

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
3 June 2010 (03.06.2010)

(10) International Publication Number
WO 2010/060964 A2

(51) International Patent Classification:
C12N 9/42 (2006.01) C12P 7/10 (2006.01)
C12N 15/62 (2006.01)

Gemenos (FR). **DEDIEU, Luc** [FR/FR]; 75 boulevard Rabateau, Patio du Rouet, F-13008 Marseille (FR).

(21) International Application Number:
PCT/EP2009/065922

(74) Agents: **PAEMEN, Liesbet Rita Johanna** et al.; De Clercq Brants & Partners, E. Gevaertdreef 10a, B-9830 Sint-Martens-Latem (BE).

(22) International Filing Date:
26 November 2009 (26.11.2009)

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(25) Filing Language: English
(26) Publication Language: English
(30) Priority Data:
08291120.7 28 November 2008 (28.11.2008) EP

(71) Applicants (for all designated States except US): **TOTAL S.A.** [FR/FR]; 2 Place Jean Millier, La Défense 6, F-92 400 Courbevoie (FR). **LE CENTRE NATIONAL DE RECHERCHE SCIENTIFIQUE (CNRS)** [FR/FR]; 3 rue Michel-Ange, F-75794 Paris Cedex 16 (FR). **L'UNIVERSITE DE LA MEDITERRANEE** [FR/FR]; Jardin du Pharo, 58 Boulevard Charles-Livon, F-13284 Marseille Cedex 7 (FR). **L'UNIVERSITE DE PROVENCE** [FR/FR]; 3 Place Victor Hugo, F-13331 Marseille Cedex 3 (FR). **L'INSTITUT NATIONAL DES SCIENCES APPLIQUEES (INSA)** [FR/FR]; 135 Avenue de Rangueil, F-31077 Toulouse Cedex 4 (FR).

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(72) Inventors; and

Declarations under Rule 4.17:

(75) Inventors/Applicants (for US only): **FIERROBE, Henri-Pierre** [FR/FR]; 31 chemin Joseph Aiguier, F-13009 Marseille (FR). **CHANAL-VIAL, Angélique** [FR/FR]; 3 place Eugène Bertrand, F-13012 Marseille (FR). **MOLINIER, Anne-Laure** [FR/FR]; 18 avenue du Marechal Foch, F-13004 Marseille (FR). **TARDIF, Chantal** [FR/FR]; 3 rotonde Roussargue, F-13420

— of inventorship (Rule 4.17(iv))

Published:

- without international search report and to be republished upon receipt of that report (Rule 48.2(g))
- with sequence listing part of description (Rule 5.2(a))



WO 2010/060964 A2

(54) Title: CELLULASE CEL5H RELATED REAGENTS AND THEIR USE IN MICROORGANISMS

(57) Abstract: The invention relates to applications of the cellulase Cel5H of *Saccharophagus degradans* and its homologues, functional fragments and/or variants and engineered forms thereof, in the context of recombinant, more particularly solventogenic microorganisms, more particularly *C. acetobutylicum*. The invention also characterises a novel domain of the Cel5H cellulase with a putative cellulose-binding module function, and its uses in chimeric proteins for depolymerisation of cellulose containing substrates.

CELLULASE CEL5H RELATED REAGENTS AND THEIR USE IN MICROORGANISMS**FIELD OF THE INVENTION**

The present invention relates to the area of biotechnology and genetic engineering and particularly concerns strategies for utilisation of cellulosic and lignocellulosic materials by means of suitable enzymatic agents and microorganisms expressing such. More specifically, the invention relates to applications of the cellulase Cel5H of *Saccharophagus degradans* and its homologues, functional fragments and/or variants in the context of solventogenic microorganisms, as well as to novel domains and reagents derived from Cel5H and homologues thereof.

BACKGROUND OF THE INVENTION

Cellulosic and lignocellulosic materials are major constituents of plant biomass, and cellulose polymers found therein can provide a significant source of glucose or other fermentable mono- and oligosaccharides that can in turn be metabolised by solventogenic microorganisms to produce useful solvents, such as ethanol, acetone or butanol.

Cellulose polymers can be hydrolysed by cellulose-depolymerising enzymes commonly known as cellulolytic enzymes or cellulases. For example, hydrolysis of native cellulose mainly involves four cellulase types: cellobiohydrolase (1,4- β -D-glucan cellobiohydrolase, EC 3.2.1.91), endo- β -1,4-glucanase (endo-1,4- β -D-glucan 4-glucanohydrolase, EC 3.2.1.4), endo-processive cellulases (EC 3.2.1.4./3.2.1.91), and β -glucosidase (EC 3.2.1.21). Cellulases and related enzymes have been widely utilised in various areas of biotechnology including in food, beer, wine, animal feeds, textile production and laundering, pulp and paper industry, agricultural industry and others (for review see Bhat 2000. Biotechnical Advances 18: 355-383).

Cellulases can vary in their characteristics, such as *inter alia* they may act as endoglucanases or as processive exoglucanases, they may generate monomers or oligomers of various lengths, they may have contrasting abilities to hydrolyse distinct cellulose forms such as crystalline cellulose, semi-crystalline cellulose, amorphous cellulose or hemicellulose, they may further display different strength of binding to cellulose substrates, different kinetic parameters, etc. Therefore, significant effort needs to be

invested into further characterisation of cellulases, so as to identify functional parts or domains thereof which may underlie the interesting properties of these enzymes. Such functional domains can be advantageously combined with other cellulases or domains thereof to generate chimeric enzymes with desired activities.

5 Further, the recombinant expression of cellulases in solventogenic microorganisms – to allow the production of useful solvents including *inter alia* ethanol by these microorganisms directly from cellulose-containing materials – is not yet satisfactorily advanced. To achieve further improvements, cellulases particularly suitable for or advantageous in such applications need to be recognised, characterised and selected.

10 10 *Saccharophagus degradans* strain 2-40, previously known as *Microbifulber degradans* strain 2-40 and deposited with the American Type Culture Collection under accession number ATCC 43961, is a marine γ -proteobacterium able to degrade at least ten different complex polysaccharides including *inter alia* cellulose (Andrykovitch & Marx 1988. Appl Environ Microbiol 54: 3-4; Ensor *et al.* 1999. J Ind Microbiol Biotechnol 23: 123-126).

15 15 The *S. degradans* genome has been sequenced and its cellulase system identified, including among many others a cellulase denoted as Cel5H (Taylor *et al.* 2006. J Bacteriol 188: 3849-61). Further, US2007/0292929 generically contemplates recombinant micro-organisms including ethanologenic bacteria expressing the cellulose degrading proteins from *S. degradans*. However, only the activity of enzymes expressed by *S. degradans* is 20 demonstrated and it is unclear whether any of these enzymes would function appropriately when expressed in ethanologenic bacteria..

SUMMARY OF THE INVENTION

25 The inventors have performed extensive characterisation of the structure and activity of the Cel5H cellulase of *S. degradans* strain 2-40 and have conclusively determined that Cel5H displays an elevated activity on crystalline cellulose. Therefore, Cel5H is advantageous for heterologous production in solvent-producing (solventogenic) microorganisms, more particularly in ethanologenic microorganisms such as ethanologenic bacteria of *Clostridium* species including *C. acetobutylicum*. Hence, expression (and optionally secretion) of Cel5H 30 in these microorganisms enables their growth on cellulose substrates comprising or

enriched in crystalline, semi-crystalline or amorphous cellulose, thereby allowing direct production of useful solvents such as ethanol from cellulose-containing substrates.

Further data generated by the inventors has indicated that Cel5H may act as an exoglucanase or endoprocessive cellulase rather than an endoglucanase as would be expected from its glycoside hydrolase family 5 (GH5) catalytic domain. Also, Cel5H appears to mainly liberate glucose, cellobiose and cellotriose from cellulose. Given the seemingly processive nature of Cel5H and the generation of low complexity cellulose derivatives thereby, its recombinant expression in solventogenic bacteria can be highly advantageous. For example, the products of Cel5H action may either be directly utilised by solventogenic microorganisms or require minimal action by further heterologous cellulase(s), thereby simplifying the design of the recombinant microorganisms and increasing the effectiveness of the process.

The invention integrates the above unexpected realisations in its various aspects.

Hence, an aspect provides a recombinant microorganism expressing the Cel5H polypeptide of *Saccharophagus degradans* strain 2-40 or a homologue thereof, or a functional fragment and/or variant of said Cel5H polypeptide or of said homologue. More particularly, the invention provides such recombinant micro-organisms which are capable of degrading cellulose or cellulose-containing material by said expressed Cel5H polypeptide.

According to particular embodiments the microorganism is a bacterium, more particularly a solventogenic bacterium.

In particular embodiments, the recombinant microorganism comprises a recombinant nucleic acid molecule encoding the Cel5H polypeptide or homologue thereof, or encoding the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, operably linked to regulatory sequences which allow for expression in said microorganism.

In further particular embodiments, the recombinant microorganism produces one or more solvents, fuels and/or chemical intermediates, more particularly solvents chosen from ethanol, acetone, butanol, propionic acid, butyric acid, ether and glycerine.

In further embodiments, the recombinant solventogenic microorganism may produce, or may be engineered to produce, at least or mainly ethanol. The industrial importance of ethanol is rapidly increasing largely due to its utility as an environmentally acceptable fuel.

Hence, in an embodiment, the recombinant solventogenic microorganism may be an ethanologenic microorganism.

In particular embodiments, the recombinant microorganism may be a bacterium. In an embodiment, the recombinant solventogenic bacterium may be an ethanologenic 5 bacterium. Alternatively, the recombinant solventogenic microorganism may be yeast, more particularly ethanologenic yeast.

For example, said bacterium may be Gram-positive, such as for example a bacterium of the *Clostridium* species. The invention thus also includes introducing a cellulase originating 10 from a Gram-negative bacterium into a Gram-positive bacterium, which has not been previously suggested. Alternatively, said bacterium may be Gram-negative.

In particular embodiments, the recombinant solventogenic and preferably ethanologenic microorganism is a *Clostridium* species, preferably *Clostridium acetobutylicum*. These bacteria are particularly proficient as solvent- and particularly ethanol-producing bacteria.

In an embodiment, the Cel5H polypeptide or homologue thereof, or the functional fragment 15 and/or variant of said Cel5H polypeptide or of said homologue, may be secreted by said microorganism. So-secreted polypeptides can advantageously act directly on and thereby depolymerise or otherwise interact with or alter cellulose polymers to which the microorganism is exposed. However, embodiments where the polypeptides are expressed intracellularly are also contemplated. For example, so-expressed polypeptides may act on 20 cellulose polymers otherwise internalised by the microorganism or may become released upon lysis of at least a fraction of the microorganisms.

The recombinant microorganism envisaged herein may comprise a recombinant nucleic acid molecule encoding the Cel5H polypeptide or homologue thereof, or encoding the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, operably 25 linked to a secretion signal sequence, optionally combined with one or more additional modules, which allows for secretion in said microorganism, and operably linked to regulatory sequences which allow for expression in said microorganism.

As noted, the invention also teaches homologues of Cel5H and functional fragments and/or variants of said homologues, and their uses in recombinant microorganisms, more 30 particularly solventogenic microorganisms. Preferably, a Cel5H homologue as contemplated herein may comprise a domain homologous to the DZ domain of Cel5H.

Particular embodiments provide the recombinant microorganism as taught herein, wherein said homologue of the Cel5H polypeptide is the ACLA polypeptide of *Pseudomonas* sp. ND137.

Further, as realised by the inventors, the domain structure of the native Cel5H polypeptide can be outlined as GH5-PSL-CBM6-EPR-DZ, wherein GH5 stands for its glycoside hydrolase family 5 domain, PSL for the polyserine linker, CBM6 for carbohydrate-binding module family 6 domain, EPR for the glutamic acid-proline-rich region and, without being limited to this interpretation, DZ represents a C-terminal domain identified by the present inventors as a putative carbohydrate-binding module.

10 Hence, in particular embodiments the recombinant microorganism as taught herein may express a functional fragment comprising one or more domains chosen from or corresponding to the GH5 domain, the CBM6 domain and the DZ domain of Cel5H. The presence of one or more said domains in the fragment can endow the latter with the functionalities attributed to the respective domains.

15 The invention further contemplates engineered forms of the Cel5H polypeptide, its homologues, fragments and/or variants, fused to one or more heterologous domains which can supply further functions and activities useful in carbohydrate polymer metabolism, and particularly cellulose metabolism, such as cellulose-depolymerising, cellulose-binding, cellulosome-forming or other activities. Hence, particular embodiments provide recombinant

20 solventogenic microorganisms as taught herein, wherein the Cel5H polypeptide or homologue thereof, or the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, is fused to one or more domains heterologous to the Cel5H polypeptide or homologue thereof, chosen from a glycoside hydrolase (GH) catalytic domain, a carbohydrate-binding module (CBM) domain, a cohesin-binding domain such as a dockerin domain, and a hydrophilic (X module) domain of a cellulosomal scaffoldin protein.

25 Additionally or alternatively, the Cel5H polypeptide or homologue thereof, or the functional fragment and/or variant of said Cel5H polypeptide or of said homologue is fused to a heterologous or its natural signal peptide.

Further, the invention also contemplates advantages of co-expressing the Cel5H or related polypeptides with other polypeptides or enzymes, preferably recombinant polypeptides or enzymes, useful in carbohydrate polymer metabolism, and particularly cellulose metabolism, such as with cellulose-depolymerising, cellulose-binding, cellulosome-forming

(such as, e.g., scaffoldin) or other polypeptides. Such co-expression, and optionally and particularly co-secretion, can provide for additive or complementary activities and functions, leading to more efficient depolymerisation and utilisation of cellulose by the present microorganisms.

5 Further embodiments thus provide recombinant microorganisms as taught herein co-expressing and optionally co-secreting the Cel5H polypeptide or homologue thereof, or the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, with one or more polypeptides (also referred to herein as co-expressed polypeptides), preferably recombinant polypeptides participating in carbohydrate polymer metabolism, and
10 particularly cellulose metabolism, particularly with one or more enzymes capable of degrading lignocellulosic material, more particularly with one or more glycoside hydrolases, even more particularly with one or more cellulases.

In particular embodiments, the catalytic action of said one or more co-expressed polypeptides, such as co-expressed enzymes, more particularly co-expressed cellulases,
15 may be additive or complementary, preferably complementary, to the enzymatic activity of the Cel5H polypeptide or homologue thereof, or the functional fragment and/or variant of said Cel5H polypeptide or of said homologue. By means of example and not limitation, a first cellulase may be considered as having activity complementary to a second cellulase if the first and second cellulases act preferentially on distinct substrates (such as, e.g.,
20 crystalline cellulose, semi-crystalline cellulose, amorphous cellulose or hemicellulose), or if the first and second cellulases produce distinct products (e.g., distinct populations of sugar monomers and/or oligomers), or if the first cellulase acts preferentially on reaction products of the second cellulase or *vice versa, etc.*

In further embodiments, the one or more co-expressed cellulases may be chosen from
25 family-5, 6, 8, 9 and 48 cellulases. In further particular embodiments, the one or more co-expressed cellulases may be chosen from *Clostridium cellulolyticum* cellulases Cel48F, Cel9G, Cel9R, Cel9P, Cel9E, Cel9H, Cel9J, Cel9M, Cel8C, Cel5N and Cel5A, and
Saccharophagus degradans strain 2-40 cellulases Cel9A, Cel9B, Cel5J, Cel5I, Cel5F,
30 Cel5D, Cel5B, Cel9G, Cel5E, Cel5A, Cel5C and Cel6A, and functional fragments and/or variants of any one of said cellulases.

Further embodiments provide recombinant microorganisms as taught herein, more particularly solventogenic recombinant microorganisms wherein the Cel5H polypeptide or

homologue thereof, or the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, and optionally the one or more co-expressed polypeptides, more particularly one or more co-expressed cellulases as taught above, are comprised in a hybrid and/or covalent cellulosome or minicellulosome (both encompassed throughout this 5 specification by the generic reference to cellulosome). Cellulosomes can provide for supramolecular organisation of *inter alia* cellulose-binding and cellulose-depolymerising activities, thereby achieving greater efficiency of carbohydrate polymer metabolism, particularly cellulose metabolism.

In a further aspect, the invention provides methods for the degradation of a substance 10 comprising cellulose, such as lignocellulosic or cellulosic material or biomass, comprising contacting said substance with the recombinant microorganism as taught herein. In particular embodiments, the substance comprises or is enriched for crystalline cellulose.

A related aspect provides methods for producing a solvent, fuel or chemical intermediate 15 from a substance comprising cellulose, such as lignocellulosic or cellulosic material or biomass, comprising treating said substance with one or more of the microorganisms as taught herein. In particular embodiments, the substance comprises or is enriched for crystalline cellulose. In most particular embodiments, the solvent is ethanol and the microorganism is an ethanologenic microorganism.

In particular embodiments, the methods of the invention comprises growing recombinant 20 microorganisms as described herein on cellulose substrates comprising or enriched in crystalline, semi-crystalline or amorphous cellulose, thus ensuring direct production of useful solvents such as ethanol from cellulose-containing substrates.

It shall also be appreciated that while in the foregoing aspects and embodiments the invention has been disclosed primarily in connection with the recombinant microorganisms, 25 the invention also encompasses aspects relating to recombinant means and reagents useful for obtaining the recombinant microorganisms, methods for obtaining the recombinant microorganisms using such recombinant means and reagents, methods for expressing the desired polypeptides in the recombinant microorganisms, as well as to the expressed polypeptides and combinations thereof *per se* and methods for preparation 30 thereof.

Accordingly, a further aspect provides isolated Cel5H polypeptides of *Saccharophagus degradans* strain 2-40 or a homologue thereof, or a functional fragment and/or variant of

said Cel5H polypeptide or of said homologue. Another aspect provides recombinant nucleic acid molecules encoding the Cel5H polypeptide or homologue thereof, or encoding the functional fragment and/or variant of said Cel5H polypeptide or of said homologue.

Further embodiments of this aspect of the invention provide recombinant nucleic acid molecules encoding the Cel5H polypeptide or homologue thereof, or encoding the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, operably linked to regulatory sequences which allow for expression in a host microorganism of interest. In particular embodiments, said host is a solventogenic microorganism. In more particular embodiments, said host is an ethanologenic microorganism as taught elsewhere in this specification. In particular embodiments of the recombinant nucleic acid molecules provided herein, the nucleic acid sequence encoding the Cel5H polypeptide or homologue thereof, or encoding the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, is adapted to a host micro-organism codon bias,

Further embodiments provide recombinant nucleic acid molecules encoding the Cel5H polypeptide or homologue thereof, or encoding the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, operably linked to a secretion signal sequence which allows for secretion in a host microorganism of interest, particularly in a recombinant microorganism, more particularly in a solventogenic and more preferably ethanologenic microorganism as taught elsewhere in this specification.

Further embodiments provide recombinant nucleic acid molecules encoding the Cel5H polypeptide or homologue thereof, or encoding the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, operably linked to a secretion signal sequence which allows for secretion in a host microorganism of interest, and operably linked to regulatory sequences which allow for expression in the host microorganism of interest, particularly in a solventogenic and more particularly in an ethanologenic microorganism as taught elsewhere in this specification.

A further aspect is a vector, such as for instance a cloning vector, a shuttle vector and/or an expression vector, comprising one or more of the recombinant nucleic acids as disclosed above.

A further aspect provides host microorganisms of interest transformed with one or more of the recombinant nucleic acids or with a vector as described above. In an embodiment, the microorganism may be a bacterium, more particularly a bacterium chosen from *Escherichia*

coli, *Salmonella typhimurium*, *Serratia marcescens*, *Salmonella typhimurium* and *Bacillus subtilis*. Such bacteria are particularly suitable for overproduction of recombinant polypeptides, for example for purposes of purification. In particular embodiments, the host microorganism is a solventogenic and preferably an ethanogenic microorganism as taught 5 elsewhere in this specification. A further aspect provides a cell lysate or cell extract directly obtained or obtainable from so-transformed microorganisms.

Hence, also provided are methods for obtaining recombinant microorganisms according to the invention, preferably solventogenic and more preferably ethanogenic microorganisms as taught elsewhere in this specification, comprising transforming a host microorganism 10 with a recombinant nucleic acid or vector as taught herein.

Further provided are methods for production, and optionally and most particularly secretion, of one or more Cel5H polypeptides or homologues thereof, or functional fragments and/or variants of said Cel5H polypeptides or of said homologues, comprising transforming a host microorganism with one or more of the respective recombinant nucleic acids or vectors as 15 taught herein, and culturing or maintaining the so-transformed microorganisms under conditions effective to cause expression of the respective polypeptides. Advantageously, the respective polypeptides are produced in a biologically active form. Optionally, the methods may comprise a step of isolating the respective polypeptides.

Hence, also contemplated is the use of the recombinant nucleic acids, vectors or host cells 20 as disclosed herein for the production, and optionally and particularly secretion of Cel5H polypeptide or homologue thereof, or functional fragment and/or variant of said Cel5H polypeptide or of said homologue. Preferably, the respective polypeptides are produced in a biologically active form.

In particular embodiments of the above polypeptides, recombinant nucleic acids, vectors, 25 transformed host microorganisms, methods and uses are characterized by one or more of the following features:

- the Cel5H homologue comprises a domain homologous to the DZ domain of Cel5H; and/or
- the Cel5H homologue is the ACLA polypeptide of *Pseudomonas* sp. ND137; and/or
- 30 - the functional fragment comprises one or more domains chosen from or corresponding to the GH5 domain, the CBM6 domain and the DZ domain of Cel5H; and/or

- the one or more Cel5H polypeptides or homologues thereof, or the functional fragments and/or variants of the one or more Cel5H polypeptides or said homologue, is fused to one or more domains heterologous to the Cel5H polypeptide or homologue thereof, chosen from a GH catalytic domain, a CBM domain, a cohesin-binding domain such as a dockerin domain, and an X module domain of a cellulosomal scaffoldin protein; and/or
- the one or more Cel5H polypeptides or homologues thereof, or the functional fragment and/or variant of said one or more Cel5H polypeptides or of said homologues, may be co-expressed and optionally co-secreted with one or more polypeptides, preferably recombinant polypeptides or enzymes, participating in carbohydrate polymer metabolism, particularly cellulose metabolism, preferably with one or more enzymes capable of degrading lignocellulosic material, more preferably with one or more heterologous cellulases as taught herein; and/or
- the one or more Cel5H polypeptides or homologues thereof, or the functional fragment and/or variant of said Cel5H polypeptides or of said homologues, and optionally the one or more co-expressed polypeptides, more particularly co-expressed cellulases, as taught above, are comprised in a hybrid and/or covalent cellulosome.

Hence, an aspect provides isolated compositions comprising the Cel5H polypeptide or a homologue thereof, or a functional fragment and/or variant of said Cel5H polypeptide or of said homologue, and further comprising one or more polypeptides participating in carbohydrate polymer metabolism, and particularly cellulose metabolism, preferably one or more enzymes capable of degrading lignocellulosic material, more preferably one or more heterologous cellulases as taught herein. In particular embodiments of said composition, the recited components are present as discrete or free polypeptides. In other embodiments, the recited polypeptides are comprised in a hybrid and/or covalent cellulosome.

The inventors have further performed a thorough analysis of the structure and organisation of the Cel5H polypeptide and have identified the presence of a new domain not previously identified as such, near the C-terminal end of the Cel5H polypeptide (termed domain DZ hereafter). Without limitation, the inventor's data suggest a putative carbohydrate-binding module (and/or oligomerisation module function) for the DZ domain and indicate that its presence may be at least partly responsible for the good ability of Cel5H to depolymerise crystalline cellulose. Therefore, it is a further realisation of the invention that the DZ domain or portions of the Cel5H polypeptide comprising the DZ domain can be employed in a

variety of enzymatic systems to improve their ability to digest and degrade crystalline cellulose.

Accordingly, a further aspect is an isolated domain (DZ domain) of the Cel5H polypeptide of *Saccharophagus degradans* strain 2-40 from amino acid 496 to amino acid 596 of said 5 Cel5H polypeptide (such as, e.g., in the exemplary, mature Cel5H polypeptide SEQ ID NO: 3 shown in Fig. 1C) or an isolated domain homologous thereto, or a functional fragment and/or variant of said DZ domain or of said homologous domain.

In particular embodiments, a domain homologous to the Cel5H DZ is a domain stretching from amino acid 463 to amino acid 558 of the ACLA polypeptide of *Pseudomonas* sp.

10 ND137 (such as, e.g., in the exemplary, mature ACLA polypeptide SEQ ID NO: 5 shown in Fig. 1E)

Particular embodiments provide for an isolated fragment of the Cel5H polypeptide or of a homologue or variant thereof, having the structure EPR-DZ, CBM6-EPR-DZ or PSL-CBM6- 15 EPR-DZ, wherein EPR, CBM6 and PSL have meanings as explained elsewhere in this specification. Such fragments comprise the DZ domain which provides for its advantageous

functions, and optionally, such fragment may include the Cel5H carbohydrate-binding module CBM6 which may further stimulate the ability to degrade crystalline cellulose or hemicellulosic substrates. Moreover, endogenous linker sequences can provide 20 advantageous means for fusing the DZ and optionally CBM6 domains to other polypeptides while retaining their structure and function.

Hence, particular embodiments provide for an isolated fragment of the ACLA polypeptide comprising a domain homologous to the DZ domain of Cel5H.

Further contemplated are uses of the isolated DZ domain or the functional fragment and/or variant thereof, or of the isolated fragment as taught above, as a carbohydrate-binding module. Such use can endow novel carbohydrate- or cellulose-binding properties on 25 various enzyme systems.

Another aspect relates to chimeric polypeptides comprising an isolated DZ domain most particularly of Cel5H or a functional fragment and/or variant thereof, or comprising an isolated fragment of Cel5H comprising the DZ domain, and further comprising one or more 30 domains heterologous to the Cel5H polypeptide or homologue thereof, chosen from a GH catalytic domain, a CBM domain, a cohesin-binding domain such as a dockerin domain,

and an X module domain of a cellulosomal scaffoldin protein. Such modular combinations allow the generation of novel cellulose-degrading enzymes with useful activity towards crystalline cellulose or hemicellulosic substrates.

Other aspects provide chimeric polypeptides comprising the isolated DZ domain or the 5 functional fragment and/or variant thereof, or an isolated fragment of Cel5H comprising the DZ domain, fused to one or more polypeptides participating in carbohydrate polymer metabolism, particularly cellulose metabolism, preferably to one or more enzymes capable of degrading lignocellulosic material, more preferably to one or more glycoside hydrolases, even more preferably to one or more cellulases, which is (are) heterologous to the Cel5H 10 polypeptide or homologue thereof. Nevertheless, fusions of additional DZ domain(s) with a Cel5H polypeptide or a homologue thereof, or fragments and/or variants thereof as described herein, are also contemplated.

A further aspect provides recombinant nucleic acid molecules encoding the isolated DZ domain or the isolated domain homologous thereto or the functional fragment and/or variant 15 thereof, or encoding an isolated fragment of Cel5H comprising the DZ domain or homologue thereof, or encoding the chimeric polypeptide as described above.

A particular embodiment of this aspect provides a recombinant nucleic acid molecule encoding the isolated DZ domain or the isolated domain homologous thereto or the functional fragment and/or variant thereof, or encoding the isolated fragment of Cel5H comprising the DZ domain or a 20 homologue thereof, or encoding the chimeric polypeptide as described above, operably linked to regulatory sequences which allow for expression in a host microorganism of interest, preferably in a solventogenic and more preferably in an ethanologenic microorganism as taught elsewhere in this specification.

A further embodiment provides a recombinant nucleic acid molecule encoding the isolated 25 DZ domain or the isolated domain homologous thereto or the functional fragment and/or variant thereof, or encoding the isolated fragment of Cel5H comprising the DZ domain or a homologue thereof, or encoding the chimeric polypeptide as described above, operably linked to a secretion signal sequence which allows for secretion in a host microorganism of interest, preferably in a solventogenic and more preferably in an ethanologenic 30 microorganism as taught elsewhere in this specification.

Yet a further embodiment of this aspect provides a recombinant nucleic acid molecule encoding the isolated DZ domain or the isolated domain homologous thereto or the

functional fragment and/or variant thereof, or encoding the isolated fragment of Cel5H comprising the DZ domain or a homologue thereof, or encoding the chimeric polypeptide as described above, operably linked to a secretion signal sequence which allows for secretion in a host microorganism of interest, and operably linked to regulatory sequences which 5 allow for expression in the host microorganism of interest, preferably in a solventogenic and more preferably in an ethanologenic microorganism as taught elsewhere in this specification.

A further aspect provides vectors, such as for instance a cloning vector, a shuttle vector and/or an expression vector, comprising any of the above recombinant nucleic acid 10 molecules.

A further aspect provides a host microorganism of interest transformed with the above described recombinant nucleic acid molecule or vector. In particular embodiments, the microorganism is a bacterium, more preferably a bacterium chosen from *Escherichia coli*, *Salmonella typhimurium*, *Serratia marcescens*, *Salmonella typhimurium* and *Bacillus subtilis*. In particular embodiments, the microorganism may be a solventogenic and is 15 preferably an ethanologenic microorganism as taught elsewhere in this specification. A further aspect provides a cell lysate or cell extract directly obtained or obtainable from so-transformed microorganisms.

In particular embodiments, the isolated DZ domain or the isolated domain homologous 20 thereto or the functional fragment and/or variant thereof, or the isolated fragment of Cel5H comprising the DZ domain or a homologue thereof, or the chimeric polypeptide as described above, and optionally the one or more co-expressed polypeptides involved in carbohydrate polymer metabolism, particularly cellulose metabolism, may be comprised in a hybrid and/or covalent cellulosome.

25 In a further aspect, the invention provides methods for the degradation of a substance comprising cellulose, such as lignocellulosic or cellulosic material or biomass, comprising contacting said substance with the isolated DZ domain or the isolated domain homologous thereto or the functional fragment and/or variant thereof, or with the isolated fragment of Cel5H comprising the DZ domain or a homologue thereof, or with the chimeric polypeptide 30 as described above, or with a recombinant microorganism expressing such. In an embodiment, the substance may comprise or be enriched for crystalline cellulose, semi-crystalline cellulose or amorphous cellulose.

A related aspect provides methods for producing a solvent, a fuel or a chemical intermediate from a substance comprising cellulose, such as lignocellulosic or cellulosic material or biomass, comprising treating said substance with the isolated DZ domain or the isolated domain homologous thereto or the functional fragment and/or variant thereof, or

5 with the isolated fragment of Cel5H comprising the DZ domain or a homologue thereof, or with the chimeric polypeptide as described above, or with a recombinant solventogenic microorganism expressing such. In an embodiment, the substance may comprise or be enriched for crystalline cellulose. In particular embodiments methods are provided for producing a solvent, more particularly ethanol and the microorganism may be an

10 ethanologenic microorganism.

A further aspect provides compositions comprising cellulose degradation products and/or solvents, more particularly ethanol, obtained or obtainable by the methods described herein. Such compositions may be crude cultivation media or (at least) partially enriched and/or purified for the product of the Cel5H polypeptide (or homologue or fragment thereof) or partially enriched and/or purified for the solvent, fuel or chemical intermediate. In particular embodiments, such compositions are characterized by the increased glucose, cellobiose and cellotriose content. In further particular embodiments compositions are provided wherein the content of glucose, cellobiose and/or cellotriose is greater than 10 wt%, greater than 20 wt%, greater than 30 wt%, greater than 40 wt%, greater than 50 wt%,

15 greater than 60 wt%, greater than 70 wt%, greater than 80 wt%, greater than 90 wt% or more.

In further particular embodiments, such compositions are characterized by the high content in solvent, fuel or chemical intermediate of interest. In further particular embodiments compositions are provided wherein the content of solvent, fuel or chemical intermediate is

25 greater than 10 wt%, greater than 20 wt%, greater than 30 wt%, greater than 40 wt%, greater than 50 wt%, greater than 60 wt%, greater than 70 wt%, greater than 80 wt%, greater than 90 wt%, greater than 90 wt%, or between 96% and 99,9 wt%. In particular embodiments the concentration of the solvent, fuel or chemical intermediate in the composition is greater than about 10 g/L, and preferably greater than about 15 g/L.

30 These and other aspects and embodiments of the invention are further explained here below and in the appended claims, as well as illustrated by non-limiting examples.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates exemplary coding nucleic acid sequence (**A**) and amino acid sequence (**B-C**) of native Cel5H polypeptide of *Saccharophagus degradans* strain 2-40 containing (**B**) or not containing (**C**) the N-terminal secretion signal sequence; exemplary amino acid sequence of native ACLA protein of *Pseudomonas* sp. ND137 containing (**D**) or not containing (**E**) the N-terminal secretion signal sequence; exemplary engineered Cel5H-coding nucleic acid sequence (**F**) and amino acid sequence (**G**) adapted to *Clostridium acetobutylicum* codon bias and containing 6xHis codons or a tag at the 3' (C-term) end.

Figure 2 schematically illustrates various fragments and derivatives of Cel5H that were or are engineered for expression, according to particular embodiments of the invention.

Figure 3 illustrates activity of the various forms of Cel5H against CMC at 37°, according to particular embodiments of the invention.

Figure 4 illustrates activity of the various forms of Cel5H against Phosphoric Acid Swollen Cellulose at 37°, according to particular embodiments of the invention.

Figure 5 illustrates activity of the various forms of Cel5H against Avicel PH101 at 37°, according to particular embodiments of the invention.

Figure 6 illustrates comparison of Cel5H avicelase activity with wild type and engineered cellulases from *C. cellulolyticum*, according to particular embodiments of the invention.

Figure 7 illustrates schematic representation of engineered forms of Cel5A from *C. cellulolyticum* comprising added domains of Cel5H, according to particular embodiments of the invention. Legend: see Fig. 2.

Figure 8 illustrates the degradation of Avicel (3.5 g/L of 20mM Tris-maleate pH 6 and 1mM CaCl₂) by different engineered forms of Cel5A from *C. cellulolyticum* comprising added domains of Cel5H, according to particular embodiments of the invention. Legend: see Fig. 2.

Figure 9 shows the degradation of Avicel by Cel5H and at least one other cellulase enzyme according to particular embodiments of the invention.

Figure 10 schematically illustrates a hybrid minicellulosome including Cel5H, according to a particular embodiment of the invention.

Figure 11 shows the degradation of PASC, the consummation of residual cellobiose and the bacterial growth of *C. acetobutylicum*, having its Cel5H signal peptide sequence replaced by the Cel5H signal peptide sequence of *C. cellulolyticum* in a fermentor with 3g/L 2YT-PASC and pH 5.5, according to particular embodiments of the invention.

5 **Figure 12** shows the consummation of residual cellobiose and the bacterial growth of *C. acetobutylicum*, having its Cel5H signal peptide sequence replaced by the Cel5H signal peptide sequence of *C. cellulolyticum* in a fermentor in the presence of 2YT and pH 5.5 but without any PASC, according to particular embodiments of the invention.

10 **Figure 13** is a schematic representation of different constructs according to particular embodiments of the invention. The constructs comprise a polypeptide of interest (cellulase “Cel5H”) fused to a carrier domain and a signal sequence. The carrier domain comprises a carbohydrate binding module (CBM3a) from a cellulosomal scaffolding protein fused to one or two X modules (Xa). Cellulase Cel5H comprises a glycoside hydrolase family 5 domain (“5”), a polyserine linker (“sss”), a carbohydrate-binding module family 6 domain (“6”), a 15 glutamic acid-proline-rich region (“eppv”) and a C-terminal domain identified by the present inventors as a putative carbohydrate-binding module (“DZ”).

20 **Figure 14** demonstrates the secretion of wild-type Cel5H and Cel5H fused to a carrier domain, compared to a control strain. The carrier domain encompasses a carbohydrate binding module (CBM3a) from a cellulosomal scaffolding protein fused to two hydrophilic domains (Xa). The activity of the culture supernatant was measured on the soluble substrate *para*-nitrophenyl-cellobiose.

25 **Figure 15** demonstrates the activity of different proteins including the fusion proteins according to particular embodiments of the invention on different celluloic substrates: (a) activity of proteins comprising Cel5H on soluble substrate *para*-nitrophenyl-cellobiose; wild-type Cel5H (full line), fusion with one X module (CBM-Xa-5H; dotted line), fusion protein with two X modules (CBM-Xa-Xa-5H, dashed line) (b) activity of proteins comprising cel5H on crystalline cellulose Avicel.

Figure 16 schematically illustrates comparison of domains of Cel5H and ACLA, according to particular embodiments of the invention.

DETAILED DESCRIPTION OF THE INVENTION

As used herein, the singular forms "a", "an", and "the" include both singular and plural referents unless the context clearly dictates otherwise.

The terms "comprising", "comprises" and "comprised of" as used herein are synonymous with "including", "includes" or "containing", "contains", and are inclusive or open-ended and do not exclude additional, non-recited members, elements or method steps.

The recitation of numerical ranges by endpoints includes all numbers and fractions subsumed within the respective ranges, as well as the recited endpoints.

The term "about" as used herein when referring to a measurable value such as a parameter, an amount, a temporal duration, and the like, is meant to encompass variations of +/-10% or less, preferably +/-5% or less, more preferably +/-1% or less, and still more preferably +/-0.1% or less of and from the specified value, insofar such variations are appropriate to perform in the disclosed invention. It is to be understood that the value to which the modifier "about" refers is itself also specifically, and preferably, disclosed.

All documents cited in the present specification are hereby incorporated by reference in their entirety.

Unless otherwise defined, all terms used in disclosing the invention, including technical and scientific terms, have the meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. By means of further guidance, term definitions are included to better appreciate the teaching of the present invention.

As used herein, the terms "Cel5H", "Cel5H polypeptide" or "Cel5H protein" refer to the polypeptide or protein of *Saccharophagus degradans* strain 2-40 commonly known under these designations in the art (see in particular Taylor *et al.* 2006, *supra*). The above designations particularly refer to such polypeptides with a native sequence, *i.e.*, polypeptides of which the primary sequence is the same as that of the Cel5H protein of *S. degradans* strain 2-40 found in or derived from nature. A skilled person understands that native sequences of Cel5H might differ between different sub-strains or sub-cultures of *S. degradans* strain 2-40 due to genetic divergence or spontaneous mutation(s). The native sequences of Cel5H may further diverge due to post-transcriptional or post-translational modifications. Accordingly, all Cel5H sequences found in or derived from nature are considered "native".

The above designations encompass the Cel5H polypeptide when forming a part of a living organism, microorganism or cells, as well as when at least partly isolated from such sources. The terms also encompass Cel5H polypeptides when produced by recombinant or synthetic means.

5 The Cel5H polypeptide is normally secretory and may exist in a precursor form comprising a cleavable N-terminal secretion signal sequence, or in a mature form lacking said signal sequence. Depending on context readily understood by a skilled person, a reference herein to the Cel5H polypeptide may encompass said mature form and/or said precursor forms.

Exemplary Cel5H polypeptides include without limitation the polypeptide encoded by the
10 nucleic acid coding sequence as annotated under the NCBI Entrez Gene database
(<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>) locus tag Sde_3237 (GeneID:
3965710), and under the NCBI Genbank (<http://www.ncbi.nlm.nih.gov/Genbank/index.html>)
accession number NC 007912 (sequence version 1, database entry updated 20July 2008),
also reproduced in Fig. 1A (SEQ ID NO: 1); and the Cel5H polypeptide as annotated under
15 the NCBI Genbank accession number YP 528706 (sequence version 1, database entry
update 20July 2008) and the Uniprot/Swissprot (<http://www.expasy.org/>) accession number
Q21FN5 (sequence version 1 entered on April 18, 2006), as exemplarily reproduced in Fig.
1B (SEQ ID NO: 2).

This sequence corresponds to a Cel5H precursor comprising the signal sequence. An
20 exemplary mature Cel5H polypeptide, *i.e.*, with the signal sequence processed away, is
represented by amino acid positions 35 to 630 of said precursor sequence, as exemplarily
reproduced in Fig. 1C (SEQ ID NO: 3).

One shall appreciate that the actual cleavage by signal peptidase may potentially occur at a
peptide bond nearby the predicted cleavage site between amino acids 34 and 35 of the
25 precursor Cel5H such as shown in Fig. 1B. For example, the actual cleavage might occur at
any of the peptide bonds within the amino acid region from position 30 to 40, preferably 31
to 39, more preferably 32 to 38, even more preferably 32 to 37, still more preferably 33 to
36, of the precursor Cel5H such as shown in Fig. 1B.

Further, unless indicated otherwise expressly or by context, whenever domains or other
30 motifs of Cel5H are identified herein by reference to amino acid positions, such reference is
to the mature form of Cel5H, such as in particular that shown in SEQ ID NO: 3.

As used herein, the term "homology" denotes structural similarity between two macromolecules, particularly between two polypeptides or polynucleotides, from same or different taxons, wherein said similarity is due to shared ancestry. Hence, the term "homologues" denotes so-related macromolecules having said structural similarity.

5 Preferably, homologues of the Cel5H polypeptide as intended herein encompass said homologues with native sequence.

A polypeptide homologue of the Cel5H polypeptide may preferably show at least about 30%, more preferably at least 40%, even more preferably at least 50%, still more preferably at least 60%, yet more preferably at least 70%, such as very preferably at least 80% or at 10 least 90% or at least 95% sequence identity as defined elsewhere in this specification to the Cel5H polypeptide. Alternatively or preferably in addition, a polypeptide homologue of the Cel5H polypeptide may show at least about 50%, more preferably at least 60%, even more preferably at least 70%, still more preferably at least 80%, yet more preferably at least 90%, such as very preferably at least 95% sequence similarity as defined elsewhere in this 15 specification to the Cel5H polypeptide. Further preferably, a polypeptide homologue of the Cel5H polypeptide may comprise one or more functional domains corresponding to domains of the Cel5H polypeptide, and preferably one or more of GH family 5 catalytic domain, CBM family 6 domain and DZ domain. Preferably, said domains may have organisation analogous to Cel5H, in particular GH5-CBM6-DZ, usually with interposing 20 linkers.

As noted, preferred homologues of the Cel5H polypeptide are the ACLA polypeptides of *Pseudomonas* sp. ND137. Exemplary ACLA polypeptides includes without limitation the sequence as annotated under the Uniprot/Swissprot accession number Q8VUT3 (sequence version 1 entered on March 1, 2002) also reproduced in Fig. 1D (SEQ ID NO: 4). This 25 sequence corresponds to ACLA precursor comprising the signal sequence. An exemplary mature ACLA polypeptide, *i.e.*, with the signal sequence processed away, is represented by amino acid positions 28 to 585 of the respective precursor, as exemplarily reproduced in Fig. 1E (SEQ ID NO: 5).

One shall appreciate that the actual cleavage by signal peptidase may potentially occur at a 30 peptide bond nearby the predicted cleavage site between amino acids 27 and 28 of the precursor ACLA such as shown in Fig. 1D (SEQ ID NO:4). For example, the actual cleavage might occur at any of the peptide bonds within the amino acid region from position

23 to 32, preferably 24 to 31, more preferably 25 to 30, even more preferably 26 to 29, of the precursor ACLA such as shown in Fig. 1D.

Depending on context readily understood by a skilled person, a reference herein to the ACLA polypeptide may encompass said mature form and/or said precursor forms. Further, 5 unless indicated otherwise expressly or by context, whenever domains or other motifs of ACLA are identified herein by reference to amino acid positions, such reference is to the mature form of ACLA, such as in particular that shown in SEQ ID NO: 5.

The term "fragment" with reference to a given polypeptide such as Cel5H or a homologue thereof, or with reference to a given domain such as the DZ domain of the Cel5H or 10 homologue thereof, generally refers to a truncated form of said polypeptide that has an N-terminal and/or C-terminal deletion of one or more amino acid residues as compared to said polypeptide such as said Cel5H or homologue or said domain, but where the remaining primary sequence of the fragment is identical to the corresponding positions in the amino acid sequence of said polypeptide such as said Cel5H or homologue or said domain.

15 For example, a fragment of a given polypeptide such as Cel5H or a homologue thereof, may include a sequence of about ≥ 5 consecutive amino acids, preferably ≥ 10 consecutive amino acids, more preferably ≥ 20 consecutive amino acids, even more preferably ≥ 30 consecutive amino acids, e.g., ≥ 40 consecutive amino acids, and most preferably ≥ 50 consecutive amino acids, e.g., ≥ 60 , ≥ 70 , ≥ 80 , ≥ 90 , ≥ 100 , ≥ 200 or ≥ 500 consecutive amino 20 acids of said polypeptide such as said Cel5H or homologue.

For example, a fragment of the DZ domain of Cel5H or a homologue thereof, may include a sequence of about ≥ 5 consecutive amino acids, preferably ≥ 10 consecutive amino acids, more preferably ≥ 20 consecutive amino acids, even more preferably ≥ 30 consecutive amino acids, e.g., ≥ 40 consecutive amino acids, and most preferably ≥ 50 consecutive 25 amino acids, e.g., ≥ 60 , ≥ 70 , ≥ 80 , or ≥ 90 , consecutive amino acids of said DZ domain or homologue thereof.

Further, a fragment of a given polypeptide such as Cel5H or a homologue thereof or DZ domain may represent at least about 30%, e.g., at least 50% or at least 70%, preferably at least 80%, e.g., at least 85%, more preferably at least 90%, and yet more preferably at 30 least 95% or even about 99% of the amino acid sequence of said polypeptide such as said Cel5H or homologue or DZ domain.

The term "variant" of a given polypeptide such as Cel5H or a homologue thereof or of a DZ domain refers to polypeptides the amino acid sequence of which is substantially identical (i.e., largely but not wholly identical) to a native sequence of said polypeptide such as said Cel5H or a homologue thereof or a DZ domain. "Substantially identical" refers to at least 5 about 85% identical, e.g., preferably at least 90% identical, e.g., at least 91% identical, 92% identical, more preferably at least 93% identical, e.g., 94% identical, even more preferably at least 95% identical, e.g., at least 96% identical, yet more preferably at least 97% identical, e.g., at least 98% identical, and most preferably at least 99% identical.

Sequence identity between two polypeptides can be determined by optimally aligning 10 (optimal alignment of two protein sequences is the alignment that maximises the sum of pair-scores less any penalty for introduced gaps; and may be preferably conducted by computerised implementations of algorithms, such as "Gap", using the algorithm of Needleman and Wunsch 1970 (J Mol Biol 48: 443-453), or "Bestfit", using the algorithm of Smith and Waterman 1981 (J Mol Biol 147: 195—197), as available in, e.g., the GCG™ v. 15 11.1.2 package from Accelrys) the amino acid sequences of the polypeptides and scoring, on one hand, the number of positions in the alignment at which the polypeptides contain the same amino acid residue and, on the other hand, the number of positions in the alignment at which the two polypeptides differ in their sequence. The two polypeptides differ in their sequence at a given position in the alignment when the polypeptides contain different 20 amino acid residues at that position (amino acid substitution), or when one of the polypeptides contains an amino acid residue at that position while the other one does not or vice versa (amino acid insertion or deletion). Sequence identity is calculated as the proportion (percentage) of positions in the alignment at which the polypeptides contain the same amino acid residue versus the total number of positions in the alignment. Further 25 suitable algorithms for performing sequence alignments and determination of sequence identity include those based on the Basic Local Alignment Search Tool (BLAST) originally described by Altschul et al. 1990 (J Mol Biol 215: 403-10), such as the "Blast 2 sequences" algorithm described by Tatusova and Madden 1999 (FEMS Microbiol Lett 174: 247-250), such as using defaults settings thereof.

30 The variant of a given polypeptide such as Cel5H or a homologue thereof or a DZ domain, as used herein, also specifically encompasses polypeptides having a certain degree of similarity to said polypeptide such as said Cel5H or a homologue thereof or a DZ domain. Preferably, such variants can be at least about 90% similar, e.g., preferably at least 91%

similar, e.g., at least 92% similar, 93% similar, more preferably at least 94% similar, e.g., 95% similar, even more preferably at least 96% similar, e.g., at least 97% similar, yet more preferably at least 98% similar, e.g., at least 99% similar.

Sequence similarity between two polypeptides can be determined by optimally aligning (see 5 above) the amino acid sequences of the polypeptides and scoring, on one hand, the number of positions in the alignment at which the polypeptides contain the same or similar (i.e., conservatively substituted) amino acid residue and, on the other hand, the number of positions in the alignment at which the two polypeptides otherwise differ in their sequence. The two polypeptides otherwise differ in their sequence at a given position in the alignment 10 when the polypeptides contain non-conservative amino acid residues at that position, or when one of the polypeptides contains an amino acid residue at that position while the other one does not or vice versa (amino acid insertion or deletion). Sequence similarity is calculated as the proportion (percentage) of positions in the alignment at which the polypeptides contain the same or similar amino acid residue versus the total number of 15 positions in the alignment.

It shall be understood that under reference to fragments and/or variants the specification also broadly contemplates fragments of variants of a given polypeptide such as Cel5H or a homologue thereof or a DZ domain, as well as variants of fragments of such polypeptide.

The term "functional" with reference to the fragments and/or variants of a given polypeptide 20 such as Cel5H or a homologue thereof or a DZ domain denotes that such fragments and variants at least partly retain the biological activity or functionality of the corresponding polypeptide such as said Cel5H or homologue thereof or DZ domain. Preferably, such functional fragments and/or variants may retain at least about 20%, e.g., at least 30%, or at least 40%, or at least 50%, e.g., at least 60%, more preferably at least 70%, e.g., at least 25 80%, yet more preferably at least 85%, still more preferably at least 90%, and most preferably at least 95% or even 100% of the activity of the corresponding polypeptide such as said Cel5H or homologue thereof or DZ domain. Potentially, such functional fragments and/or variants may even have higher activity than the corresponding polypeptide such as said Cel5H or homologue or DZ domain.

30 Functional fragments and/or variants of a given polypeptide such as Cel5H or homologue thereof or DZ domain may be functionally equivalent to said polypeptide in at least one and preferably more or all aspects of its biological activity or function. Relevant aspects of

biological function of said Cel5H or homologue thereof include in particular but without limitation cellulose-depolymerising or cellulase activity, activity profile with respect to different cellulose substrates (e.g., crystalline, semi-crystalline or amorphous cellulose or hemicellulose), ability or preference to degrade crystalline cellulose substrates, and ability 5 to interact with or bind to cellulose substrates. Relevant aspects of biological function of said DZ domain of Cel5H or homologue thereof include in particular but without limitation cellulose-binding and particularly crystalline cellulose-binding, and/or oligomerisation activities. The biological activity of a given fragment and/or variant of Cel5H or of homologue thereof or of DZ domain, including the above aspects of such activity, can be 10 determined by standard tests, such as for example the enzyme activity tests presented in the examples section.

The terms "solventogenic" or "solvent-producing" have their art-established meaning and in particular denote the ability of microorganisms such as bacteria (i.e., solventogenic bacteria) to produce one or more non-gaseous organic liquids or solvents, such as *inter alia* 15 ethanol, acetone, butanol, propionic acid, butyric acid, ether or glycerine, from a carbohydrate source such as for example hexoses, pentoses or oligosaccharides. In particular, the term encompasses naturally occurring solventogenic organisms, solventogenic organisms with naturally occurring or induced mutations, and solventogenic organisms which have been genetically modified.

20 The terms "ethanogenic" or "ethanol-producing" is intended to denote the ability of microorganisms such as bacteria (i.e., ethanogenic bacteria) to produce at least or preferably mainly ethanol from a carbohydrate source such as for example hexose, pentose or oligosaccharides, more preferably to produce ethanol from carbohydrates as the most abundant non-gaseous fermentation product. In particular, the term encompasses naturally 25 occurring ethanogenic organisms, ethanogenic organisms with naturally occurring or induced mutations, and ethanogenic organisms which have been genetically modified. Examples of suitable solventogenic microorganism include but are not limited to strains of Clostridium and Zymomonas, such as, but not limited to those exemplified herein.

30 The term "isolating" with reference to a particular component generally denotes separating that component from at least one other component of a composition from which the former component is thereby "isolated". In particular, the terms "isolating" and "isolated" may refer herein to increasing amount of a desired protein(s) or polypeptide(s) in a sample. The

relative amount may be expressed as the ratio between the concentration or amount of the desired protein(s) or polypeptide(s) to be isolated and the concentration or amount of total proteins in the sample. In addition, the term may encompass separating desired protein(s) or polypeptide(s) from non-protein cellular components (e.g., cell walls, lipids, nucleic acids, 5 etc.).

The terms "recombinant nucleic acid" or "recombinant nucleic acid molecule" as used herein generally refer to nucleic acid molecules (such as, e.g., DNA, cDNA or RNA molecules) comprising segments generated and/or joined together using recombinant DNA technology, such as for example molecular cloning and nucleic acid amplification. Usually, 10 a recombinant nucleic acid molecule may comprise one or more non-naturally occurring sequences, and/or may comprise segments corresponding to naturally occurring sequences that are not positioned as they would be positioned in a source genome which has not been modified. When a recombinant nucleic acid molecule replicates in the host organism into which it has been introduced, the progeny nucleic acid molecule(s) are also 15 encompassed within the term "recombinant nucleic acid molecule".

In particular embodiments, a recombinant nucleic acid molecule can be stably integrated into the genome of a host organism, such as for example integrated at one or more random positions or integrated in a targeted manner, such as, e.g., by means of homologous recombination, or the recombinant nucleic acid molecule can be present as or comprised 20 within an extra-chromosomal element, wherein the latter may be auto-replicating.

Standard reference works setting forth the general principles of recombinant DNA technology include Molecular Cloning: A Laboratory Manual, 2nd ed., vol. 1-3, ed. Sambrook *et al.*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989; Current Protocols in Molecular Biology, ed. Ausubel *et al.*, Greene Publishing and Wiley- 25 Interscience, New York, 1992 (with periodic updates); Innis *et al.*, PCR Protocols: A Guide to Methods and Applications, Academic Press: San Diego, 1990. General principles of microbiology are set forth, for example, in Davis, B. D. et al., Microbiology, 3d edition, Harper & Row, publishers, Philadelphia, Pa. (1980).

The term "recombinant polypeptide" as used herein refers to a polypeptide or protein 30 produced by a host organism through the expression of a recombinant nucleic acid molecule, which has been introduced into said host organism or an ancestor thereof, and which comprises a sequence encoding said polypeptide or protein.

Further, the terms "recombinant" or "transformed" as used herein with reference to host organisms, microorganisms or cells, encompass such host organisms, microorganisms or cells into which a recombinant nucleic acid molecule has been introduced, as well as the recombinant progeny of such host organisms, microorganism or cells.

5 Hence, the term "transformation" encompasses the introduction or transfer of a foreign nucleic acid such as a recombinant nucleic acid into a host organism, microorganism or cell. The so-introduced nucleic acid may be preferably maintained throughout the further growth and cell division of said host organism, microorganism or cell. Any conventional gene transfer methods may be used to achieve transformation, such as without limitation 10 electroporation, electropermeation, chemical transformation, lipofection, virus- or bacteriophage-mediated transfection, etc.

By "encoding" is meant that a nucleic acid sequence or part(s) thereof corresponds, by virtue of the genetic code of an organism in question to a particular amino acid sequence, e.g., the amino acid sequence of a desired polypeptide or protein. By means of example, 15 nucleic acids "encoding" a particular polypeptide or protein may encompass genomic, hnRNA, pre-mRNA, mRNA, cDNA, recombinant or synthetic nucleic acids.

Preferably, a nucleic acid encoding a particular polypeptide or protein may comprise an open reading frame (ORF) encoding said polypeptide or protein. An "open reading frame" or "ORF" refers to a succession of coding nucleotide triplets (codons) starting with a 20 translation initiation codon and closing with a translation termination codon known *per se*, and not containing any internal in-frame translation termination codon, and potentially capable of encoding a polypeptide. Hence, the term may be synonymous with "coding sequence" as used in the art.

In an embodiment, the nucleic acid sequence or ORF encoding the present polypeptide(s) 25 may be codon optimised as known *per se* for expression in a particular organism, e.g., microorganism, more particularly a bacterium of interest. Codon usage bias and codon frequencies from various organisms are available, for example via the Codon Usage Database (<http://www.kazusa.or.jp/codon/>) described by Nakamura et al. 2000 (Nucl Acids Res 28: 292).

30 Expression of polypeptides of interest in recombinant microorganisms as taught herein can be achieved through operably linking the nucleic acid sequences or ORF(s) which encode said polypeptides with regulatory sequences allowing for expression in said

microorganisms. In this regard, the term "expressing" a polypeptide when referring to a recombinant organism encompasses recombinant organisms capable of expressing the polypeptide, e.g. under suitable culture conditions or upon addition of inducers (e.g. where inducible regulatory sequences are used).

5 An "operable linkage" is a linkage in which regulatory nucleic acid sequences and sequences sought to be expressed are connected in such a way as to permit said expression. For example, sequences, such as, e.g., a promoter and an ORF, may be said to be operably linked if the nature of the linkage between said sequences does not: (1) result in the introduction of a frame-shift mutation, (2) interfere with the ability of the
10 promoter to direct the transcription of the ORF, (3) interfere with the ability of the ORF to be transcribed from the promoter sequence.

The precise nature of regulatory sequences or elements required for expression may vary between organisms, but typically include a promoter and a transcription terminator, and optionally an enhancer.

15 Reference to a "promoter" or "enhancer" is to be taken in its broadest context and includes transcriptional regulatory sequences required for accurate transcription initiation and where applicable accurate spatial and/or temporal control of gene expression or its response to, e.g., internal or external (e.g., exogenous) stimuli. More particularly, "promoter" may depict a region on a nucleic acid molecule, preferably DNA molecule, to which an RNA
20 polymerase binds and initiates transcription. A promoter is preferably, but not necessarily, positioned upstream, *i.e.*, 5', of the sequence the transcription of which it controls. Typically, in prokaryotes a promoter region may contain both the promoter *per se* and sequences which, when transcribed into RNA, will signal the initiation of protein synthesis (e.g., Shine-Dalgarno sequence).

25 In embodiments, promoters contemplated herein may be constitutive or inducible.

By means of example, constitutive promoters suitable for expression in *Clostridium* species such as *Clostridium acetobutylicum* include without limitation the promoter of the thiolase gene of *C. acetobutylicum* (P_{thl} promoter), the promoter of the acetoacetate decarboxylase
30 gene (P_{adc} promoter), the promoter of the phosphotransbutyrylase gene (P_{ptb} promoter), the xylose-inducible promoter of the *Staphylococcus xylosus* xylose operon (P_{xy} promoter)

Hence, in particular embodiments a promoter is provided which is a promoter controlled by an operator. As used herein, the term "operator" refers to a nucleotide sequence, preferably DNA sequence, which controls the initiation and/or maintenance of transcription of a sequence from a promoter. Typically, an operator may be generally placed between a 5 promoter and a downstream sequence the transcription of which the promoters controls. Usually, an operator is capable of binding a repressor polypeptide, whereby it reduces the transcription from the said promoter. A useful repressor can alternate between a state in which it binds the operator and a state in which it does not and such alternation may be advantageously controlled by an external condition, e.g., an external substance or a 10 particular metabolite. Accordingly, in host cells comprising a compatible repressor, the inclusion of an operator in the nucleic acid of the invention may allow to control the activity of the promoter and expression there from. Operator sequences may be generally derived from bacterial chromosomes.

It shall also be appreciated that the nucleic acids of the invention may encode one or more 15 than one polypeptides of interest. Moreover, where expression of two or more polypeptides is intended, said expression may be from independent expression units, or may be from a single expression unit (*i.e.*, multi-cistronic expression) comprising two or more sequential ORFs controlled by common regulatory elements. By means of example, an expression unit encoding two or more polypeptides may be generated by associating translation initiation 20 codons of the respective ORFs with sequences controlling translation initiation (*e.g.*, ribosome entry sequences).

The terms "terminator" or "transcription terminator" refer generally to a sequence element at the end of a transcriptional unit which signals termination of transcription. For example, a 25 terminator is usually positioned downstream of, *i.e.*, 3' of ORF(s) encoding a polypeptide of interest. For instance, where a recombinant nucleic acid contains two or more ORFs, *e.g.*, successively ordered and forming together a multi-cistronic transcription unit, a transcription terminator may be advantageously positioned 3' to the most downstream ORF.

In a preferred embodiment, regulatory sequences controlling the expression of 30 polypeptide(s) as taught herein may be advantageously provided on the recombinant nucleic acid to be used for transforming the present microorganisms. However, also contemplated are embodiments in which the recombinant nucleic acid provides for coding sequence(s) lacking one or more regulatory sequences, whereas the required regulatory

sequences are provided by the transformed microorganism (such as, e.g., wherein the recombinant nucleic acid inserts into a chromosomal or episomal region comprising suitable regulatory sequences).

Clostridium and *Clostridium acetobutylicum* as used herein refer to bacterial taxons known 5 as such in the art, and particularly encompass bacteria, bacterial species, strains, subspecies and cultures belonging to said taxons, as well as modified or engineered, such as for example mutated or genetically engineered, derivatives of naturally-occurring or wild-type specimens. By means of guidance and not limitation, isolated strains of *Clostridium* available from the American Type Culture Collection (ATCC) can be ordered *inter alia* 10 under accession numbers ATCC 824, ATCC 4259, ATCC 39236 or ATCC 43084 for *Clostridium acetobutylicum*. Isolated strains of *Clostridium beijerinckii* can, for example be ordered *inter alia* under accession numbers ATCC 858 and ATCC 25752.

As noted, the polypeptides expressed by the present solventogenic microorganisms may be 15 targeted for secretion. To achieve secretion, a nucleic acid sequence encoding a polypeptide may be operably linked to a sequence encoding a secretion signal sequence. In this connection, "operably linked" denotes that the sequence encoding the signal sequence and the sequence encoding the polypeptide to be secreted are connected in frame or in phase, such that upon expression the signal peptide facilitates the secretion of 20 the polypeptide so-linked thereto.

It shall be appreciated that suitable signal sequences may depend on the type of microorganism in which secretion is desired. For example, distinct signal sequences may be required in Gram-positive bacteria vs. Gram-negative bacteria.

By means of example and not limitation, secretion in Gram-positive bacteria, and in 25 particular in *Clostridium* such as *C. acetobutylicum*, may be achieved using the signal sequence of the Cel5A precursor polypeptide of *C. cellulolyticum* (exemplary sequence: Genbank acc. no. AAA51444, seq version 1 revised on October 31, 1994), or of the CipC precursor scaffolding protein of *C. cellulolyticum* (exemplary sequence: Genbank acc. no. AAC28899, seq. version 2 revised on December 5, 2005), or of the CipA precursor 30 scaffolding protein of *C. acetobutylicum* (exemplary sequence: Genbank acc. no. AAK78886, seq. version 1 revised on January 19, 2006).

Further without limitation, secretion in Gram-negative bacteria, may be achieved via the Sec pathway, e.g., using the signal sequence of the gluconolactonase precursor polypeptide of *Z. mobilis* (exemplary sequence: Genbank acc. no. CAA47637, seq. version 1 revised on October 19, 2006) or of the carbohydrate selective porine OprB (exemplary sequence: Genbank acc. no. AAV88688, seq. version 1 revised on January 25, 2005), or via the twin-arginine translocation (Tat) pathway, e.g., using the signal sequence of the gluco fructose oxidoreductase precursor polypeptide of *Z. mobilis* (exemplary sequence: Genbank acc. no. CAB02496, seq. version 1 revised on October 19, 2006).

It shall also be appreciated that native (or homologous, endogenous) signal peptides of polypeptides to be expressed by the microorganisms as taught herein may be employed, insofar they are functional in said microorganisms. Hence, by means of example, secretion of Cel5H or related polypeptides may be achieved using the endogenous or homologous secretion signal sequence of the Cel5H precursor polypeptide. In particular, since *S. degradans* is a Gram-negative bacterium, the endogenous or homologous signal sequence of Cel5H may be particularly suited for secretion thereof in other Gram-negative bacteria. In other embodiments, polypeptides such as Cel5H and related polypeptides may be secreted using heterologous signal sequences.

The domain structure of the native Cel5H polypeptide can be outlined as GH5-PSL-CBM6-EPR-DZ, with schematic indication of the domain boundaries in mature Cel5H shown in Fig. 9. Functional fragments of Cel5H or homologues thereof preferably contain one or more of the GH5, CBM6 and DZ domains, or domains corresponding thereto (e.g., analogous domains within a homologue and/or variant of Cel5H). Preferably, domains comprised in such fragments may be functional as intended herein. Hence, in particular embodiments the invention contemplates fragments such as those corresponding to of the following domain organisations: GH5, GH5-PSL, GH5-PSL-CBM6, GH5-PSL-CBM6-EPR, PSL-CBM6, PSL-CBM6-EPR, PSL-CBM6-EPR-DZ, CBM6, CBM6-EPR, CBM6-EPR-DZ, and EPR-DZ, DZ. The invention further envisages fragments corresponding to combinations of the domains of mature Cel5H, such as, but not limited to EPR-DZ-EPR-DZ.

As noted, the invention also relates to various fusions of Cel5H and related polypeptides with useful heterologous domains. In this context, the term "heterologous" denotes that said domain(s) are derived from polypeptides other than the Cel5H polypeptide or the homologue thereof. Accordingly, "heterologous" when referring to combinations of domains

refers to the fact that the different domains do not originate from (or do not naturally occur in) the same protein. Nevertheless, fusions of additional Cel5H-derived domain(s) with Cel5H polypeptide or homologue thereof, or fragments and/or variants thereof, can also be contemplated.

5 The term “fusion” is broadly contemplated herein as encompassing *inter alia* genetic fusions (i.e., fusing distinct parts into a single open reading frame) as well as fusions induced by chemical and/or physical means (e.g., chemical cross-linking or photo-coupling). Furthermore, fusion includes direct fusion or fusion via a linker. Suitable linkers may include without limitation peptide sequences of at least 3, preferably at least 4 or 5, more preferably 10 at least 7, and yet more preferably at least 12 amino acids, such as between 3 and 15 amino acids, and preferably comprising, being enriched for or consisting of non-charged amino acids, preferably chosen from glycine, serine, cysteine, asparagine, tyrosine, glutamine, alanine, valine, proline, threonine, and more preferably glycine or serine.

15 Exemplary glycoside hydrolase (GH) catalytic domains may be derived from glycoside hydrolases (glycosidases) known *per se* or future ones identified and isolated *inter alia* using established procedures. Exemplary glycosidases include those grouped under the EC 3.2.1. “Glycosidases” classification of NC-IUBMB, more preferably glycosidases capable of depolymerising cellulosic and lingo-cellulosic substrates, such as, e.g., cellulases (EC 3.2.1.4.), cellulose 1,4- β -cellobiosidases (EC 3.2.1.91) and processive endocellulases (EC 20 3.2.1.4./EC 3.2.1.91.).

25 The term carbohydrate-binding module(s) (CBM) is known in the art and commonly broadly covers all non-catalytic sugar- or carbohydrate-binding modules found in or derived from glycoside hydrolases. CBM modules preferred herein may include those binding specifically to cellulose or cellulosic substrates and alternatively referred to as cellulose-binding domains (CBD). CMB and their structural and functional properties are reviewed *inter alia* in Tomme *et al.* 1995 (“Cellulose-binding domains: classification and properties”, in Enzymatic Degradation of Insoluble Polysaccharides, Saddler JN & Penner M, eds., pp. 142-163, American Chemical Society, Washington) and Boraston *et al.* 2004 (“Carbohydrate-binding modules: fine tuning polysaccharide recognition”. Biochem J 382: 769-81).

30 Cohesin-binding domains as used herein refer generally to all polypeptide domains capable of specifically binding to one or more receptor (cohesin) domains of cellulosomal scaffoldin subunits. Exemplary cohesin-binding domains include dockerin domains found in or derived

from glycoside hydrolases that assemble to cellulosomes. Both type I and type II dockerin domains are contemplated herein. Dockerin domains are explained *inter alia* in Lytle *et al.* 2001 ("Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain". *J Mol Biol* 307: 745-753), Adams *et al.* 2005 ("Structural characterization of type II dockerin module from the cellulosome of *Clostridium thermocellum*: calcium-induced effects on conformation and target recognition". *Biochemistry* 44: 2173-82) and in Bayer *et al.* 1998 ("Cellulosomes-structure and ultrastructure". *J Struct Biol* 124: 221-234). A suitable assay for determining dockerin-cohesin interactions, such as to guide the choice of a suitable dockerin domain herein, is described *inter alia* in Haimovitz *et al.* 2008 ("Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules". *Proteomics* 8: 968-979).

Further contemplated is co-expression and optionally co-secretion of polypeptides taught herein with other polypeptides or enzymes, preferably recombinant polypeptides or enzymes, useful in carbohydrate polymer metabolism, particularly in cellulose metabolism. The term "co-expression" generally relates to expression of two or more proteins or polypeptides (said to be "co-expressed") by the same microorganism, particularly by a same cell of said microorganism. Suitable systems for achieving co-expression of two or more polypeptides are known *per se*. Without limitation, co-expressed polypeptides may be encoded by separate cistrons controlled by the same or different regulatory elements, or may be encoded by a single multi-cistronic element; such cistrons may be on a chromosome or on same or different epigenetic elements, or combinations thereof etc.

The terms "glycoside hydrolase" or "glycosidase" generally encompass enzymes and enzyme complexes capable of hydrolysing glycosidic linkages, more particularly O-, N- or S- glycosidic linkages, even more preferably such linkages in oligosaccharide and/or polysaccharide substrates. The terms encompass exoglycosidases as well as endoglycosidases. The terms particularly encompass glycosidases grouped under the EC 3.2.1. "Glycosidases" classification of NC-IUBMB.

The term "cellulase" generally encompasses enzymes and enzyme complexes capable of hydrolysing cellulose and cellulose-containing substrates. The term encompasses without limitation the following cellulase types: cellobiohydrolase(1,4- β -D-glucan cellobiohydrolase, EC 3.2.1.91), endo- β -1,4-glucanase (endo-1,4- β -D-glucan 4-glucanohydrolase, EC 3.2.1.4),

processive endocellulase (EC 3.2.1.4./EC 3.2.1.91.) and β -glucosidase (EC 3.2.1.21). Cellulases further encompass endoglucanases as well as (processive) exoglucanases, and cover enzymes which produce monomers and/or oligomers of various lengths. The term may also encompass cellobiases, oxidative cellulases, glucosidases, cellulose phosphorylases, and other enzymes commonly denoted thereby. The term encompasses enzymes capable of degrading various forms of cellulose, such as without limitation crystalline cellulose, semi-crystalline cellulose, amorphous cellulose, hemicellulose and/or chemically or biologically modified cellulose forms.

By means of example and not limitation, the cellulolytic system of *Clostridium cellulolyticum* 10 has been reviewed by Bélaich *et al.* 1997 ("The cellulolytic system of *Clostridium cellulolyticum*". J Biotechnol 57: 3-14); and of *Saccharophagus degradans* in Taylor *et al.* 2006 (*supra*).

The invention further contemplates engineered cellulosomes comprising polypeptides, particularly Cel5H and related polypeptides (e.g., DZ-domain and polypeptides comprising 15 such), as taught herein. As known in the art, native cellulosomes represent extracellular, multienzyme cellulolytic complexes of certain bacterial taxons, such as *inter alia Clostridium* and *Bacteroides*. The basic organisation of native cellulosomes involves an organising polypeptide (scaffoldin), commonly comprising one or more same or distinct carbohydrate-binding modules (CBM) and one or more same or distinct receptor (cohesin) domains that 20 facilitate integration of glycosidases or cellulolytic enzymes into the cellulosome complex through interaction with cognate dockerin domains in said enzymes.

These structural features of native cellulosomes may also be employed in engineered cellulosomes, whereby desired carbohydrate-binding and carbohydrate-depolymerising activities may be integrated into engineered cellulosomes. Techniques for producing 25 engineered cellulosomes are established in the art (see, e.g., Fierobe *et al.* 2002 ("Degradation of cellulose substrates by cellulosome chimeras. Substrate targeting versus proximity of enzyme components". J Biol Chem 277: 49621-30); Fierobe *et al.* 2005 ("Action of designer cellulosomes on homogeneous versus complex substrates: controlled incorporation of three distinct enzymes into a defined trifunctional scaffoldin". J Biol Chem 280: 16325-34); and Perret *et al.* 2004 ("Production of heterologous and chimeric scaffoldins by *Clostridium acetobutylicum* ATCC 824". J Bacteriol 186: 253-7)).

Commonly, an engineered cellulosome may contain a hybrid or chimeric scaffoldin polypeptide comprising one or more CBM modules of the same or divergent carbohydrate substrate specificity, and one or more cohesin domains capable of binding the same or divergent cognate dockerin domains. Enzymes, such as particularly glycosidases or 5 cellulolytic enzymes, may be incorporated into the cellulosome by means of cognate dockerin modules (normally present in and/or engineered into said enzymes) for the cohesin domains found in the scaffoldin polypeptide. Said enzymes may also comprise CBM or other non-catalytic domains. Such engineered cellulosomes are commonly known in the art as "chimeric" or "hybrid" cellulosomes or "minicellulosomes".

10 Alternatively or additionally, the one or more enzymes, such as glycosidases or cellulolytic enzymes, may be covalently linked with the scaffoldin backbone, such as, e.g., by being expressed as a part of the same polypeptide chain (i.e., genetic fusion). Such cellulosomes are commonly known as "covalent" (Mingardon et al. 2007, "Exploration of new geometries in cellulosome-like chimeras" Appl. Environ. Microbiol. 73: 7138-7149).

15 Hence, engineered cellulosomes, such as hybrid, chimeric, mini-, or covalent cellulosomes, or any combination of these modalities, or any other engineered cellulosome forms, are contemplated for use with the polypeptides taught herein.

The term "cellulose" is known *per se* in the art. By means of further explanation and not limitation, the term may encompass all forms of cellulose, such as without limitation 20 naturally and non-naturally occurring cellulose forms, such as, e.g., crystalline cellulose, semi-crystalline cellulose, microcrystalline cellulose, amorphous cellulose, hemicellulose, regenerated cellulose, and/or chemically or biologically modified or derivatised cellulose forms, such as, e.g., ethers or esters thereof, *etc.*

A substance enriched in crystalline cellulose may comprise, without limitation, at least about 25 1% (w/w), e.g., $\geq 5\%$ (w/w), preferably at least about 10% (w/w), e.g., $\geq 20\%$, $\geq 30\%$ or $\geq 40\%$ (w/w), more preferably at least about 50% (w/w), e.g., $\geq 60\%$ or $\geq 70\%$ (w/w), yet more preferably at least about 80% (w/w), e.g., $\geq 90\%$ (w/w) of crystalline cellulose, semi-crystalline and/or microcrystalline cellulose as understood in the art, more preferably of crystalline cellulose.

30 Methods and processes for treating or contacting cellulose-containing substances with recombinant microorganisms such as taught herein are generally known in the art, and may without limitation comprise industrial fermentation processes of various scale. The material

to be so-fermented may be provided, e.g., in solid, semi-solid, liquid or soluble form, or as a homogenate or extract, or as a (aqueous) dispersion or suspension, or as (partly) isolated or purified and/or pre-treated (e.g., by dilute-acid pre-treatment, sodium hydroxide pretreatment, lime pre-treatment, pre-treatment with organic solvents with water, 5 hydrothermal processes, or AFEX) lignocellulosic substrates or cellulosic substrates or cellulose, or in other suitable forms used *per se* for depolymerisation and/or fermentation of cellulosic materials and biomass. Accordingly, the invention also contemplates starter cultures comprising the microorganisms of the invention, optionally in combination with one or more other desired microorganisms. Biotechnological approaches for cellulose utilisation 10 are reviewed *inter alia* in Lynd *et al.* 2002 ("Microbial cellulose utilization: fundamentals and biotechnology". *Microbiol Mol Biol Rev* 66: 506-77), and in Himmel *et al.* 2007 ("Biomass recalcitrance: engineering plants and enzymes for biofuels production" *Science* 315: 804-807).

15 The above aspects and embodiments are further supported by the following non-limiting examples.

EXAMPLES

Example 1: Identification of the novel DZ domain of Cel5H

The published analysis of the putative cellulases produced by *S. degradans* (Taylor *et al.* 20 2006, *supra*) indicated the following organisation for Cel5H (from N-terminal to C-terminal ends of the enzyme): signal sequence/catalytic module (GH-family 5)/serine-rich linker/carbohydrate binding module belonging to CBM-family 6/glutamic acid- and proline-rich-linker, in other denotation signal sequence-GH5-PSL-CBM6-EPR.

We have performed a more thorough sequence analysis which revealed the existence of a 25 novel about 100-aminoacid long domain termed DZ domain at the C-terminus of Cel5H, *i.e.*, from position 496 to position 596 in SEQ ID NO: 2 (**Fig. 1B**). Hence, the organisation of Cel5H can be denoted as signal sequence-GH5-PSL-CBM6-EPR-DZ. Since the DZ domain is rich in aromatic residues (Trp, Tyr) it may plausibly represent a new type of carbohydrate-binding module (CBM).

A search for homologous domains in non-redundant DNA sequence database revealed a highly homologous domain in the putative cellulase encoded by the *ac/A* gene from the bacterium *Pseudomonas* sp. ND137 (positions 465 to 558 of SEQ ID NO: 4, **Fig. 1D**). Other proteins containing this particular domain originate from *S. degradans*.

5 Another atypical feature of Cel5H is related to the GH5 catalytic module of which the length is estimated to 300 amino acids whereas other family-5 enzymes display catalytic modules of 380-450 residues. This module is therefore unusually short and lacks an important stretch of highly conserved residues among family-5 enzymes.

10 **Example 2: Production of entire (wild-type), truncated Cel5H, and engineered forms of Cel5H in *E. coli***

Using molecular biology techniques the DNA encoding the mature form of Cel5H (without signal sequence) was amplified and cloned in an *E. coli* expression vector (pET22b(+), Novagen). The resulting vector was used to transform the *E. coli* strain BL21 (DE3) 15 (Novagen). Similarly, modified genes encoding truncated or engineered forms of Cel5H were obtained by PCR and cloned in pET22b(+), prior to transformation of *E. coli* strain BL21 (DE3). In all cases, six His codons were grafted at the C-terminus extremity of the recombinant proteins to facilitate their purification on Nickel resin (Ni-NTA, Qiagen). The various engineered forms of Cel5H are summarised in **Figure 2**.

20 The recombinant strains were grown in Luria Bertani medium and the expression of the cloned genes was triggered using IPTG as the inducer. The synthesis of the recombinant Cel5H was verified by denaturing polyacrylamide gel electrophoresis (SDS-PAGE). The cultures were centrifuged and the harvested cells were broken in a French press.

Example 3: Purification and characterisation of the various forms of Cel5H and evaluation of activity on various substrates including carboxymethyl cellulose (CMC), phosphoric acid-swollen cellulose (PASC), Avicel and a raw substrate, the hatched straw

5 Cel5H, Cel5Ht and Cel5Ht-CBM3a were purified by loading the crude extract on Ni-NTA (Qiagen), and elution of the protein of interest using increasing concentrations of imidazolium. Purification was achieved using FPLC Q-sepharose (Hitrap Q HP resin, GE Healthcare) as described below.

10 Alternatively Cel5H but also Cel5H-CBM3a and CBM3a-Cel5H constructs were purified using a different procedure. The crude extract containing these proteins was incubated with PASC at 4°C, and centrifuged. The cellulose pellet was washed several times by centrifugation in phosphate buffer, and the proteins specifically bound to the cellulase were eluted from the matrix using pure water or 1% triethylamine. The purification was achieved by loading the fraction containing the protein of interest onto an anion exchanger resin 15 (Hitrap Q HP, GE healthcare) at pH 8 using an FPLC device. The protein was eluted from the column using a linear gradient of NaCl. In all cases, purity was checked by SDS-PAGE, and the protein concentrated, divided into aliquots and stored at -80°C.

20 Activity of the purified enzymes was tested on CMC, PASC, Avicel, and hatched straw using standard conditions (37°C).

For testing of activity on CMC, the enzyme concentration was 1 nM and substrate was at 8 g/L CMC (pH 6.0). Activity of the various forms of Cel5H against CMC at 37° is shown in **Figure 3**. The activity pattern obtained in all cases suggests an exo mode of action for Cel5H rather than endo mode of action. As shown in **Figure 3**, all enzymes are 25 characterised by an important release of sugars within the first minutes or seconds of the kinetics, then the enzymatic activity is considerably reduced (observed plateau after 5 minutes), which is more typical of an exo-acting or processive enzyme.

Based on initial velocity, the specific activity of Cel5H was estimated at 20,000 iu/µmol on CMC.

For testing of activity on PASC, the enzyme concentration was 5 nM and substrate was at 3.5 g/L (pH 6.0). Activity of the various forms of Cel5H against PASC at 37° is shown in **Figure 4**.

5 Removal of the DZ domain (Cel5Ht) reduced the specific activity by approx. 70%, showing that this domain plays an important role during the hydrolysis of PASC by Cel5H. Based on these kinetics, the specific activity of Cel5Hwt on PASC is approx. 1,000 iu/μmol. Grafting a CBM3a at the N-terminus (CBM3a-Cel5H), the C-terminus (Cel5H-CBM3a) or instead of the DZ module (Cel5Ht-CBM3a) further reduced the activity of the enzyme.

10 For testing of activity on Avicel, the enzyme concentration was 100 nM and substrate was at 3.5 g/L (pH 6.0). Activity of the various forms of Cel5H against Avicel at 37° is shown in **Figure 5**.

15 As observed for the two other substrates, the activity of wild type Cel5H is higher than the engineered forms, and especially compared to the truncated form that lacks the C-terminal DZ domain, further confirming the role of the DZ domain in degradation of crystalline cellulose. Its replacement by the powerful CBM3a only partly restores the avicelase activity compared to wild type Cel5H. The activity of Cel5H on Avicel is elevated, being calculated on data obtained during the first hour of the kinetic reveals as a specific activity of 8.5 ± 2 iu/μmol.

20 The enzymatic activity of Cel5H was also tested on the soluble chromogenic substrate, para-nitrophenyl-β-D-celllobioside. Upon cleavage by the enzyme of the link between the para-nitrophenyl group and the cellobiose, para-nitrophenol and cellobiose is liberated and its concentration can be directly determined by means of spectrophotometric analysis. The specific activity was about 14 ui/μmol for Cel5H on this substrate. This activity is about 13 times higher than the activity of Cel5A from *C. cellulolyticum*, which also belongs to the 25 family 5 of glycoside hydrolases.

Example 4: Comparison of Cel5H avicelase activity with wild type and engineered cellulases from *C. cellulolyticum*

Using an assay design as described in example 3, the Cel5H avicelase activity was compared with wild type and engineered cellulases from *C. cellulolyticum*.

5 The enzyme concentration was 100 nM and substrate was at 3.5 g/L (pH 6.0).

As shown in **Figure 6**, Cel5Hwt is almost 6-fold more active than Cel9G, one of the most active cellulases from *C. cellulolyticum* on Avicel. Compared to the engineered form of

Cel9G which contains a CBM3a and an X2 module (CBM-X-Cel9G, Mingardon *et al.* 2007.

Appl Environ Microbiol 73: 7138-49.), Cel5Hwt still remains 25 % more active on crystalline

10 cellulose. On the contrary, the covalent cellulosome (which includes the catalytic modules of Cel9G and Cel48F and the powerful CBM3a; Mingardon *et al.* 2007, *supra*) is merely 36% more active than Cel5H.

Analysis by dionex of the soluble sugars released by Cel5Hwt on Avicel indicates that after 24 hours of incubation the latter enzyme mainly releases cellobiose (67 %), cellotriose (26.5

15 %) and glucose (6.5 %), *i.e.*, soluble sugars of relatively low complexity, but no cellotetraose.

The activity of Cel5H (Cel5Hwt) was also tested on hatched straw (enzyme and substrate at 100 nM and 3.5 g/L, respectively). After 24 hours of incubation, approx. 70 µM of soluble sugars were released by Cel5Hwt. Analysis of the soluble sugars by dionex indicated that

20 glucose (90 %) was mainly released by Cel5H on this raw substrate.

All kinetic experiments described above were performed at 37°C in 20 mM Tris-maleate pH 6.0, 1 mM CaCl₂ (azide 0.01% w/v) buffer. The avicelase activity of Cel5H was also measured in presence of 1 % sodium chloride (171 mM NaCl) in the same buffer. In the presence of NaCl, Cel5H was found to be 10 % more active.

25 The activity of the enzyme is considerably reduced at 50°C compared to 37°C, possibly due to unfolding of the enzyme at 50°C.

In conclusion, Cel5H displays an elevated activity on crystalline cellulose and is therefore a suitable candidate for heterologous production in solvent producing bacteria such as *C. acetobutylicum*, to enable the growth of this strain on crystalline cellulose or amorphous

30 cellulose. Our results suggest the DZ module acts as a cellulose binding module (CBM), which may advantageously anchor the cellulase at the surface of the substrate. Our data

obtained so far indicate that this enzyme is likely to be an exoglucanase or an endoprocessive cellulase (and not an endoglucanase). It is worth noticing that family-5 cellulases are usually described as endoglucanases. Cel5H does not release any cellotetraose from pure cellulose, but produces cellobiose, glucose and cellotriose.

5

Example 5: 4-Grafting accessory domains and/or linkers of Cel5H to other enzyme(s)

The DZ module of *S. degradans* alone or with the upstream CBM6 and optionally a linker (see Cel5Hwt in **Figure 1**) was adjoined to other enzyme(s) that lack such modules, using 10 molecular genetic techniques.

The cellulase selected for these experiments was Cel5A from *C. cellulolyticum*, which analogously to Cel5H also displays a family-5 catalytic module at the N-terminus. Cel5A is a genuine endoglucanase with moderate activity on Avicel (Fierobe *et al.* 1991. J Bacteriol 173: 7956-62). The hybrid enzymes which were constructed are shown in **Figure 7**.

15 The various engineered forms of Cel5A were purified and their activity was tested on Avicel and compared to Cel5Awt and Cel5At (**Figure 8**). DZ

The observed activity profiles indicated that the presence of a DZ domain ameliorates Cel5A activity (i.e. hydrolysis of Avicel by Cel5A). In particular, after 24 hours of incubation, 20 the quantity of sugars liberated by Cel5At-DZ or Cel5At-DZ2 is three times higher than the quantity of sugars liberated by Cel5At. These results suggest that the enzyme is fixed or attached by the DZ domain onto the cellulose. Hence, it is suggested that the DZ domain is a new type of fixation domain.

However, the observed activities are 20% lower than the activity of Cel5At-CBM3. The 25 scaffoldin CBM domain thus appears to be also relevant for the amelioration of Cel5A activity on crystalline cellulose.

The addition of two DZ domains does not significantly increase enzyme activity compared to enzyme constructs having only one DZ domain.

It is also noted that a Cel5A construct having one or more additional DZ domains, is still

30 four times less active than wild type Cel5H cellulase on crystalline cellulose.

Therefore, a Cel5A construct, comprising both the CBM domain (CBM6) and the DZ domain from Cel5H is made. The introduction of a CBM6 domain in between the catalytic Cel5A domain and the Cel5H DZ domain increases enzyme activity with respect to the degradation of crystalline cellulose.

Example 6: Cel5H activity is complementary or additive to the activity of other cellulases on crystalline cellulose, in the free state or incorporated into a hybrid cellulosome

10 Kinetic experiments were performed to investigate whether Cel5H activity is complementary or additive to that of other cellulase enzymes in the degradation of pure crystalline cellulose.

In these experiments, the following enzymes were tested:

- wild type enzymes from *C. cellulolyticum* : Cel5A (endo), Cel9G (endo), Cel9M (endo),

15 Cel9P (endo), Cel48F (processive) et Cel9E (processive).

- Modified enzymes from *C. cellulolyticum* : Cip0-G (endo)

- Enzymes isolated from organisms other than *C. cellulolyticum* : Cel6A (exo) de *Neocallimastix patriciarum* (mushroom).

The amount of soluble sugars that were liberated by the enzymes after 24 hours incubation

20 on Avicel crystalline cellulose (3.5 g/L tampon, 20 mMTris-Maleate pH 6, 1mM CaCl₂ 1) are shown in **Figure 9**. For each combination of enzymes, the first bar indicates the sum of the individual activities of both enzymes, whereas the second bar indicates the activity of a mixture of both enzymes (each enzyme being present at a concentration of 0.1mM).

These results indicate that the activities of other celluloses are in most cases

25 complementary or additive to the activity of the Cel5H cellulose.

Further experiments include the degradation of Avicel by hybrid minicellulosomes comprising Cel5H appended with a suitable dockerin module (such as, e.g., in **Figure 2**), one or two other complementary cellulases bearing cognate dockerins and bound to a hybrid scaffoldin displaying an optional CBM3a module (see **Figure 10**). Increased degradation of substrate compared to the Cel5H or other cellulases alone is demonstrated.

5

Example 7: Production and secretion of Cel5H (wild type, engineered, with modified signal sequence) alone or with other cellulases (and/or scaffoldin) by *C. acetobutylicum*

10 The wild type gene encoding Cel5H was amplified by PCR from the strain *S. degradans*, and cloned in the *E. coli*-*C. acetobutylicum* pSOS952 shuttle vector. This plasmid is an expression vector for *C. acetobutylicum* (Perret *et al.* 2004. J Bacteriol 186: 253-7). The expression of the gene of interest is under the control of the strong and constitutive promoter (*pthl*) of the gene encoding the thiolase enzyme. Two *lac* operators were
15 introduced upstream and downstream of the *pthl* to prevent any expression of the gene in *E. coli*. The vector was subsequently methylated (*in vivo* using the strain ER2275[pAN1] and *in vitro* using the methylase CpG) and used to electro-transform *C. acetobutylicum* strain as formerly described. Recombinant clones were obtained and PCR tests on colonies indicated that the vector remains intact in the solventogenic *Clostridium*. Several clones
20 were grown in 2YTC medium, but no additional CMCase activity was detected in the culture supernatant compared to control strain harbouring the “empty” pSOS952 (CMCase test on plates).

Since the GC content of *S. degradans* is around 46%, whereas *C. acetobutylicum* displays a GC content of 31%, we predicted the wild-type gene encoding Cel5H may not be adapted
25 to *C. acetobutylicum* codon bias. A synthetic gene encoding wild-type Cel5H but adapted to *C. acetobutylicum* codon bias was made and cloned as described above in pSOS952.

The codon optimised Cel5H coding sequence for *C. acetobutylicum* is shown in **Figure 1F** (SEQ ID NO: 6) and the corresponding polypeptide sequence is shown in **Figure 1G** (SEQ ID NO: 7). The sequences contain an engineered, C-terminal 6xHis tag to facilitate
30 detection and/or isolation.

Transformation of *C. acetobutylicum* with this new vector (after methylation) generated recombinant colonies. Analysis of the supernatant after growth in 2YTC revealed an improved CMCase activity, thus showing that active Cel5H was secreted by the recombinant strain. The secretion yield seems to presently remain below 1 mg/L.

5 *S. degradans* is a Gram-negative bacterium which displays two membranes (outer membrane and cytoplasmic membrane) whereas *C. acetobutylicum* (Gram-positive bacterium) only possesses a cytoplasmic membrane. The native signal sequence of Cel5H is thus designed to address secretion across two membranes but may not be suitable for efficient secretion by *C. acetobutylicum*. For this reason, using molecular biology 10 techniques, we replaced the native signal sequence of Cel5H by that of Cel5A from *C. cellulolyticum*, since the latter is well secreted by *C. acetobutylicum* (up to 5 mg/L). We transformed *C. acetobutylicum* with the vector containing the hybrid gene. The secretion 15 yield of Cel5H using the Cel5A signal sequence was investigated. It was shown that secretion levels were increased by the replacement of the native signal sequence of Cel5H by that of Cel5A from *C. cellulolyticum*. Due to the strong Cel5H activity on pNP-cellobioside, we were able to measure secretion levels of about 0.5 mg/L.

A culture of the recombinant bacterial strain was grown in a fermentor in an environment of 3 g/L 2YT-PASC with a pH fixed at 5,5. We measured the degradation of PASC, the consummation of residual cellobiose and bacterial growth. The results are shown in Figure 20 **11 and 12**.

A significant bacterial growth was observed during the first 40 hours until an OD value of 0.45 was reached. Subsequently, cell lysis occurred. At the same time, an increasing degradation of PASC was observed. After 117 hours, 68% of the PASC was consumed but no mono- or oligosaccharides could be detected. This suggested that the sugars liberated 25 by Cel5H by hydrolysis of PASC were immediately consumed by the bacteria. These results showed that we succeeded in constructing a strain which is capable of growing on PASC-supplemented medium.

A second culture of the recombinant bacterial strain was grown in a fermentor in an environment of 2YT (however not containing any PASC) with a pH fixed at 5,5. The goal of 30 this experiment was to check the correlation between bacterial growth and the consummation of PASC. The results are shown in Figure 12. As we expected, no bacterial growth could be observed in this case.

Therefore, we can conclude that the observed bacterial growth is due to the degradation of PASC and subsequent consummation of liberated sugars.

Furthermore, a third culture is set up, to check whether the observed bacterial growth is due to the activity of Cel5H and not to the activity of the endogenous enzymatic system of *C. acetobutylicum*. Therefore, a culture of a strain containing an empty vector (pSOS) is grown in a fermentor in an environment of 3 g/L 2YT-PASC with a pH fixed at 5,5.

Finally, a fourth culture is grown in a fermentor in an environment of 5 g/L 2YT-PASC with a pH fixed at 5,5 to check whether the PASC concentration is the limiting factor for bacterial growth. Thus, a higher concentration of PASC may potentially result in higher OD values (i.e. higher than the maximal OD value of 0.45 observed in the previous experiment).

The gene encoding Cel5H is cloned in combination with genes encoding other cellulases that act independently from Cel5H in the free state. Alternatively, the genes encoding efficient minicellulosomes comprising of Cel5H (appended with a dockerin), a scaffoldin and one or two other cellulase(s) bearing suitable dockerins are cloned and expressed in *C. acetobutylicum*.

When three or four different genes are cloned, a second expression vector called pMTL is also used. This vector was shown to be compatible with pSOS952 in *C. acetobutylicum*. Alternatively, the genes are integrated in the chromosome at specific sites (such as taught in *inter alia* patented PCT/EP/2006/066997), or randomly.

Moreover, cellulases with added modules of Cel5H (particularly DZ and optionally CBM6, see Example 4) are cloned and expressed in *C. acetobutylicum*, to achieve an increased degradation and utilisation of crystalline cellulose.

We use both wild-type and various engineered strains of *C. acetobutylicum*, such as for example strains with modified secretion machinery, optimised for the secretion of heterologous proteins, and/or modified metabolism for optimised production of ethanol (or butanol).

The corresponding recombinant *C. acetobutylicum* are shown to produce and secrete the engineered enzymes, and display ability to grow on and/or degrade one or more cellulosic substrates.

Example 8: Improvement of the secretion of cel5H by *C. acetobutylicum*

Various constructions were made and are schematically represented on **FIG. 13**, wherein modules obtained from different scaffolding proteins were used.

In a first construct, the synthetic polynucleic acid adapted to *C. acetobutylicum* codon bias and encoding the cellulase Cel5H from *Saccharophagus degradans*, was fused to a polynucleic acid encoding a carrier domain and comprising the CBM3a module, the first (Xa) and the second (Xa') X modules obtained from the CipA protein of *C. acetobutylicum* and the signal peptide of the CipC scaffolding protein of *C. cellulolyticum*. Suitable linker sequences are used to link the different modules to one another.

10 The domain structure of the native Cel5H polypeptide can be outlined as GH5-PSL-CBM6-EPR-DZ, wherein GH5 stands for its glycoside hydrolase family 5 domain, PSL for the polyserine linker, CBM6 for carbohydrate-binding module family 6 domain, EPR for the glutamic acid-proline-rich region and, without being limited to this interpretation, DZ represents a C-terminal domain identified by the present inventors as a putative
15 carbohydrate-binding module.

This construct was constructed using Overlap Extension PCR technique, and cloned in the shuttle expression vector pSOS952, that confers resistance to the antibiotic erythromycin, thereby generating the plasmid pSOS952-CBM-Xa-Xa'-5H. The constructs were checked
20 by sequencing, and methylated *in vivo* and *in vitro*. The methylated vector was subsequently used to electrotransform *C. acetobutylicum* strain ATTC 824.

The secretion by *C. acetobutylicum* of the wild-type Cel5H protein appended with the signal peptide of the scaffoldin CipC from *C. cellulolyticum* was 0.5-0.9 mg/L (values based on the activity of the culture supernatant on *para*-nitrophenyl-celllobioside). However, *C. acetobutylicum* strains bearing the two constructs encoding the fusion protein comprising the Cel5H protein linked to the carrier domain secreted a fusion protein containing the cellulase 5H in their growth medium in significantly higher amounts, more particularly up to 6.1 mg/L (value based on the activity of the culture supernatant on *para*-nitrophenyl-Celllobioside, see **Figure 14(3)**). These results demonstrate the production and secretion of
30 Cel5H by *C. acetobutylicum*.

Example 9: Demonstration of the activity of the fusion proteins according to the invention on cellulose.

Using molecular biology techniques the DNA encoding the different protein constructs described in Example 8 was amplified and cloned in an *E. coli* expression vector 5 (pET22b(+), Novagen). The resulting vector was used to transform the *E. coli* strain BL21 (DE3) (Novagen). In all cases, six His codons were grafted at the C-terminus extremity of the recombinant proteins to facilitate their purification on Nickel resin (Ni-NTA, Qiagen).

The recombinant strains were grown in Luria Bertani medium and the expression of the cloned genes was triggered using IPTG as the inducer. The synthesis of the recombinant 10 proteins was verified by denaturing polyacrylamide gel electrophoresis (SDS-PAGE). The cultures were centrifuged and the harvested cells were broken in a French press.

The recombinant proteins were purified by loading the crude extract on Ni-NTA (Qiagen), and elution of the protein of interest using increasing concentrations of imidazolium. Purification was achieved using FPLC Q-sepharose (Hitrap Q HP resin, GE Healthcare).

15 Activity of the purified cel5H enzyme was tested on *para*-nitrophenyl-cellobioside and the results are presented in Figure 15a. Alternatively, the activity was measured on Avicel (microcrystalline cellulose) using standard conditions (37°C). The results are illustrated in Figure 15b.

20 **Example 10: The cloning and expression strategies as described in examples 2 to 7 are used with Cel5H homologues, particularly with ACLA from *Pseudomonas* sp ND137 encoded by the gene *acIA***

The same strategy is applied with enzymes which display an important overall homology with Cel5H from *S. degradans*, and in particular possess a DZ domain.

25 This is exemplified by the enzyme (ACLA) encoded by the gene *acIA* from *Pseudomonas* sp ND137. **Figure 16** compares the general organisation of ACLA with that of Cel5H. The gene encoding ACLA has already been cloned and over-expressed in *E. coli*, but the corresponding enzyme was not fully characterized (Aoki & Kamei 2006. European Journal of Phycology 41: 321-328).

With respect to the high degree of homology between these two enzymes (with the exception of the linkers), they are expected to share similar properties/activities. Consequently, the cloning and expression strategies described in Examples 2 to 7 are applied to ACLA, and producing a further cellulose-degrading system particularly in *C. acetobutylicum*.

Example 11: Evaluation of the activity of ACLA on various substrates including para-nitrophenyl- β -D-celllobioside (pNPC), phosphoric acid-swollen cellulose (PASC) and Avicel

Activity of the purified ACLA enzyme was tested on a variety of substrates using standard conditions (37°C) and compared to the activity of purified recombinant Cel5H on the same substrates.

The enzymatic activity of ACLA was tested on the soluble chromogenic substrate, para-nitrophenyl- β -D-celllobioside (pNPC) and compared to that of Cel5H (at 95% purity). Upon cleavage by the enzyme of the link between the para-nitrophenyl group and the cellobiose, para-nitrophenol and cellobiose is liberated and its concentration can be directly determined by means of spectrophotometric analysis. The enzymatic activity of ACLA on para-nitrophenyl- β -D-celllobioside was 17,9 ui/ μ mol, whereas the enzymatic activity of Cel5H on the same substrate was 18,2 ui/ μ mol. Accordingly, ACLA showed 98% of enzymatic activity compared to Cel5H on para-nitrophenyl- β -D-celllobioside substrate.

For testing of activity on PASC (amorphous cellulose), the enzyme concentration used was 5 nM and the substrate concentration was at 3.5 g/L (pH 6.0). The enzymatic activity of ACLA on PASC was 780 ui/ μ mol, whereas the enzymatic activity of Cel5H on the same substrate was 1600 ui/ μ mol. Accordingly, ACLA showed 49% of enzymatic activity compared to Cel5H on PASC.

For testing of activity on Avicel (crystalline cellulose), the enzyme concentration is 100 nM and substrate is at 3.5 g/L (pH 6.0). The enzymatic activity of ACLA on Avicel was 56 μ M/24h, whereas the enzymatic activity of Cel5H on the same substrate was 140 μ M/24h. Accordingly, ACLA showed 40% of enzymatic activity compared to Cel5H on Avicel.

CLAIMS

1. A recombinant bacterium expressing the Cel5H polypeptide of *Saccharophagus degradans* strain 2-40 or a homologue thereof, or a functional fragment and/or variant of said Cel5H polypeptide or of said homologue, which recombinant bacterium is capable of degrading cellulose.
5
2. The recombinant bacterium according to claim 1 comprising a recombinant nucleic acid molecule encoding the Cel5H polypeptide or homologue thereof, or encoding the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, operably linked to regulatory sequences which allow for expression in said 10 microorganism.
3. The recombinant bacterium according to any of claims 1 or 2, which is a solventogenic bacterium, more particularly an ethanologenic bacterium.
4. The recombinant bacterium according to any one of claims 1 to 3, which is *Clostridium acetobutylicum*.
15
5. The recombinant bacterium according to any of claims 1 to 4, wherein the Cel5H polypeptide or homologue thereof, or the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, is fused to one or more domains heterologous to the Cel5H polypeptide or homologue thereof, chosen from a glycoside hydrolase (GH) catalytic domain, a carbohydrate-binding module (CBM) 20 domain, a cohesin-binding domain such as a dockerin domain, and a hydrophilic (X module) domain of a cellulosomal scaffoldin protein.
6. The recombinant bacterium according to any of claims 1 to 5, wherein the Cel5H polypeptide or homologue thereof, or the functional fragment and/or variant of said Cel5H polypeptide or of said homologue, and optionally the one or more enzymes 25 capable of degrading a substance containing cellulose, are comprised in a hybrid and/or covalent cellulosome or minicellulosome.
7. An isolated domain (DZ domain) of the Cel5H polypeptide of *Saccharophagus degradans* strain 2-40 from amino acid 496 to amino acid 596 of said Cel5H

polypeptide or an isolated domain homologous thereto, or a functional fragment and/or variant of said DZ domain or of said homologous domain.

8. An isolated fragment of the Cel5H polypeptide or of a homologue or variant thereof, having the structure EPR-DZ, CBM6-EPR-DZ or PSL-CBM6-EPR-DZ, wherein EPR stands for a glutamic acid-proline-rich region and PSL stands for polyserine linker.
9. A chimeric polypeptide comprising the isolated domain or functional fragment and/or variant thereof according to claim 7, or the isolated fragment according to claim 8 of the Cel5H polypeptide, and further comprising one or more domains heterologous to the Cel5H polypeptide or homologue thereof, chosen from a GH catalytic domain, a CBM domain, a cohesin-binding domain such as a dockerin domain, and an hydrophilic (X module) domain of a cellulosomal scaffoldin protein.
10. A recombinant nucleic acid encoding the isolated domain or functional fragment and/or variant thereof according to claim 7, or the isolated fragment according to claim 8 of the Cel5H polypeptide, or the chimeric polypeptide according to claim 9.
11. The recombinant nucleic acid according to claim 10, further comprising regulatory sequences which allow for expression in a host bacterium.
12. A host bacterium transformed with the recombinant nucleic acid according to claim 10 or 11.
13. The host bacterium according to claim 12, which is *Clostridium acetobutylicum*.
14. A method for the degradation of a substance comprising cellulose, comprising contacting said substance with the recombinant bacterium according to any of claims 1 to 6 or with the isolated domain or functional fragment and/or variant thereof according to claim 7 or the isolated fragment according to claim 8 of the Cel5H polypeptide, or the chimeric polypeptide according to claim 9, or the chimeric peptide according to claim 9, or the host bacterium according to claims 12 or 13.
15. A method for producing a solvent, fuel or chemical intermediate from a substance comprising cellulose, comprising treating said substance with the recombinant

microorganism according to any of claims 1 to 6 or with the host bacterium according to claim 11 or 12.

16. The method according to claim 14 or 15, wherein said substance comprises or is enriched in crystalline cellulose.

FIG 1A

ATGAAATCAGCAACCACAAATCAATCGAGGGCACGCAGTAGCGCCTTAAAAATAT
GTTGGCGGCATCGCTCGCAGGTTAGGGCTACTATCAGCTTCTGCATTTGCCGAT
GTAGCCCCGCTAACCGTAGACGGCAATAAAATTCTTAGCGGTGCCAGCAAGCCA
GTTTGCCGGTAATAGCTTATTTGGCTAACATGGCTGGGCGGTGAGAAGTAT
TACACGGCCGGTACCGTTGAATGGCTAAAGCAAGACTGGGCAGTAATTAGTTC
GCGCCGCAATGGGTGTCGATGAAAACGGCGGCTACTAGAACGACCCAGCAGGAA
ACAAAGCGAAAGTAACAAACCGTTGTAGATGCAGCCATCGCTAACGATATGTATGTA
ATTATCGATTGGCACAGCCACCACGCCAAGACTACCAAAACCAAGCCATTAGCT
TTTCCAAGATATGGCTCGCACCTACCGTAACAACAACAGTTATATACGAAATT
TATAACGAGCCATTACAGGTTCTGGAGCGGCACCATCAAGCCTACGCAGAAG
CGGTATTGGCGCAATCGCGCAATCGACCCAGATAACCTTATTATTGTGGGCAC
GCCTACTTGGTCGCAGGATGTAGACGTAGCCTCGCGCACCCATCACCGAGTA
CAGCAACATTGCCTACACTATTCACTTTATGCAGGGCACCCACAAACAATCCCTAC
GCGATAAAAGCACAAACCGCATTAAATAATGGTATTGCTTGTGCTACCGAATGG
GGTACAGTAATGCCAACGGTGACGGCGGTGTAGACGCAGCCAAACTGATCGT
TGGATGCAGTTTTAAAGCGAATCATATAAGCCATGCCAACTGGCCTAAACGA
TAAAGCCGAAGGCTCTCTGCATTAAAGCCTGGCTCTAACGCAAACGGCGGCTGG
AGCAATTCCGACTTAACC GCCCTCTGGTACCTATGTTAAAAACTTAATTAAAATG
GAACGACGGCTCACCGAGCAGCAGCTCATCTAGCAGCACCAGTTCTTCAAGC
AGCTCCTCGTCTAGTAGCTCATCTAGCAGCTTCTAGTAGTTCTGGCG
TACCAATTACCGCGCGCATTGAAGCAGAAAACTACGATAGCGCACCGGTAGAA
ACCACTGCAGGTAATAGCGGCTACCCACCAATTGTTGTTGATAAAGGTATGGCG
TAGATGTAGAAAACTCTACTGAAGGTGTTGTAATATTGGCTGGACTGCCAGG
CGAAAAAGTAACTTACAACATTGGCAATGCCGATGGCAGTACGATATTGCATTGC
GCGTAGCCTCTATGGATGCCAACCGTATCTGTGCATGTAACAAACAGCCT
AGCAGATACCGTAACCACACAAGGTGGCGCTGGCAGGCATGGACTACCGAAAC
CATTCTAACGTGTATATCCCCTAACACTCGGTAAATTACCGTTGAGTTACGATAG
TGGCTCTAACCTAAACTTTAAACATTACCGAAAGCTGGGTACCGAACCCACTG
TAGAACCACCCGTTGAGCCGCCAGTAGAACCAACCCGTAGACAACGGTAACCTCC
ATGTAACGACGGTAACTCTACGCTTGCCAACAAACGGCGCCTCCATTAAACCTAAC
CAAGGAGCGTGTGTTAAATAACATTACCGCTGGGCGATATTGTTAGGCACCT
GGAGCGGCAACGGTACCTCGATACGACTAGACTGCAATAACAACGTAAT
GAGTGTATTGCACAAAAACTTAATGACTTACTGCTGTAGACACCCGCAACAATGA
ACTGCGCACACTACATTATGTAACAAAGCCCCTAGCAGCTACACCCTGCAATT
GGTAGCTGGTAG (SEQ ID NO : 1)

FIG 1B

MKSATTNQSRARSSAFKNMLAASLAGLGLLSASAFADVAPLTDGNKILSGGQQASF
AGNSLFWSNNGWGGEKYYTAGTVEWLKQDWGSNLVRAAMGVDENGGYLEDPAGN
KAKVTTVVDAAIANDMYIIDWHSHHAEDYQNQAISFFQDMARTYGNNNNVIEIYN
LQVWSWSGTIKPYAEAVIGAIRAIDPDNLIVGTPWSQDVVASRDPTQYSNIA
YTIHFYAGTHKQSLRDKAQTALNNGIALFATEWGTVNANGDGGVDAETDRWMQFF
KANHIS
HANWALNDKAEGSSALKPGSNANGWSNSDLTASGTYVKNLIK
TWNDGSPSSSSSS
STSSSSSSSSSSSSSSSSSSSSSSGGT
NL
PARIEAENYDSAPVETTAGNSGSPTNCS
YKG
MGVD
VENSTE
GACNIG
WTAA
GEK
V
TYN
IGNAD
GTY
DIAL
RVA
SM
DAG
KRIS
V
H
V
NSL
ADTV
TTQ
GGGW
QAW
TT
T
IS
NVY
IP
SNS
VIT
VE
FY
D
SGS
NL
N
FL
N
I
TE
SS
G
TE
PP
VE
PP
VE
PP
V
D
NG
N
F
P
C
N
D
G
N
S
T
L
A
N
G
A
S
I
N
L
N
Q
G
A
C
V
K
Y
N
H
G
W
G
D
I
R
L
G
T
W
SG
N
G
T
I
R
Y
D
V
L
D
C
N
N
N
V
M
S
D
I
A
Q
K
L
N
D
F
T
A
V
D
T
A
T
M
N
C
A
H
Y
I
Y
V
K
Q
A
P
S
S
Y
T
L
Q
F
G
S
W
(SEQ ID NO: 2)

FIG 1C

FIG 1D

MKHS LHQRFLLPVTLAALSLSASMTNADVAPISTNGNQLLFGGAVDVSAGPSLFWSN
NGWGGEKFYNAAGAVASAQQDWNAEIIRAAAMGVDEPGGYLEDASANLNRVRAVVDA
AIANDMYVIIDWHSHHAESEYTQAAVSFFQQMASEYQGQHDNVIYEIYNEPLSVSWSNTI
KPYAEQVIGAIRAVDPDNLIVVGTPTWSQDVAAANDPITNYNNIAYTLHFYAGHTHQY
LRDKAQYALDMGIPLFVTEWGTVNANGDGGVAYNETNTWMDFLKANNISHANWALN
DKAEGSSALVTGTNPSCGNWADNQYTASGTFVRDIVERDWSDGDPPDPTCTRINMP
GTIEAESFCDMGQTESTTDTGGGLNIGWTDAGDWTSYEVNVPAAGRYKVSYRVA
AQNSGMLQLEAAGGFPTYGSITTPVTGGWQSQTISHEV DLPAGDQDLAIAVVSGG
WNLNWIKVEPAGGSSSSSSSSSSSSSSSSSSSSSTSGCTANATSITGNTITVSEGQ
CIRYEHTWGSQQLGSWSAAAGTTYDVINCNGQVIADVAQVQNGFSTVATGTNHCNLY
VYVKQAPTSFDLQFGSW (SEQ ID NO: 4)

FIG 1E

DVAPISTNGNQLLFGGAVIDSVAGPSLFWSNNGWGKEFYNAGAVASAQQDWNAEII
RAAMGVDEPGGYLEDASANLNRVRAVVDAAIANDMYVIIDWHSHHAESEYTQAAVSFF
QQMASEYQHQHDNVIYEIYNEPLSVWSNTIKPYAEQVIGAIRAVDPDNLIVGTPPTWS
QDVDAAANDPITNYNNIAYLHFYAGTHTQYLRDKAQYALDMGIPLFVTEWGTVNANG
DGGVAYNETNTWMDFLKANNISHANWALNDKAEGSSALVTGTNPSGNWADNQYTA
SGTFVRDIVRDWSDGDPVDPDPTCTRINMPGTIEAESFCMDGQTESTTDGGGLNI
GWTDAGDWTSYEVNVPAAAGRYKVSYRAAAQNSGMLQLEAAGGFPTYGSITTPVTG
GWQSWQTISHEVDPAGDQDLIAVSGGWNLNWIKVEPAGGSSSSSSSSSSSSSS
SSSSSSSTSGCDTANATSITGNTITVSEGQCIRYEHTW GSLQLGSWSAAAGTTYDVIN
CNGQVIADVAQVQNGFSTVATGTNHCNLVYVKQAPTSFDLQFGSW (SEQ ID NO: 5)

FIG 1F

ATGAAAAGTGCAACAACAAATCAAAGTAGAGCAAGAAGTAGTCATTAAAAATAT
GCTTGAGCAAGTTAGCAGGACTTGGACTTTAAGTGCAGTCATTGCAGATG
TTGCACCTTAACAGTTGATGGAAATAAAACTTACTGGAGGACAACAGGCATCA
TTTGCAAGGAAATAGTTATTTGGAGTAATAATGGATGGGGTGGAGAAAAATATTA
TACAGCTGGAACAGTTGAATGGCTAAACAAGATTGGGAAGTAATCTTGTAGAG
CAGCAATGGGAGTTGATGAAAATGGTGGATATTAGAACATCCTGCAGGAAATAA
GCAAAAGTTACAACAGTTGATGCAGCAATAGCAAATGATATGTATGTTATAATA
GATTGGCATAGTCATCATGCAGAAGATTCAAAATCAAGCAATAAGTTTTTCAA
GATATGGCAAGGACTTATGGAAATAATAATGGTATATGAAATATATAATGAA
CCTTACAGGTTCATGGTCTGAAACAATAAAACCTTATGCTGAAGCAGTTAGG
TGCAATAAGAGCAATAGATCCTGATAATCTTATAATAGTTGGAACACCTACATGGT
CTCAAGATGTTGATGTTGCAAGTAGAGATCCTATAACACAATATAGTAATATAGCAT
ATACTATACATTTATGCAGGAACACATAACAAAGTTAACAGGATAAAAGCACA
CAGCATTAAATAATGGAATAGCATTATGGCAACAGAATGGGAACAGTTAATGCA
AATGGTATGGGAGTTGATGCAGCTGAAACTGATAGATGGATGCAATTAA
AGCAAATCATATAAGTCATGCAAATTGGCATTAAATGATAAAAGCAGAAGGATCAT
CAGCACTAAACCTGGAAGTAATGCTAATGGTGGATGGCAAATAGTGTGTTAAC
GCAAGTGGAACTTATGTTAAAATCTTATAAAACATGGAATGATGGAAGTCCTAG
TAGTAGTAGTCAGTAGTACAAGTAGTCATCATCAAGTTCAAGTTCAAGTAG
TAGTTCTAGTCAGTTCTAGTGGAGGAAACAATTACCTGCAAGAAT
AGAAGCAGAAAATTATGATAGTGCACCTGTTGAAACTACAGCAGGAAATTCA
GTCCTACAAATTGTTAGTTATAAAGGAATGGGAGTAGATGTTGAAATTCA
GGTGTGTTAATATAGGATGGACAGCAGCTGGAGAAAAGTTACATATAATAGG
AAATGCAGATGGAACTTATGATAGTCAGTAACTAGCAGTAACTACAGTACT
AAAGAATATCAGTTCATGTTATAATTCAATTAGCAGATAAGTACTACACA
GAGGATGGCAAGCATGGACAACAGAAACAATCAAATGTTATATACCTAGTAAT
AGTGTATAACAGTTGAATTGATAGTGGAAAGTAATTAAATTCTTAATATAA
CAGAAAGTAGTGGAACAGAACCTCCTGTTGAACCACCAAGTAGAACCCAC
ACCTCCAGTAGATAATGGAAATTTCCTGTAATGATGGAAATTCAACACTG
TAATGGTGCAAGTATAATTGAATCAAGGTGCATGTGTTAAATATAATCATGG
GGGAGATATAAGACTTGGAACATGGTCAGGAAATGGAACAATAAGATATGATG
TTGATTGTTAATAATAATGTAATGAGTGTATAGCAGACAGAAATTAAATG
CAGTTGATAACAGCTACAATGAATTGTGCTCATTATATATGTA
GTAGTTTACACTCAATTGGAAAGTTGGCATCATCACCATCATTAAATTAATT
A (SEQ ID NO: 6)

FIG 1G

MKSATTNQSRARSSAFKNMLAASLAGLGLLSASAFADVAPLTVDGNKILSGGQQASF
AGNSLFWSNNGWGKEKYTAGTVEWLKQDWGSNLVRAAMGVDENGGYLEDPAGN
KAKVTTVDAAIANDMYVIIDWHSHHAEDYQNQAISFFQDMARTYGNNNNVIYEIYNEP
LQVWSGSGTICKPYAEAVIGAIRAIDPDNLIIVGPTWSQDVASRDPTQYSNIAYTIFY
AGTHKQSLRDKAQTAQNNGIALFATEWGTVNANGDGGVDAETDRWMQFFKANHIS
HANWALNDKAEGSSALKPGSNANGWSNSDLTASGTYVKNLIKWTNDGSPSSSSSS
STSSSSSSSSSSSSSSSSSSSSSSGGTNLPARIEAENYDSAPVETTAGNSGSPTNCS
YKGMGVDEVNSTEGACNIGWTAAGEKVTYNIGNADGTYDIALRVAAMDAGKRISVHV
NNSLADTVTTQGGGWQAWTTETISNVYIIPSNSVITVEFYDSGSNLNFLNITESSGTEPP
VEPPVEPPVEPPVDNGNFPCCNDGNSTLANNGASINLNQGACVKYNHGWDIRLGTW
SGNGTIRYDVLDCCNNVMSDIAQKLNDFTAVDTATMNCAYIYVKQAPSSYTLQFGS
WHHHHHH (SEQ ID NO: 7)

FIG 2

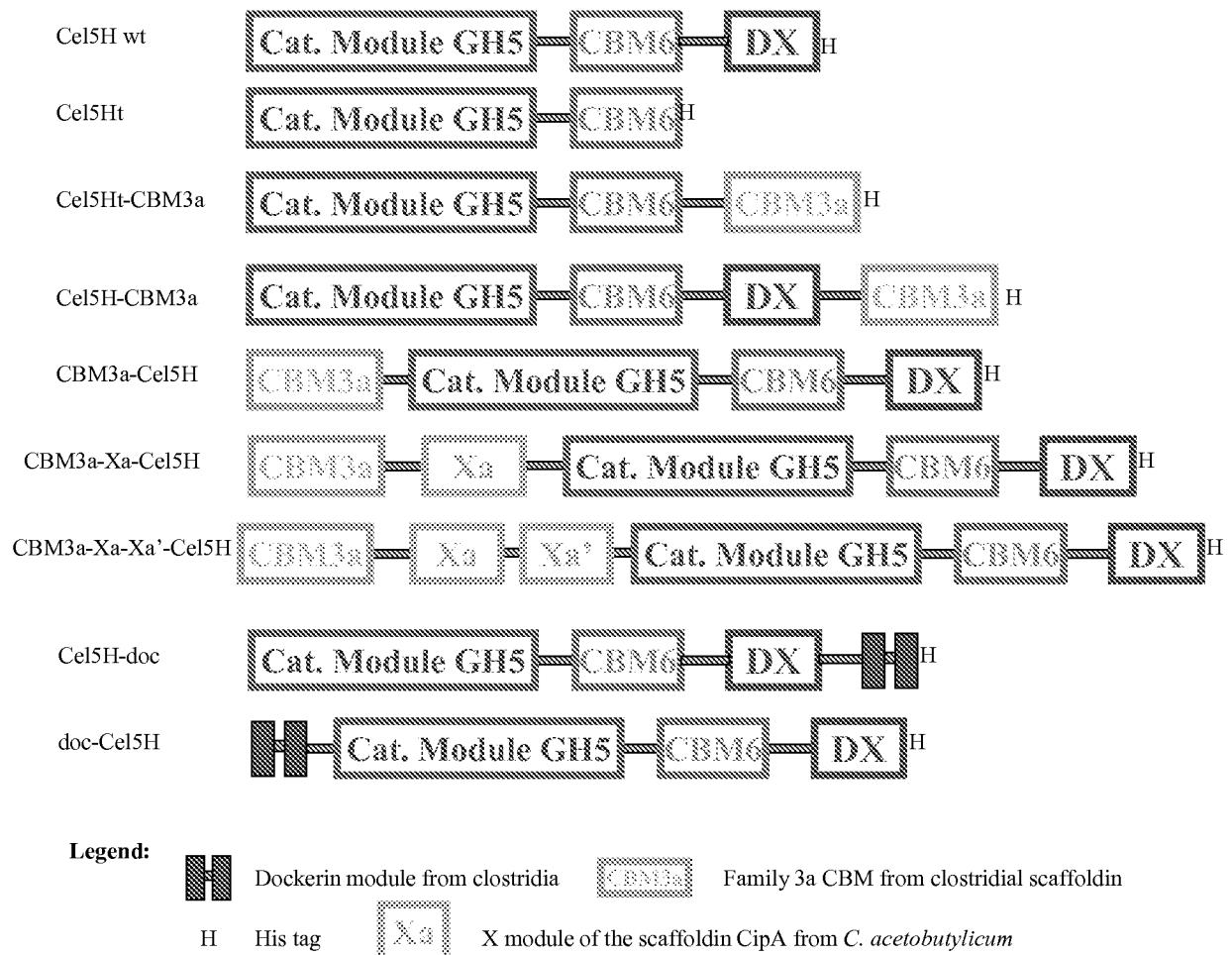


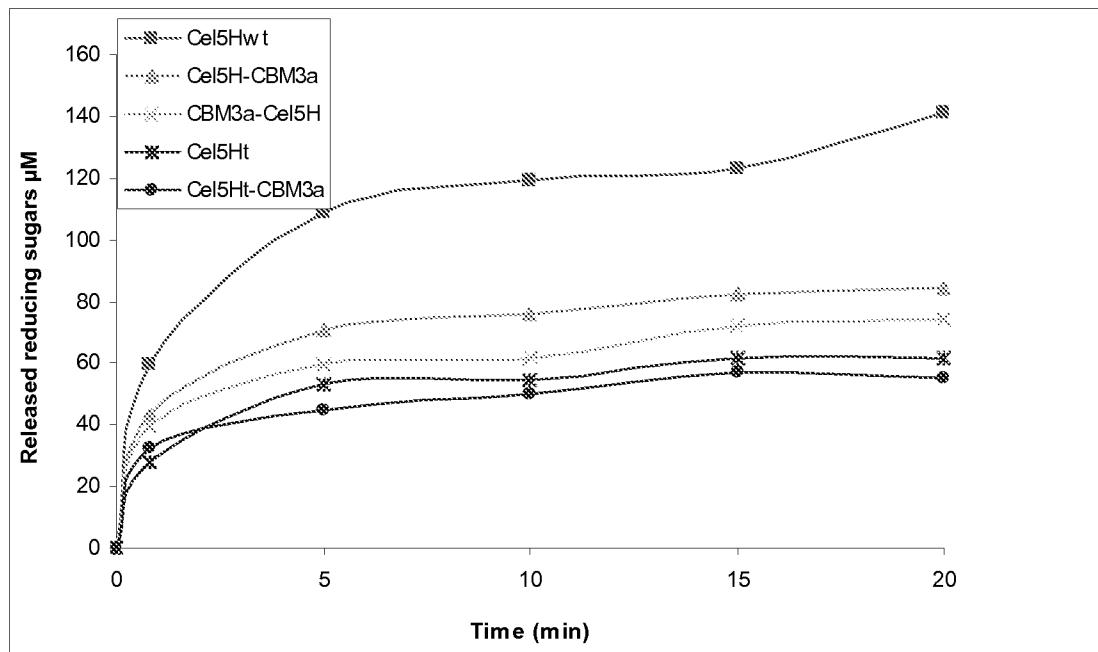
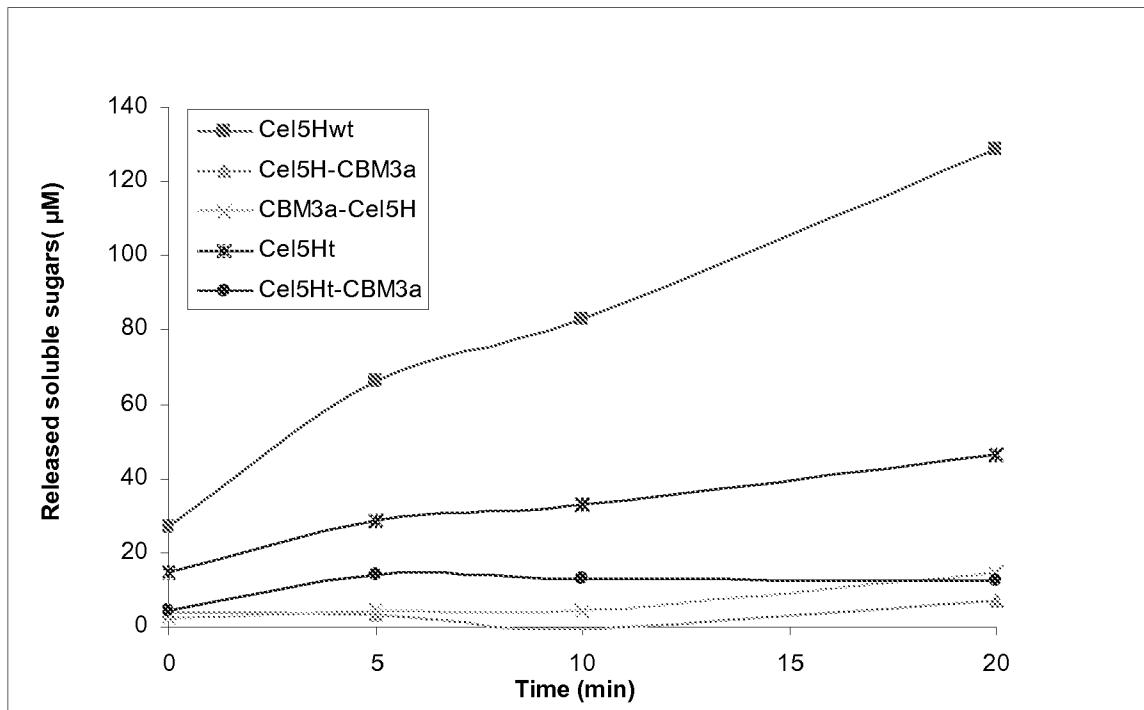
FIG 3**FIG 4**

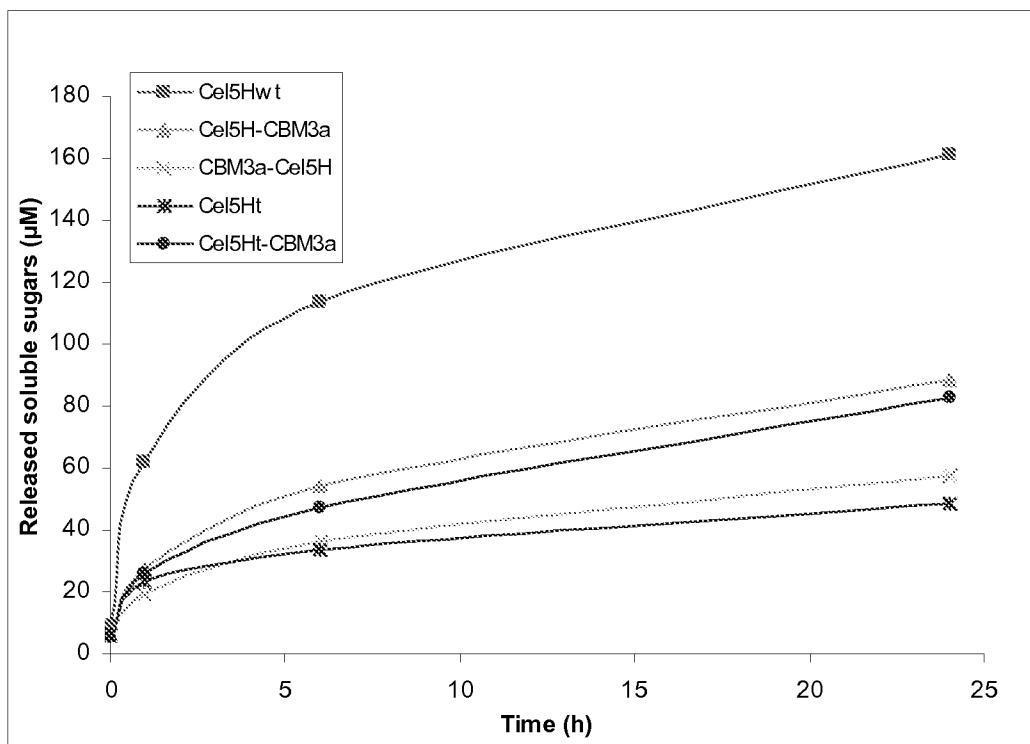
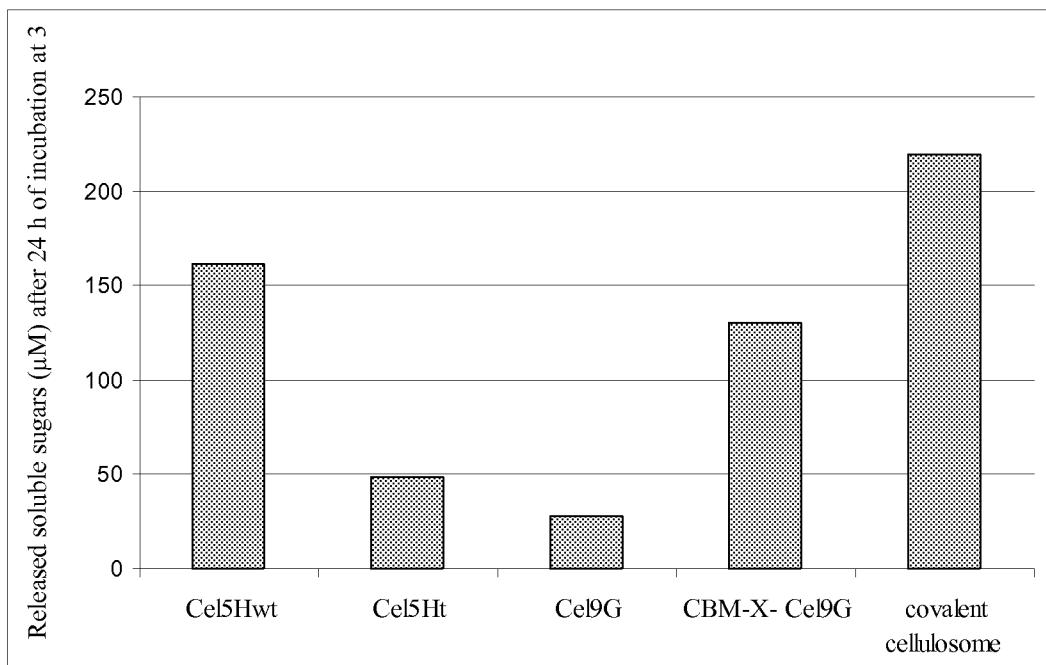
FIG 5**FIG 6**

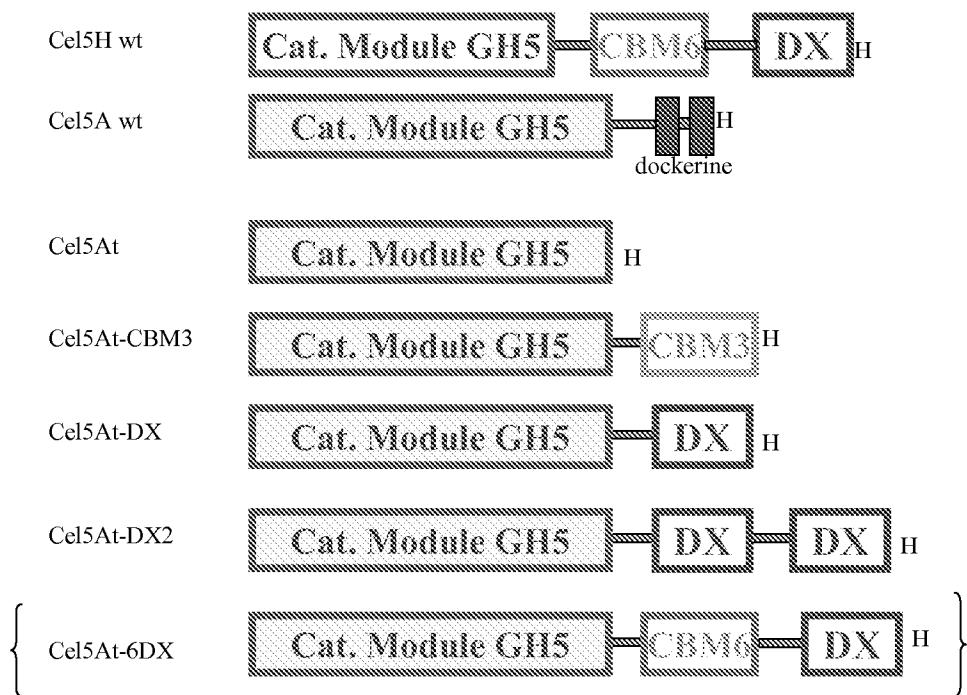
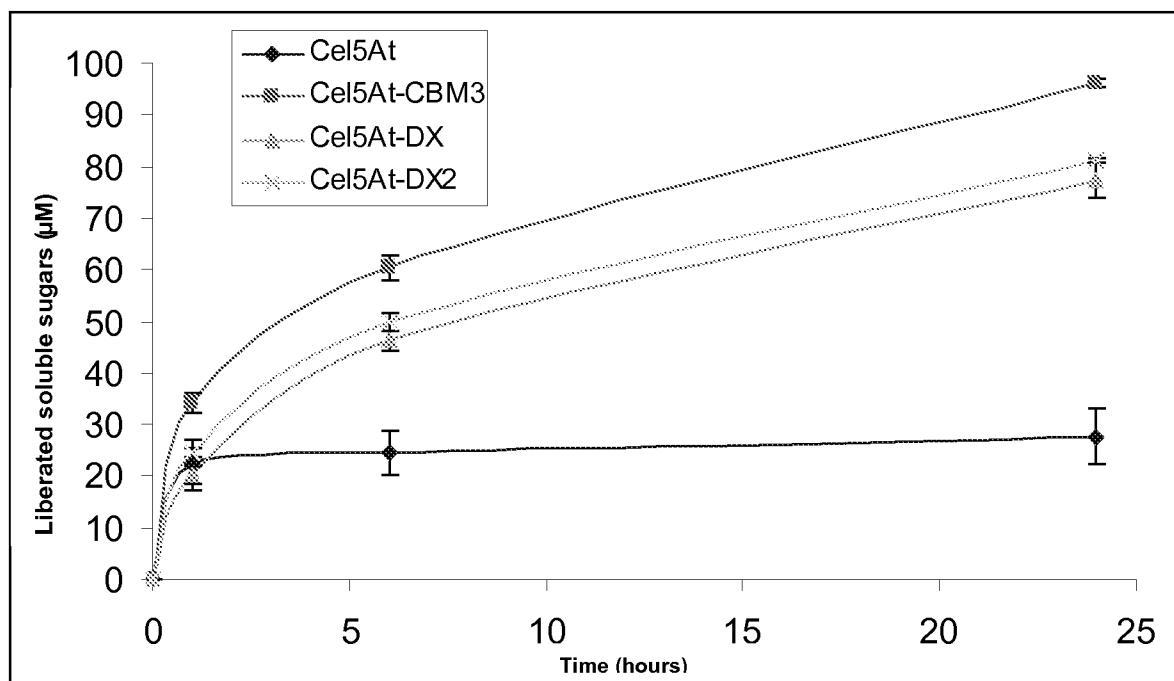
FIG 7**FIG 8**

FIG 9

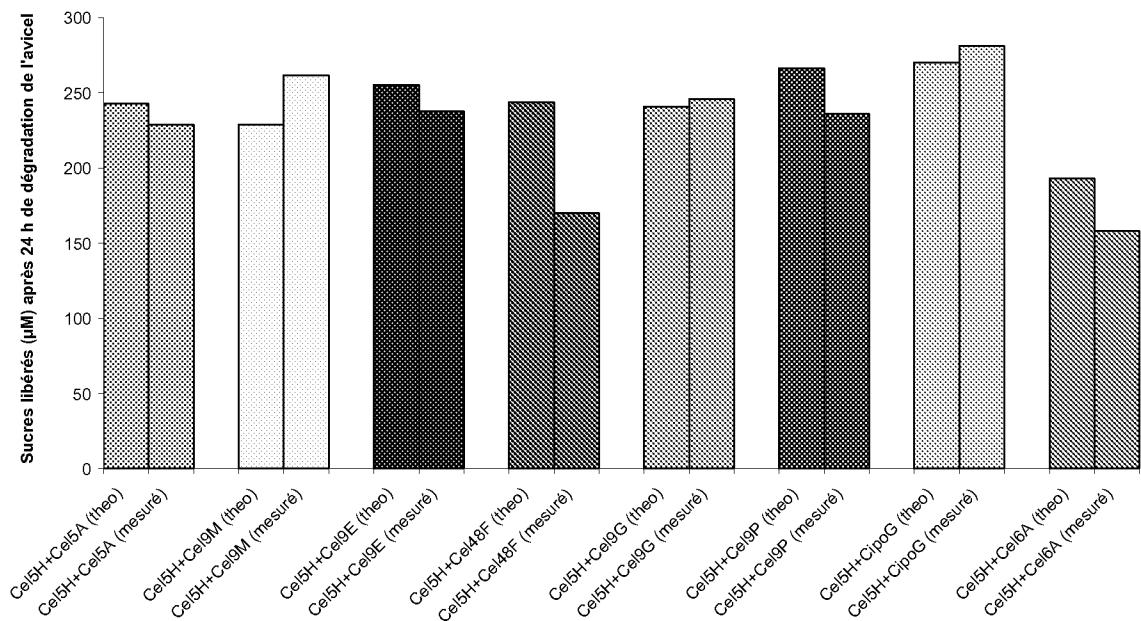


FIG 10

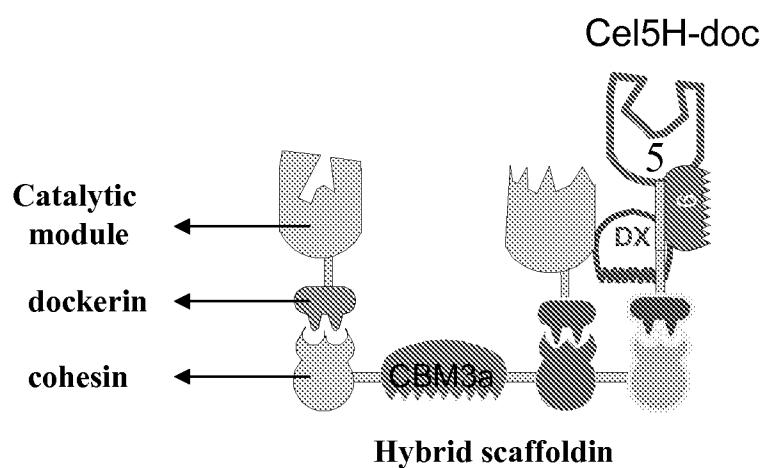


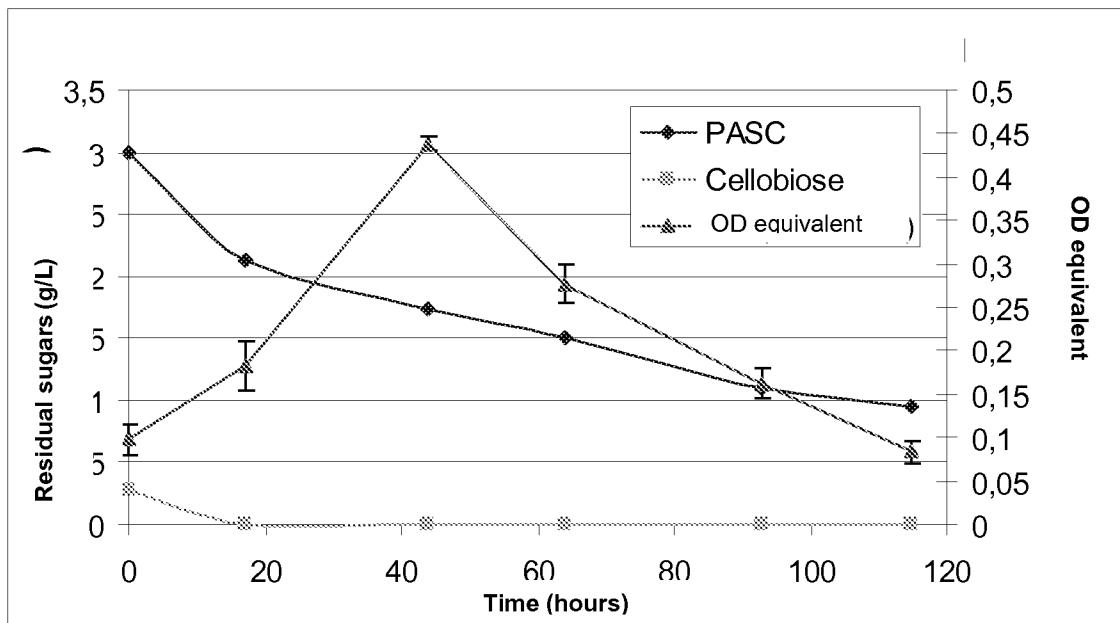
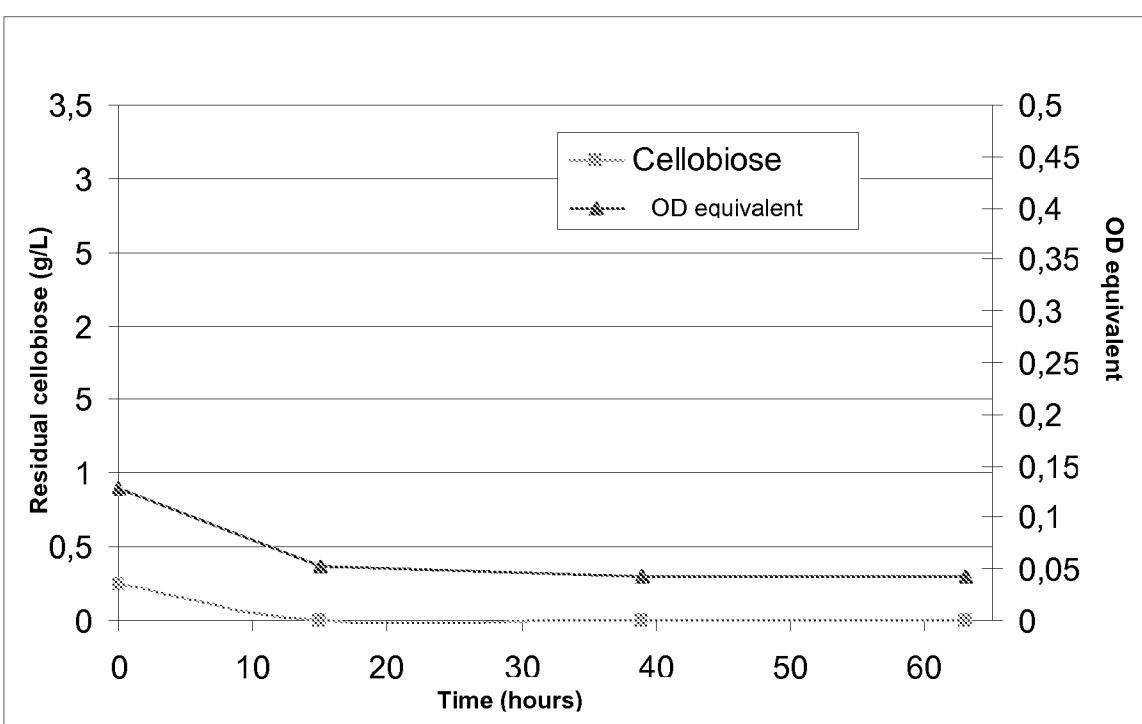
FIG 11**FIG 12**

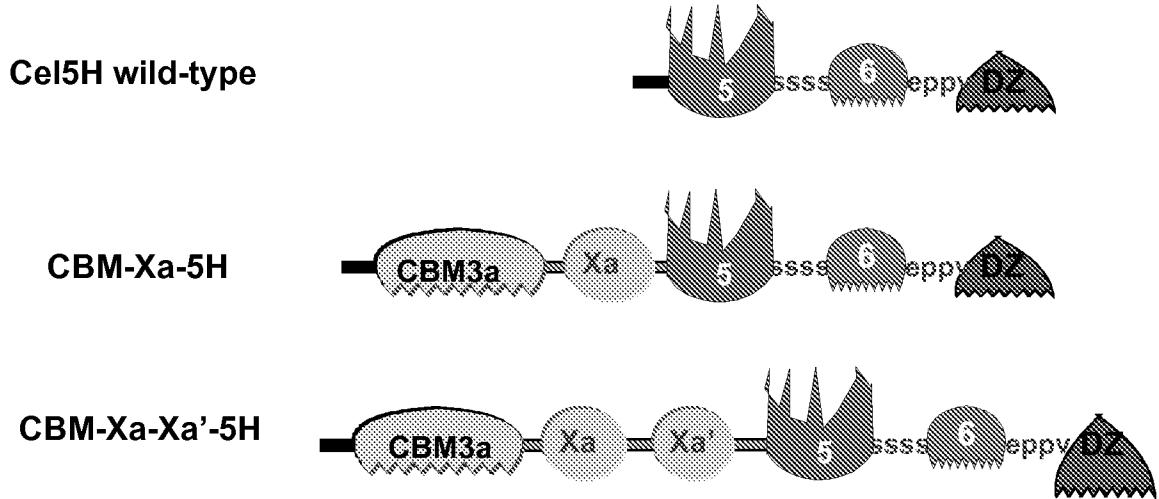
FIG. 13

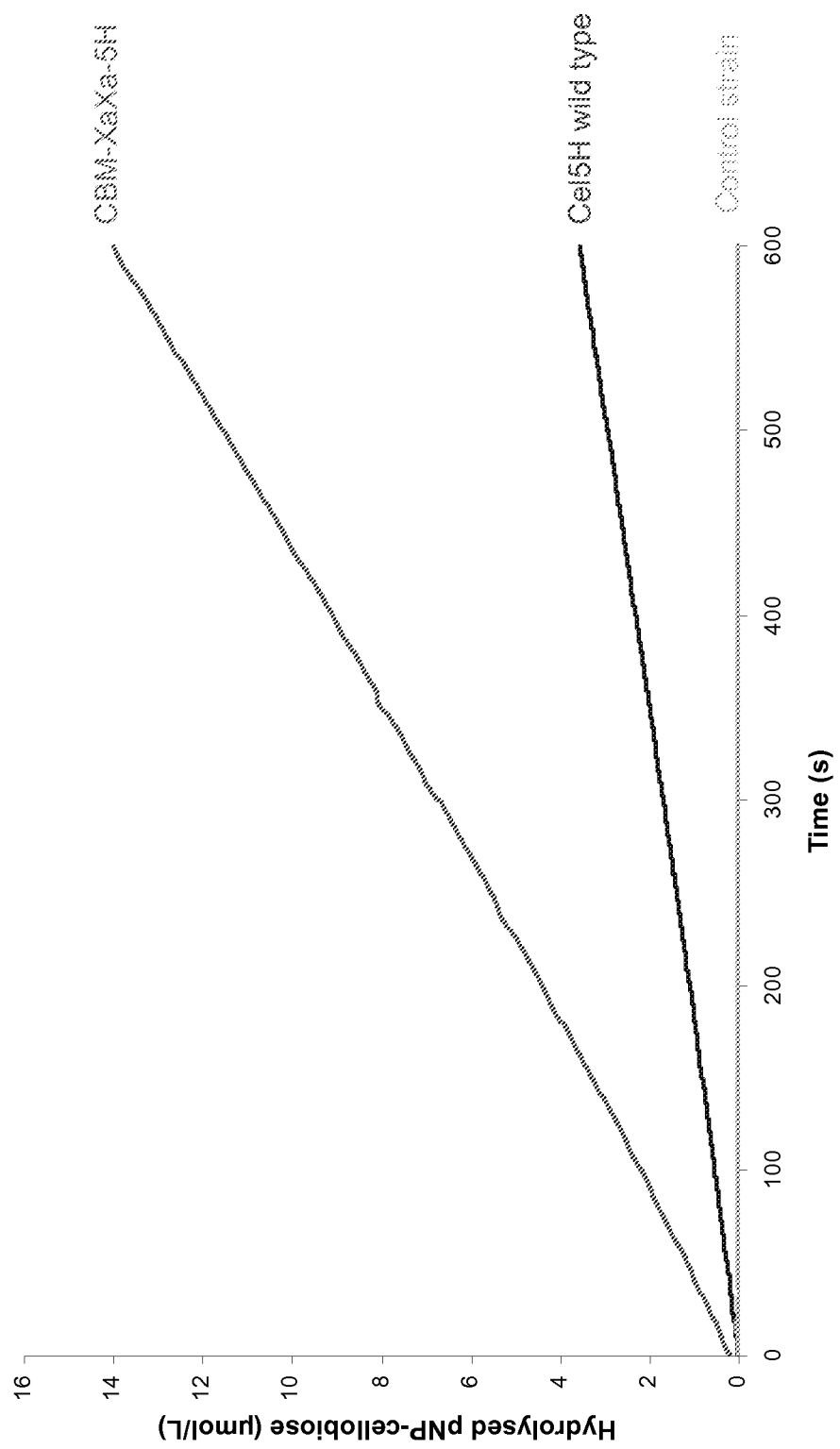
FIG 14

FIG 15a

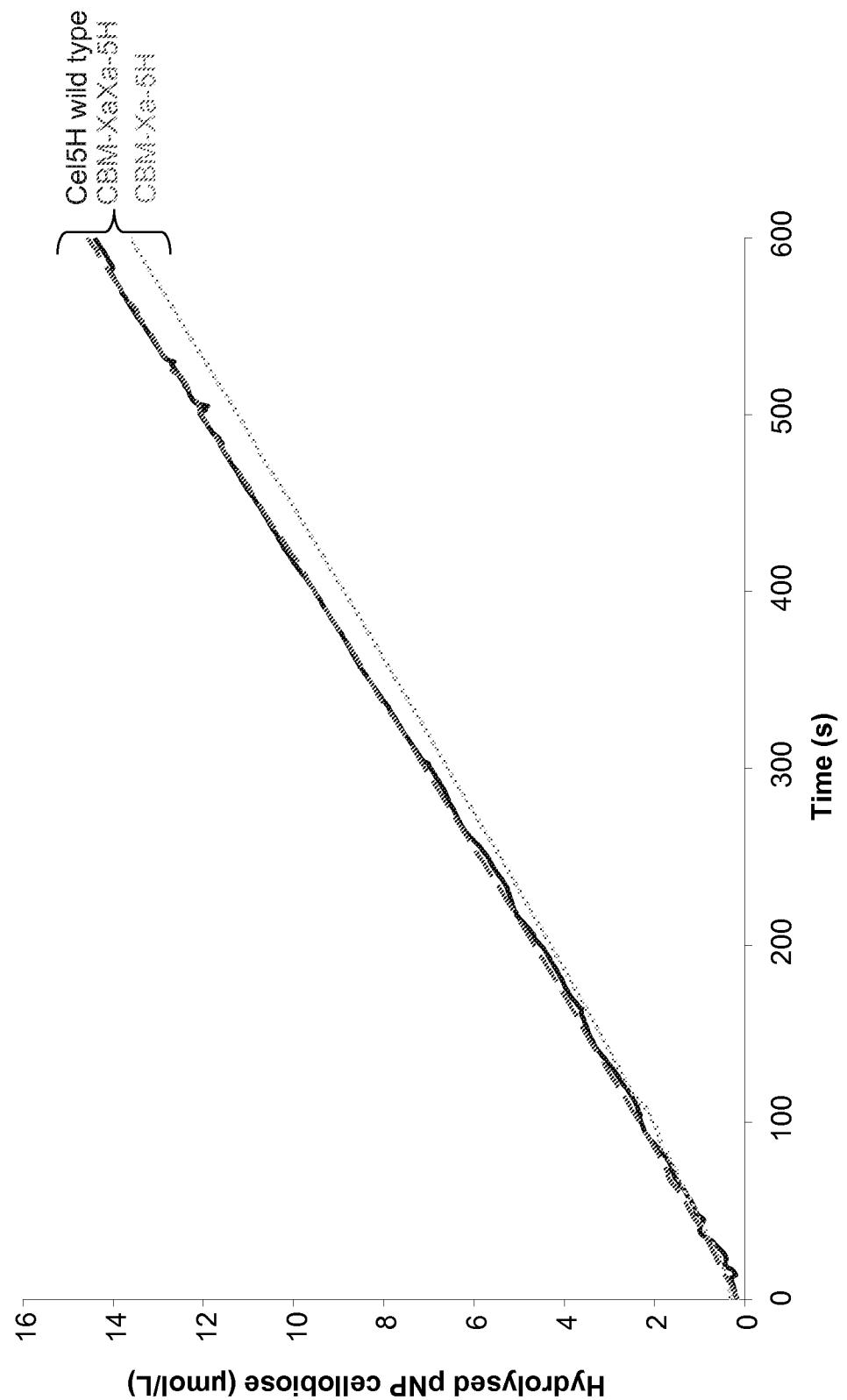


FIG. 15b

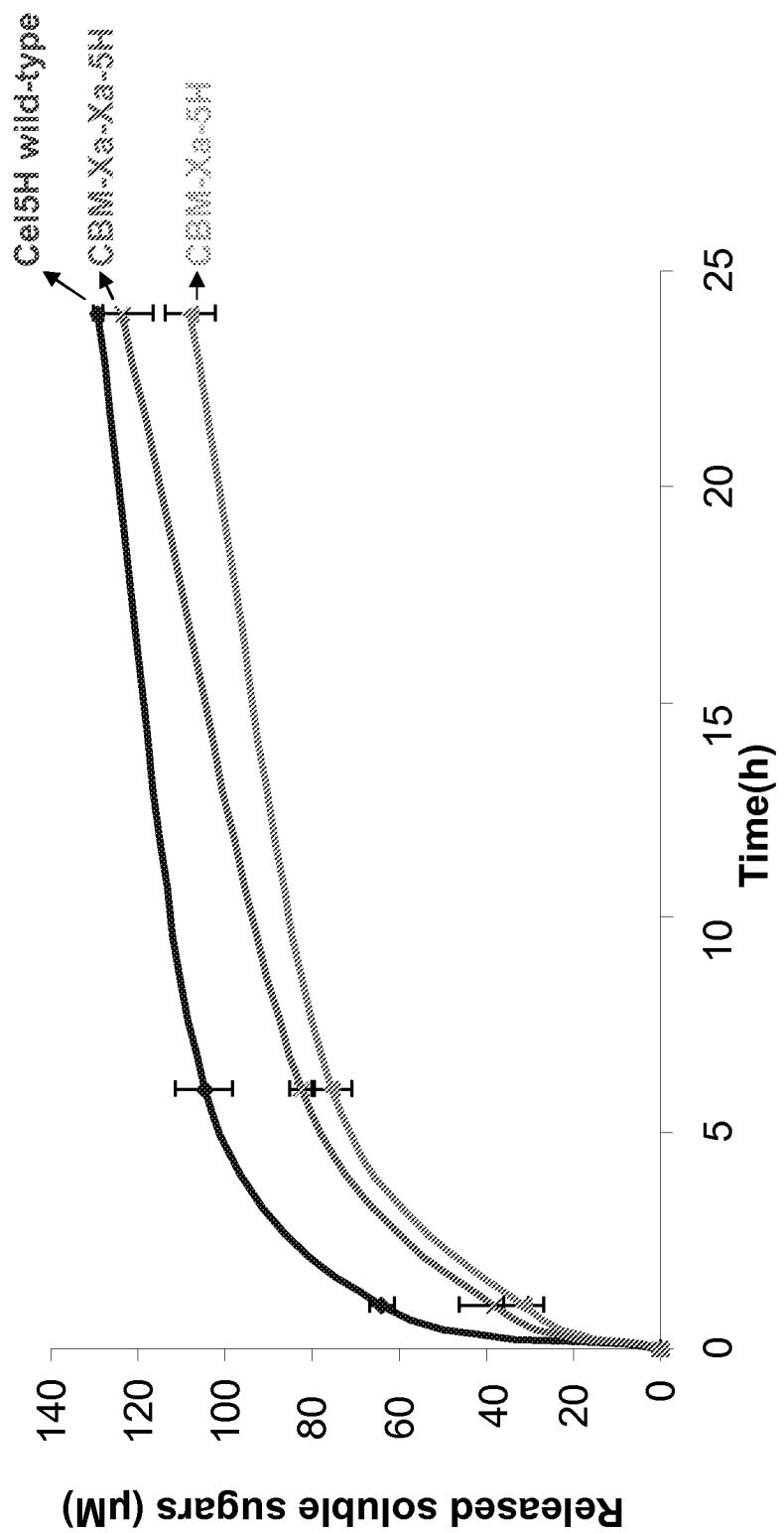


FIG 16

