1652–1662.

[16] Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2, 953–971.

[17] Bendtsen, J.D., Nielsen, H., Widdick, D., Palmer, T. and Brunak, S. (2005) Prediction of twin-arginine signal peptides. BMC Bioinform. 6, 167.

[18] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580.

[19] Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform. 9, 299–306.

[20] Ferrer, M., Golyshina, O.V., Chernikova, T.N., Khachane, A.N., Reyes-Duarte, D., Santos, V.A., Strompl, C., Elborough, K., Jarvis, G., Neef, A., Yakimov, M.M., Timmis, K.N. and Golyshin, P.N. (2005) Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. Environ. Microbiol. 7, 1996–2010.

[21] Walter, J., Mangold, M. and Tannock, G.W. (2005) Construction, analysis, and beta-glucanase screening of a bacterial artificial chromosome library from the large-bowel microbiota of mice. Appl. Environ. Microbiol. 71, 2347–2354.

[22] Bartlett, G.J., Porter, C.T., Borkakoti, N. and Thornton, J.M. (2002) Analysis of catalytic residues in enzyme active sites. J. Mol. Biol. 324, 105–121.

[23] Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 37, D233–D238.

[24] Henrissat, B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. 280 (Pt 2), 309–316.

[25] Hudson, E.R., Pan, D.A., James, J., Lucocq, J.M., Hawley, S.A., Green, K.A., Baba, O., Terashima, T. and Hardie, D.G. (2003) A novel domain in AMP-activated protein kinase causes glycogen storage bodies similar to those seen in hereditary cardiac arrhythmias. Curr. Biol. 13, 861–866.

[26] Rigden, D.J., Mello, L.V. and Galperin, M.Y. (2004) The PA14 domain, a conserved all-beta domain in bacterial toxins, enzymes, adhesins and signaling molecules. Trends Biochem. Sci. 29, 335–339.

[27] Simpson, D.L., Rosen, S.D. and Barondes, S.H. (1974) Discoidin, a developmentally regulated carbohydrate-binding protein from Dictyostelium discoideum. Purification and characterization. Biochemistry 13, 3487–3493.

[28] Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R. and Gordon, J.I. (2007) The human microbiome project. Nature 449, 804–810.

[29] Comstock, L.E. (2009) Importance of glycans to the host-bacteroides mutualism in the mammalian intestine. Cell. Host Microbe 5, 522–526.

[30] Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.H., Westover, B.P., Weatherford, J., Buhler, J.D. and Gordon, J.I. (2005) Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science 307, 1955–1959.

[31] Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C., Hooper, L.V. and Gordon, J.I. (2003) A genomic view of the human-bacteroides thetaiotaomicron symbiosis. Science 299, 2074–2076.

[32] Backhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A. and Gordon, J.I. (2005) Host-bacterial mutualism in the human intestine. Science 307, 1915–1920.

[33] Derrien, M., Vaughan, E.E., Plugge, C.M. and de Vos, W.M. (2004) *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. Int. J. Syst. Evol. Microbiol. 54, 1469–1476.

[34] Peters, C., Schmidt, B., Rommerskirch, W., Rupp, K., Zuhlsdorf, M., Vingron, M., Meyer, H.E., Pohlmann, R. and von Figura, K. (1990) Phylogenetic conservation of arylsulfatases. CDNA cloning and expression of human arylsulfatase B. J. Biol. Chem. 265, 3374–3381.

[35] Nieuw Amerongen, A.V., Bolscher, J.G., Bloemena, E. and Veerman, E.C. (1998) Sulfomucins in the human body. Biol. Chem. 379, 1–18.

[36] Corfield, A.P., Wagner, S.A., O'Donnell, L.J., Durdey, P., Mountford, R.A. and Clamp, J.R. (1993) The roles of enteric bacterial sialidase, sialate O-acetyl esterase and glycosulfatase in the degradation of human colonic mucin. Glycoconj. J. 10, 72–81.

[37] Wang, P. and Granados, R.R. (1997) An intestinal mucin is the target substrate for a baculovirus enhancin. Proc. Natl. Acad. Sci. USA 94, 6977–6982.

[38] Malik, A. and Ahmad, S. (2007) Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. BMC Struct. Biol. 7, 1.

[39] Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R. and Bateman, A. (2010) The Pfam protein families database. Nucleic Acids Res. 38, D211–D222.

[40] Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C. and Baker, D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34, 82–95.
[41] Simons, K.T., Kooperberg, C., Huang, E. and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. 268, 209–225.
[42] Wu, S., Skolnick, J. and Zhang, Y. (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. 5, 17.
[43] Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinform. 9, 40.
[44] Bumbaca, D., Littlejohn, J.E., Nayakanti, H., Lucas, A.H., Rigden, D.J., Galperin, M.Y. and Jedrzejas, M.J. (2007) Genome-based identification and characterization of a putative mucin-binding protein from the surface of Streptococcus pneumoniae. Proteins 66, 547–558.
[45] Kankainen, M., Paulin, L., Tynkkynen, S., von Ossowski, I., Reunanen, J., Partanen, P., Satokari, R., Vesterlund, S., Hendrickx, A.P., Lebeer, S., De Keersmaecker, J., Vanderleyden, J., Hamalainen, T., Laukkanen, S., Salovuori, N., Ritari, J., Alatalo, E., Korpela, R., Mattila-Sandholm, T., Lassig, A., Hatakka, K., Kinnunen, K.T., Karjalainen, H., Saxelin, M., Laakso, K., Surakka, A., Palva, A., Salusjarvi, T., Auvinen, P. and de Vos, W.M. (2009) Comparative genomic analysis of Lactobacillus rhamnosus GG reveals pili containing a human-mucus binding protein. Proc. Natl. Acad. Sci. USA 106, 17193–17198.
[46] Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D. and Gordon, J.I. (2005) Obesity alters gut microbial ecology. Proc. Natl. Acad. Sci. USA 102, 11070–11075.
[47] Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R. and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444, 1027–1031.
[48] Tsai, F. and Coyle, W.J. (2009) The microbiome and obesity: is obesity linked to our gut flora? Curr. Gastroenterol. Rep. 11, 307–313.
[49] Cole, C., Barber, J.D. and Barton, G.J. (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res. 36, W197–W201.