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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 September 2010 (16.09.2010)

PCT

(10) International Publication Number
WO 2010/102982 A1

(51) International Patent Classification:
C12P 21/02 (2006.01) *C12N 15/67* (2006.01)

(21) International Application Number:
PCT/EP2010/052918

(22) International Filing Date:
8 March 2010 (08.03.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09154783.6 10 March 2009 (10.03.2009) EP

(71) Applicant *(for all designated States except US)*: **DSM IP Assets B.V.** [NL/NL]; Het Overloon 1, NL-6411 TE Heerlen (NL).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LAAN, VAN DER, Jan Metske** [NL/NL]; Leursebaan 364, NL-4839 AP Breda (NL). **WU, Liang** [NL/NL]; Sweelinckstraat 36, NL-2625 VK Delft (NL). **ROUBOS, Johannes, Andries** [NL/NL]; Freule Wttewaall van Stoetwegensingel 45, NL-2642 DB Pijnacker (NL). **PARENICOVA, Lucie** [CZ/NL]; Böttgerwater 44, NL-2497 ZJ Den Haag (NL). **LOS, Alrik Pieter** [NL/NL]; Ko van Dijkstraat 6, NL-2548 ZV 's-Gravenhage (NL). **PEIJ, VAN, Noël Nicolaas Maria Elisabeth** [NL/NL]; Poldermeesterstraat 7, NL-2645 KJ Delfgauw (NL). **PEL, Herman Jan**; Rijksweg 131, 3784 LV Terschuur (NL).

(74) Agents: **HAAFT, TEN, Petrus Johannes Fredrik** et al.; DSM Intellectual Property, Delft Office (600-0240), P.O. Box 1, NL-2600 MA Delft (NL).

(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, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(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).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with (an) indication(s) in relation to deposited biological material furnished under Rule 13bis separately from the description (Rules 13bis.4(d)(i) and 48.2(a)(viii))*

— *with sequence listing part of description (Rule 5.2(a))*

WO 2010/102982 A1

(54) Title: METHOD FOR IMPROVING THE YIELD OF A POLYPEPTIDE

(57) Abstract: The present invention relates to a method for improving protein yield. The method comprises modifying the value of a set of relevant protein features to fall within an optimal range or to become more close to an optimal value for one or more protein features in the eukaryotic host.

METHOD FOR IMPROVING THE YIELD OF A POLYPEPTIDE

Field of the invention

The present invention relates to a method for improving the yield of a polypeptide. In particular, it relates to a method for improving the yield of a polypeptide by modifying the backbone of the polypeptide.

Background of the invention

Recent rapid developments in genome and meta-genome sequencing has resulted in a large number of genes which represent a wealth of potentially very interesting proteins. Problems to express these genes at a significant level hamper the exploration of the functionality of the proteins encoded by those genes and as a consequence prevent the potential exploitation of such proteins in an economical feasible way. Since in many cases the discovered genes originate from organism which are less suitable for large scale production or which are rather inaccessible to the present genetic engineering tools, it is highly desired to use well established production hosts for which gene transfer systems and well developed genetic engineering tools are available. In particular eukaryotes such as filamentous fungi and yeasts are widely used as cell factories in the production of proteins, in particular the production of extracellular proteins. Because of a long tradition of utilization several of these species are generally regarded as safe (GRAS), which makes them very interesting for manufacture of products for human use. However, despite substantial improvements, the production levels obtained for heterologous genes are often much lower than observed for homologous genes. Often there is no expression of protein at all.

Various techniques exist to increase levels of protein production. These include application of strong promoters, increase of copy number, optimal Kozak sequence, mRNA stabilizing elements, optimized codon usage (WO2008/000632) and gene. These strategies however generally do not guarantee that proteins can be produced at detectable levels. To date the most successful approach for producing heterologous proteins is to express them as translational fusion with an efficiently secreted

homologous protein. Nevertheless production levels still lag significantly behind and in many cases expression levels are problematically low. In general low expression in the fermentation leads to lower yields in the recovery. Even if expression is optimized, the final mature protein product may still result in very low production yields due to large losses in the downstream processing. This may be the case when the expressed protein remains associated with the biomass. This results in high losses or alternatively requires use of costly, and sometimes undesirable use of detergents to solubilise the proteins.

Brief description of the figures

Figure 1 Figure 1 depicts a plasmid map of *K. lactis* expression vector pKLPGE-WT (construction described in Example 1). Figure 1 provides also a representative map for other pKLPGE- expression plasmids. Indicated are the LAC4 promoter relative to the PGE encoding gene and the *amd*S selection marker cassette. The *E. coli* DNA can be removed by digestion with restriction enzyme *Sac*II, prior to transformation.

Figure 2 depicts a plasmid map of expression vector pANPGE-3 (construction described in Example 1). Figure 2 provides also a representative map for other pANPGE- expression plasmids. In addition are indicated sequences of the *gla*A promoter and the truncated GlaA and PGE encoding sequences encoding variant PGE enzymes according a method of the invention. The *E. coli* DNA can be removed by digestion with restriction enzyme *Not*I, prior to transformation of the *A. niger* strains.

Figure 3 depicts a plasmid map of expression vector pGBFINZDU-WT (construction described in Example 1). Figure 3 provides also a representative map for other pGBFINZDU-, pGBFINZTB- and pGBFINZTC- plasmids. Indicated are the *gla*A flanking regions relative to the *amd*S selection marker cassette. In addition are indicated sequences of the *gla*A promoter and the ZDU, ZTB and ZTC sequences encoding variant enzymes according a method of the invention. The *E. coli* DNA can be removed by digestion with restriction enzyme *Not*I, prior to transformation of the *A. niger* strains.

Figure 4 SDS-PAGE and western blot analysis of *A.niger* WT6 and the PGE mutant transformants pANPGE12#16 (A) and pANPGE13#30 (B). Supernatant of day 2 (D2) and day 3 (D3) of the cultures was analyzed. The horizontal lines that are at the

14kDa and 97 kDa are for alignment of the SDS-PAGE and Western blot. The marker size on the left-hand side correspond to the SDS-PAGE stained marker and the marker on the right-hand side corresponds to the Western blot marker.

Figure 5 depicts chitinase activity in culture broth of *A. niger* strains after 3 days of fermentation expressing different *ZDU* constructs, all under control of the *gla*A promoter. Depicted is the chitinase activity in culture broth of *A. niger* strains expressing variant SDU constructs wherein signal sequence, N-terminus and protein designs have been modified. Details about the different constructs can be found in Table 6. Relative chitinase activities are depicted as OD590 measurements. For all transformant groups indicated, three transformants were isolated and cultivated independently.

Figure 6 depicts SDS-PAGE analysis of culture broth of *A. niger* WT6 and ZDU strains after 4 days of fermentation expressing variant *ZDU* constructs, all under control of the *gla*A promoter. Details about the different constructs and ZDU proteins expressed can be found in Table 6. For all transformant groups indicated, three transformants were isolated and cultivated independently.

Figure 7 depicts SDS-PAGE analysis of culture broth of *A. niger* WT6 and ZTB-strains after 4 days of fermentation expressing variant *ZTB* constructs, all under control of the *gla*A promoter. Details about the different constructs and ZTB proteins expressed can be found in Table 7. For all transformant groups indicated, three transformants were isolated and cultivated independently.

Figure 8 depicts SDS-PAGE analysis of culture broth of *A. niger* WT6 and ZTC-strains after 5 days of fermentation expressing variant *ZTC* constructs, all under control of the *gla*A promoter. Details about the different constructs and ZTC proteins expressed can be found in Table 8. For the ZTC-WT transformant group indicated, three transformants were isolated and cultivated independently, for the other two strain types two strains.

Figure 9 depicts local protein features.

Description of SEQ ID numbers

SEQ ID NO: 1: cDNA codon-pair optimized (CPO) pregastric esterase (PGE); processed, i.e. without signal sequence coding part

SEQ ID NO: 2: protein calf pregastric esterase (PGE), including signal sequence

SEQ ID NO: 3: DNA PGE protein feature optimized (PFO) variant KL8, 1 extra glycosylation site added

SEQ ID NO: 4: protein PGE PFO variant KL8, 1 extra glycosylation site added

SEQ ID NO: 5: DNA PGE PFO variant KL9, 5 extra glycosylation sites added

SEQ ID NO: 6: protein PGE PFO variant KL9, 5 extra glycosylation sites added

SEQ ID NO: 7: DNA PGE PFO variant KL11, pI shift of 6.96 to 7.74

SEQ ID NO: 8: protein PGE PFO variant KL11, pI shift of 6.96 to 7.74

SEQ ID NO: 9: DNA PGE PFO variant KL12, pI shift from 6.96 to 6.7

SEQ ID NO: 10: protein PGE PFO variant KL12, pI shift from 6.96 to 6.7

SEQ ID NO: 11: DNA PGE, PGE variant with native signal sequence fused to α-MAT factor signal pre(pro-)sequence

SEQ ID NO: 12: DNA PGE AN3, CPO gene tAG fusion with Kex site (KR)

SEQ ID NO: 13: DNA PGE variant AN12, pI shift from 6.96 to 4.6

SEQ ID NO: 14: protein PGE variant AN12, pI shift from 6.96 to 4.6

SEQ ID NO: 15: DNA PGE variant AN13, pI shift from 6.96 to 4.88

SEQ ID NO: 16: protein PGE variant AN13, pI shift from 6.96 to 4.88

SEQ ID NO: 17: DNA chitinase (ZDU) wild-type

SEQ ID NO: 18: protein chitinase (ZDU) wild-type

SEQ ID NO: 19: DNA chitinase variant ZDU-6

SEQ ID NO: 20: protein chitinase variant ZDU-6

SEQ ID NO: 21: DNA chitinase variant ZDU-7

SEQ ID NO: 22: protein chitinase variant ZDU-7

SEQ ID NO: 23: DNA beta-glucosidase wild-type ZTB-WT

SEQ ID NO: 24: protein beta-glucosidase wild-type ZTB-WT

SEQ ID NO: 25: DNA beta-glucosidase variant ZTB-4

SEQ ID NO: 26: protein beta-glucosidase variant ZTB-4

SEQ ID NO: 27: DNA endoglucanase wild-type ZTC-WT

SEQ ID NO: 28: protein endoglucanase wild-type ZTC-WT

SEQ ID NO: 29: DNA endoglucanase variant ZTC-5

SEQ ID NO: 30: protein endoglucanase variant ZTC-5

Detailed description

The present invention relates to a method for improving the secretion of a polypeptide of interest by a eukaryotic host cell by modifying the value of a set of relevant protein features in the amino acid backbone of the polypeptide to fall within an optimal range or to become more close to an optimal value for one or more protein features in the eukaryotic host.

One advantage is that proteins with interesting functionalities which before were not secreted or were only secreted in such low amounts that commercial application was unattractive, now become available for industrial processes because of their improved secretion. Another advantage is that downstream processing and recovery of polypeptides become easier since the designed polypeptides are already separated from the biomass.

In the present context, protein features are properties that can be computationally derived from the protein amino acid sequence and DNA sequence.

Modification of a polypeptide is herein defined as any event resulting in a change in the amino acid sequence of the polypeptide. A modification is construed as one or more modifications. Modification may be accomplished by the introduction (insertion), substitution or removal (deletion) of one or more amino acids in the polypeptide backbone.

In the present context, the term 'secretion' refers to the appearance of a polypeptide in the extracellular medium, typically the growth medium or production medium. The polypeptide which is secreted is free from the biomass. The level of secretion may be measured by methods known in the art, including by activity assays (units of activity), specific activity (units per weight protein), quantitative PAGE analysis, quantitative mass spectrometry and antibody assays.

The expression 'improvement of the secretion of a polypeptide' refers to an increase in the amount of polypeptide which is secreted in the extracellular medium of a cell. The improvement may be reflected by the fact that a polypeptide which is normally not secreted, such as for example an intracellular polypeptide, is now secreted. The improvement may also reside in the fact that a polypeptide which is expected to be secreted, for example because it contains a signal sequence, and which is not secreted,

is now secreted. Improvement is of course always measured with reference to identical host genetic background and identical culture or fermentation conditions. In these cases, improved secretion may be clear from, for example, the appearance of a protein band in a polyacrylamide gel, where there was no band visible before improvement.

Alternatively, the improvement may be reflected by the fact that a polypeptide which is secreted in very low amounts, shows increased levels of secretion.

In one embodiment, the amount of polypeptide secreted is determined by measuring the activity of the polypeptide in the extracellular medium. In comparison to the situation before improvement, the activity in the extracellular medium may be increased by at least 5%, at least 10%, at least 15% or at least 20%. Preferably the activity is increased by at least 25%, at least 30%, at least 35% or at least 40%. In a more preferred embodiment, the activity is at least 45%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, at least 100%, at least 200%, at least 500% or at least 1000% increased. The activity may be increased from no activity to some activity in the extracellular medium.

Any eukaryotic cell may be used in the method of the invention. Preferably, the eukaryotic cell is a mammalian, insect, plant, fungal, or algal cell. Preferred mammalian cells include e.g. Chinese hamster ovary (CHO) cells, COS cells, 293 cells, PerC6 cells, and hybridomas. Preferred insect cells include e.g. Sf9 and Sf21 cells and derivatives thereof. More preferably, the eukaryotic cell is a fungal cell, i.e. a yeast cell, such as *Candida, Hansenula, Kluyveromyces, Pichia, Saccharomyces, Schizosaccharomyces*, or *Yarrowia strain.* More preferably from *Kluyveromyces lactis*, *S. cerevisiae*, *Hansenula polymorpha*, *Yarrowia lipolytica* and *Pichia pastoris*, or a filamentous fungal cell. Most preferably, the eukaryotic cell is a filamentous fungal cell.

"Filamentous fungi" include all filamentous forms of the subdivision *Eumycota* and *Oomycota* (as defined by Hawksworth *et al.*, In, Ainsworth and Bisby's Dictionary of The Fungi, 8th edition, 1995, CAB International, University Press, Cambridge, UK). The filamentous fungi are characterized by a mycelial wall composed of chitin, cellulose, glucan, chitosan, mannan, and other complex polysaccharides. Vegetative growth is by hyphal elongation and carbon catabolism is obligately aerobic. Filamentous fungal strains include, but are not limited to, strains of *Acremonium, Agaricus, Aspergillus, Aureobasidium, Chrysosporium, Coprinus, Cryptococcus, Filibasidium, Fusarium, Humicola, Magnaporthe, Mucor, Myceliophthora, Neocallimastix, Neurospora, Paecilomyces, Penicillium, Piromyces,*

Panerochaete, Pleurotus, Schizophyllum, Talaromyces, Thermoascus, Thielavia, Tolypocladium, and *Trichoderma*.

Preferred filamentous fungal cells belong to a species of an *Aspergillus, Chrysosporium, Penicillium, Talaromyces* or *Trichoderma* genus, and most preferably a species of *Aspergillus niger*, *Aspergillus awamori, Aspergillus foetidus, Aspergillus sojae*, *Aspergillus fumigatus, Talaromyces emersonii*, *Aspergillus oryzae*, *Chrysosporium lucknowense, Trichoderma reesei* or *Penicillium chrysogenum*. When the host cell according to the invention is an *Aspergillus* host cell, the host cell preferably is CBS 513.88 or a derivative thereof.

Several strains of filamentous fungi are readily accessible to the public in a number of culture collections, such as the American Type Culture Collection (ATCC), Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSM), Centraalbureau Voor Schimmelcultures (CBS), and Agricultural Research Service Patent Culture Collection, Northern Regional Research Center (NRRL) *Aspergillus niger* CBS 513.88, *Aspergillus oryzae* ATCC 20423, IFO 4177, ATCC 1011, ATCC 9576, ATCC14488-14491, ATCC 11601, ATCC12892, *P. chrysogenum* CBS 455.95, P*enicillium citrinum* ATCC 38065, *Penicillium chrysogenum* P2, *Talaromyces emersonii* CBS 124.902, *Acremonium chrysogenum* ATCC 36225 or ATCC 48272, *Trichoderma reesei* ATCC 26921 or ATCC 56765 or ATCC 26921, *Aspergillus sojae* ATCC11906, *Chrysosporium lucknowense* ATCC44006 and derivatives thereof.

In one embodiment of the invention, *A.niger* or *K.lactis* is used.

In one embodiment, the eukaryotic cell is a host cell in which the polypeptide is produced by recombinant technology. Suitable methods for transforming or transfecting host cells can be found in Sambrook, et al. (*Molecular Cloning: A Laboratory Manual, 2nd,ed. Cold Spring Harbor Laboratory,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989), Davis et al., *Basic Methods in Molecular Biology* (1986) and other laboratory manuals. Accordingly, the present invention also relates to a method for the production of a polypeptide of interest by applying a method according to the invention to improve the secretion of the polypeptide to the polypeptide of interest and producing the polypeptide modified according to the invention by recombinant technology. The present invention also relates to said recombinantly produced polypeptide. The present invention also relates to a polypeptide obtainable by a method

according to the invention to improve the secretion of the polypeptide; preferably said polypeptide is obtained by a method according to the invention to improve the secretion of the polypeptide.

The polypeptide of interest of which the secretion is improved according to a method of the invention may be any polypeptide having a biological activity of interest. The polypeptide may be a collagen or gelatin, or a variant or hybrid thereof. The polypeptide may be an antibody or parts thereof, an antigen, a clotting factor, an enzyme, a hormone or a hormone variant, a receptor or parts thereof, a regulatory protein, a structural protein, a reporter, or a transport protein such as serum albumin, e.g. Bovine Serum Albumin and Human Serum Albumin, or such as a transferrin, e.g. lactoferrin, a protein involved in secretion process, a protein involved in folding process, a chaperone, a peptide amino acid transporter, a glycosylation factor, a transcription factor, a synthetic peptide or oligopeptide, a protein which in its native form is an intracellular protein and is secreted by methods known in the art such as fusion to a signal peptide and fusion to a polypeptide that is already secreted in its native form. Such intracellular protein may be an enzyme such as a protease, ceramidases, epoxide hydrolase, aminopeptidase, acylases, aldolase, hydroxylase, aminopeptidase, lipase. The polypeptide may be an enzyme secreted extracellularly in its native form. Such enzymes may belong to the groups of oxidoreductase, transferase, hydrolase, lyase, isomerase, ligase, catalase, cellulase, chitinase, cutinase, deoxyribonuclease, dextranase, esterase. The enzyme may be a carbohydrase, e.g. cellulases such as endoglucanases, β-glucanases, cellobiohydrolases or β-glucosidases, hemicellulases or pectinolytic enzymes such as xylanases, xylosidases, mannanases, galactanases, galactosidases, pectin methyl esterases, pectin lyases, pectate lyases, endo polygalacturonases, exopolygalacturonases rhamnogalacturonases, arabanases, arabinofuranosidases, arabinoxylan hydrolases, galacturonases, lyases, or amylolytic enzymes; hydrolase, isomerase, or ligase, phosphatases such as phytases, esterases such as lipases, proteolytic enzymes, oxidoreductases such as oxidases, transferases, or isomerases. The enzyme may be a phytase. The enzyme may be an aminopeptidase, asparaginase, amylase, carbohydrase, carboxypeptidase, endo-protease, metallo-protease, serine-protease catalase, chitinase, cutinase, cyclodextrin glycosyltransferase, deoxyribonuclease, esterase, alpha-galactosidase, beta-galactosidase, glucoamylase, alpha-glucosidase, beta-glucosidase, haloperoxidase, protein deaminase, invertase, laccase, lipase, mannosidase, mutanase, oxidase, pectinolytic enzyme, peroxidase,

phospholipase, polyphenoloxidase, ribonuclease, transglutaminase, or glucose oxidase, hexose oxidase, monooxygenase. The polypeptide of which the secretion is improved may be homologous or heterologous to the host cell. A suitable example of a homologous polypeptide is an *Aspergillus niger* protein which is cloned into and produced by an *Aspergillus niger*. Suitable examples of heterologous expression include a bacterial polypeptide, for example from *E. coli* or *Bacillus*, cloned into and produced by a filamentous fungus or a yeast, or a mammalian protein, for example from bovine or goat, which is cloned into and produced by a filamentous fungus or a yeast, or a filamentous fungal polypeptide which is cloned and produced by a yeast, or a filamentous fungal protein which is cloned into and produced by another fungus. Preferably, the nucleic acids encoding the polypeptides are optimized, for example by codon pair optimization, for expression in the relevant host cell. Codon-pair optimization is a method wherein the nucleotide sequences encoding a polypeptide have been modified with respect to their codon-usage, in particular the codon-pairs that are used, to obtain improved expression of the nucleotide sequence encoding the polypeptide and/or improved production of the encoded polypeptide. Codon pairs are defined as a set of two subsequent triplets (codons) in a coding sequence. Codon-pair optimization is preferably performed as described in WO2008/000632.

Preferably, the specificity of the modified polypeptide is substantially the same as before the improvement of secretion. This means for example that substrate specificity or binding specificity is substantially maintained. In this context, the term “substantially maintained” means that more than 60%, more than 65%, more than 70% or more than 75% of the specificity is maintained. Preferably more than 80%, 85% or 90% of the specificity is maintained. Most preferably, more than 95%, 96%, 97%, 98% or 99% of the specificity is maintained.

According to the method of the invention, the level of activity in the extracellular medium is increased, which is an indication of improved secretion. However, specific activity of the modified polypeptide does not have to be increased, as long as as it is not decreased. Therefore, specific activity is preferably substantially the same as or higher than before the improvement of secretion. In a preferred embodiment, specific activity is substantially the same as before improvement. In this context the phrase ‘substantially the same level of activity’ refers to a level of activity which differs less than 15%, preferably less than 12% or less than 10%, more preferably less than 8%, less than 6% or less than 4% from the level of activity of the parent polypeptide.

In the present context, the terms 'polypeptide' and protein' are used interchangeably. Any type of polypeptide may have its secretion improved by the method of the invention. In a preferred embodiment, the polypeptide is one of the list cited earlier herein.

According to the method of the invention, the value of a set of relevant protein features in the amino acid backbone is modified to fall within an optimal range or to become more close to an optimal value for one or more protein features in the eukaryotic host.

The amount of change of a protein feature between a modified polypeptide and a reference polypeptide can be defined in two ways: relative improvement (RI) and normalized relative improvement (RI_N)

RI of a protein feature is defined in terms of absolute deviation (D) of a protein feature from the optimal value:

$RI = (D_{REF} - D_{PFO}) / D_{REF}$,

where $D = | F_{POI} - F_{OPT} |$, F_{POI} is the value of the feature of the protein of interest, being either the reference or the PFO, F_{OPT} is the optimal feature value.

RI_N is defined in terms of normalized deviation (D_N) to make sense which features matter substantially. D_N takes into account the upper bound (UB) and lower bound (LB) of a feature value (see Table 1).

$RI_N = D_{N, REF} - D_{N, PFO}$,

where $D_N = (F_{POI} - F_{OPT}) / (UB - F_{OPT})$ if $F_{POI} > F_{OPT}$

$D_N = (F_{POI} - F_{OPT}) / (LB - F_{OPT})$, if $F_{POI} < F_{OPT}$

According to the method of the invention, modifications are made to the polypeptide backbone. In this context, the term "backbone" refers to the regular structure which is formed when amino acids are linked together through peptide bonds and form a sequence of covalently linked amino acids. In the present invention, preferably the backbone of the mature polypeptide is modified. In the context of the present invention "mature polypeptide" is defined herein as a polypeptide that is in its final functional form following translation and any post-translational modifications, such as N-terminal processing, C-terminal truncation, glycosylation, phosphorylation, etc. The polypeptide before modification is referred to as the parent or reference or wild-type polypeptide to

distinguish it from the modified polypeptide which results from it. The terms "parent-", "wild-type-" and "reference-polypeptide" are used interchangeably herein. When the polypeptide is a chimeric polypeptide, i.e. a translational fusion with an efficiently secreted polypeptide, preferably a polypeptide native to the host cell, the entire chimeric polypeptide may be modified according to the invention. When the chimeric polypeptide comprises an efficiently secreted polypeptide as a leader polypeptide fused to polypeptide of interest, the polypeptide of interest is preferably modified.

As is known to the person skilled in the art it is possible that the N-termini of the mature polypeptide might be heterogeneous as well as the C–terminus of the mature polypeptide due to processing errors during maturation. In particular such processing errors might occur upon overexpression of the polypeptide. In addition, exo-protease activity might give rise to heterogeneity. The extent to which heterogeneity occurs depends also on the host and fermentation protocols that are used. Such N-terminal and C-terminal processing artefacts might lead to shorter polypeptides or longer polypeptides compared with the expected mature polypeptide.

In one embodiment of the invention, the method comprises:

(i) determining an optimal range and an optimal value for one or more protein features in the eukaryotic host, and

(ii) determining a set of relevant protein features in the eukaryotic host, which features will improve the secretion of the polypeptide by the eukaryotic host if one or more of these relevant features is modified in the amino acid backbone of the polypeptide, and

(iii) modifying the value of the relevant protein features to fall within the optimal range or more close to the optimal value as determined in (i), wherein (i) and (ii) may be performed in any order.

Any method may be used to determine the set of relevant features. In one embodiment, a relevant set of features to improve the secretion of a polypeptide is determined as follows:

(i) collecting or creating a dataset **S**, which contains the secretion levels of a suitable amount of proteins in a certain eukaryotic host and the amino acid and DNA sequences of these proteins. Dataset **S** may contain secreted proteins (**S+**). Preferably, dataset **S** also contains non-secreted proteins (**S-**). For example, one

can express all predicted secreted proteins in *A. niger* (Tsang et al., 2009, Fungal Genetics and Biology, 46: S153-160). The proteins that are secreted belong to the set **S+**, while the proteins that are not secreted belong to the set **S-**. Any method can be used to measure the level of secretion. Alternatively, the set **S-** may contain non-secretary proteins known in the literature in the eukaryotic host. The proteins in **S** may be homologous or heterologous to the eukaryotic host.

(ii) Computing protein features (**F**) for all proteins in the dataset **S**. **F** may be derived both from the DNA sequence and the amino acid sequences of these proteins;

(iii) Using statistical classification methods to select a subset of protein features computed in ii) (**Fs**) that gives the best performance of a statistical classifier to distinguish between **S+** and **S-**, according to a suitably defined classifier performance criterion. **Fs** might be derived both from the DNA sequence (**Fs_DNA**) and the amino acid sequence (**Fs_AA**);

The protein features in **Fs_AA** are the relevant features for modification to improve protein secretion in the corresponding eukaryote host.

Since preferably, the backbone of the mature polypeptide is modified according to the method of the invention, the protein features are preferably computed from a set of mature proteins.

Standard statistical classification methods, which are well known in the art, can be used, such as Linear Discriminant Classifier (LDC), Quadratic Discriminant Classifier (QDC), Nearest Mean Classifier (NMC), 1-/k-Nearest Neighbour classifiers, support vector machine and decision tree, etc (Webb, Statistical Pattern Recognition, 2nd ed, John Wiley & sons).. When applying such methods, the dataset S might be divided into a training dataset and a validation dataset and validation schemes well known in the art (such as 10-fold cross validation) may be used.

Any classifier performance measures known in the art may be used, for example, specificity, sensitivity, accuracy, precision and Area Under the Receiver Operation Characteristics (ROC) curve.

Any suitable method may be used to determine an optimal range or an optimal value of protein features.

In one embodiment, an optimal range or an optimal value of protein features for a eukaryotic host are determined as follows:

i) Collecting or creating a dataset **S**, which contains the secretion levels of a suitable amount of proteins in a certain eukaryotic host and the amino acid and DNA sequences of these proteins. Dataset **S** may contain secreted proteins (**S+**). Preferably, dataset **S** also contains non-secreted proteins (**S-**). For example, one can express all predicted secreted proteins in *A. niger* (Tsang et al., 2009, Fungal Genetics and Biology, 46: S153-160). The proteins that are secreted belongs to the set **S+**, while the proteins that are not secreted belongs to the set **S-**. Any method can be used to measure the level of secretion. Alternatively, the set **S-** may contain non-secretary proteins known in the literature in the eukaryotic host. The proteins in **S** may be homologous or heterologous to the eukaryotic host.

ii) Computing protein features (**F**) for all proteins in the dataset **S**. **F** may be derived both from the DNA sequence and the amino acid sequences of these proteins;

iii) Determining an optimal value (**F_opt**) for each feature for the corresponding eukaryote host: The optimal value may also be obtained by computing measures of central tendency of each protein feature computed from **S+**. Any measures of central tendency can be used, for example, geometric mean, harmonic mean, arithmetic mean, trimmed mean, most frequent value and the median. The computed measure for central tendency is an optimal value for the feature for the corresponding eukaryotic host. Alternatively, fit a probability distribution for each protein feature computed from **S+** such that the distribution of the feature values is well described by the chosen probability distribution. Any probability distribution can be used, for example normal distribution, exponential distribution, or lognormal distribution can be used. The mean of the probability distribution is an optimal value for the feature for the corresponding eukaryote host.

iv) Determining an optimal range of each feature for the corresponding eukaryote host: considering the set **S+** containing only secreted proteins, a lower bound of the optimal range for a protein feature is defined as the value corresponding to the 0.3-, 0.2-, 0.15 or preferably the 0.10- and 0.05-quantile of the protein feature computed from **S+**. Here the value 0.3, 0.2, 0.15, etc. refers to cumulative probabilities. Quantiles corresponding to a certain cumulative probability can be computed by any statistical methods, for example, using the quantile function of

the Statistical Toolbox, Matlab R2007a (The Mathworks Inc). An upper bound of the optimal range of a protein feature is defined as the value corresponding to the 0.7-, 0.8-, 0.85 or preferably the 0.90- and 0.95-quantile of the protein feature computed from **S+**. Alternatively, considering the whole dataset **S** containing both secreted and non-secreted proteins, a lower bound of the optimal range for a protein feature may be defined as a value of the protein feature below which 70%, 80%, 85%, preferably 90% or 95% of the proteins in **S** is not secreted; an upper bound of the optimal range of a protein feature is defined as a value of the protein feature above which 70%, 80%, 85%, preferably 90% or 95% of the proteins in **S** is not secreted.

The set of relevant features and optimal ranges and values will vary from host cell to host cell. For A. *niger* the relevant protein features (**Fs_AA**) to be modified to increase protein secretion include, but are not limited to,: basic amino acid frequency, polar amino acid frequency, non-polar amino acid frequency, tiny amino acid frequency, small amino acid frequency, charged amino acid frequency, net charge (at pH 7.2), isoelectric point, frequency of asparagine, arginine, isoleucine, cysteine, histidine, glutamine, valine, lysine, glycine, threonine and leucine, turn (as calculated by Garnier), PEST motif as calculated by EPESTFIND, local feature (LF) values for pI, in particular LF1 and LF6, LF values for Gravy score, in particular LF2 and LF4, LF values for aroma score, in particular LF3, LF4 and LF6, atomic composition w.r.t. sulphur (S) and localization features (*e.g.* predicted by MultiLoc localization prediction tool).

Net charge has the same unit as the charge of a proton. Net / net positive / net negative / total charge per length have the same unit as the charge of a proton, but normalized to the length of the polypeptide.

The net charge of a polypeptide is herein estimated assuming that all amino acids are fully exposed to the solvent, that neighboring peptides have no influence on the pK of any given amino acid, and that the constitutive amino acids, as well as the N- and C-termini, are unmodified. Different programs can be used to calculate the net charge of a polypeptide at a particular pH (by default pH = 7.2), for example, by using the 'isoelectric' function of the Bioinformatics Toolbox of Matlab (version R2008b), or by using the 'pepstats' function of the EMBOSS Explorer, available at http://emboss.sourceforge.net/.

The net charge per length is herein defined as the net charge of a polypeptide divided by the length of the polypeptide.

The net positive charge per length is herein defined as the net positive charge of a polypeptide calculated by summing up the partial charges of the N-terminus and all lysine, arginine and histidine residues of a polypeptide at pH 7.2. The net positive charge per length is determined by dividing the net positive charge of a polypeptide by the length of the polypeptide.

The net negative charge per length is herein defined as the net negative charge of a polypeptide calculated by summing up the partial charges of the C-terminus and all aspartate, glutamate, cysteine and tyrosine residues of a polypeptide at pH 7.2. The net negative charge per length is determined by dividing the net negative charge of a polypeptide by the length of the polypeptide.

The total charge per length is herein defined as the total charge of a polypeptide calculated by subtracting the net positive charge of the polypeptide (a positive number) by the net negative charge of the polypeptide (a negative number). The total charge per length is determined by dividing the total charge of a polypeptide by the length of the polypeptide.

The gravy score is herein defined as the hydropathy index of a polypeptide as defined by Kyte and Doolittle (1982). Each amino acid has a hydrophobicity score between 4.6 and -4.6. 4.6 is assigned to the most hydrophobic and -4.6 to the most hydrophilic proteins. The GRAVY score of a polypeptide is preferably determined according to Kyte and Doolittle (1982). Kyte, J. and Doolittle, R. 1982 A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157:** 105-132.

The Aroma score of a polypeptide is calculated herein by summing the frequencies of the three aromatic amino acids, Phe, Tyr and Trp in the polypeptide.

The aliphatic index is herein defined as the relative volume occupied by aliphatic side chains. The aliphatic index of a polypeptide (AI) is calculated according to the formula of Ikai (1980): AI = f_Ala + a f_Val + b (f_Ile + f_Leu). Amino acids alanine, valine, isoleucine and leucine have aliphatic side chains.

Where *a* is the relative volume of the valine side chain (a=2.9) and *b* is the relative volume of the leucine and isoleucine side chains (b=3.9). f_Ala, f_Val, f_Ile and f_Leu are frequencies of alanine, valine, isoleucine and leucine in the polypeptide, respectively. Ikai, A.J. 1980 Thermostability and aliphatic index of globular proteins. *J. Biochem.*, **88:** 1895-1898

For GRAVY and aliphatic one could also refer to *Protein Identification and Analysis Tools on the ExPASy Server;* Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.; (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005). pp. 571-607.

Classes of amino acids based on their physio-chemical properties:
Acidic: D, E
Aliphatic: A, I, L, V
Aromatic: F, W, Y
Basic: H, K, R
Charged: D, E, H, K, R
Non-polar: A, C, F, G, I, L, M, P, V, W, Y
Polar :D, E, H, K, N, Q, R, S, T
Small : A, C, D, G, N, P, S, T, V
Tiny: A, C, T, S, G

The features based on the composition of single elements in a sequence are calculated from the frequency *fi* of the element *i*. Frequency and fraction are herein used interchangeably. The frequency is defined as number of times *ni* an element *i* occurs in a sequence divided by the total number of elements in the sequence. Single elements e.g. amino acids in the sequences can be combined to multiple elements e.g. tiny, acidic.

The surface accessibility of an amino acid residue within a polypeptide can be determined by any method known in the art.

If the polypeptide has an experimentally solved structure, the solvent accessible surface area (ASA) is given in Å2 and the area is calculated by rolling a sphere the size of a water molecule over the protein surface [1]. The ASA is then transformed to a relative surface area (RSA), which is calculated as the ASA of a given amino acid residue in the polypeptide chain, relative to the maximal possible exposure of that residue in the centre of a tri-peptide flanked with either glycine [2] or alanine [3]. A residue with an RSA greater than a threshold value alpha (RSA >= alpha, 0<=alpha<=1) is said to be exposed, while a residue with an RSA less than a threshold value beta

(RSA <= alpha, 0<=beta<=1) is said to be buried. Preferably, alpha >= 0.25, more preferably alpha = 0.25. Preferably beta <=0.25, more preferably beta = 0.25.

The surface accessibility can also be predicted from the amino acid sequence of a polypeptide, if the structure of the polypeptide is not available. Different methods are available in the literature to predict the surface accessibility from the amino acid sequence of a polypeptide, for example, as described in [3], [4], [5] and [6]. Preferably, the RSA is predicted using the so-called NetSurfP method described in [4], which can be accessed online http://www.cbs.dtu.dk/services/NetSurfP/. In this application, surface accessibility is predicted from the amino acid sequence of the mature protein. The definition of exposed and buried residues is the same as before.

[1] Connolly M: Analytical molecular surface calculation. Journal of Applied Crystallography 1983, 16(5):548-558.

[2] Chothia C: The nature of the accessible and buried surfaces in proteins. J Mol Biol 1976, 105(1):1-12.

[3] Ahmad S, Gromiha MM, Sarai A: Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003, 50(4):629-635.

[4] Bent Petersen et al: A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Structural Biology 2009, 9: 51.

[5] Dor O, Zhou Y: Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins 2007, 68(1):76-81.

[6] Faraggi E, Xue B, Zhou Y: Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 2009, 74(4):847-856.

Optimal values and ranges for *A. niger* are presented in Table 1.

Table 1A Lower bound (LB), upper bound (UB) and optimal values (F_{OPT}) of protein features

	Whole protein			Mature protein		
	F_{OPT}	LB	UB	F_{OPT}	LB	UB
pI	4.46	3.46	6.21	4.39	3.86	5.26
Net charge	-21.20	-66.10	-4.7	-22.11	-36.26	-9.50
Net charge per length	-0.040	-0.150	-0.01	-0.047	-0.080	-0.024
Net positive	NA	NA	NA	0.063	0.044	0.082

charge per length						
Net negative charge per length	NA	NA	NA	-0.113	-0.137	-0.091
Total charge per length	NA	NA	NA	0.176	0.147	0.212
Gravy score	-0.22	-0.55	0.06	-0.280	-0.439	-0.118
Aroma score	0.110	0.070	0.17	0.111	0.083	0.133
Aliphatic index	0.750	0.450	0.93	0.728	0.606	0.822
Tiny	0.356	0.264	0.505	0.351	0.305	0.430
Small	0.588	0.525	0.714	0.588	0.550	0.664
Polar	0.443	0.398	0.512	0.452	0.425	0.493
Non-polar	0.557	0.488	0.602	0.548	0.507	0.575
Charged	0.181	0.096	0.271	0.187	0.156	0.224
Acidic	0.106	0.075	0.207	0.110	0.088	0.133
Basic	0.075	0.020	0.119	0.077	0.053	0.100
Aliphatic	0.186	0.110	0.236	0.181	0.150	0.208
Ala	0.089	0.042	0.155	0.081	0.059	0.106
Arg	0.029	0.006	0.049	0.029	0.014	0.045
Asn	0.055	0.032	0.099	0.056	0.041	0.074
Asp	0.061	0.040	0.142	0.063	0.048	0.085
Cys	0.013	0.000	0.066	0.013	0.005	0.024
Gln	0.035	0.012	0.09	0.036	0.020	0.050
Glu	0.042	0.019	0.078	0.044	0.031	0.063
Gly	0.087	0.068	0.136	0.088	0.072	0.108
His	0.017	0.000	0.072	0.018	0.008	0.031
Ile	0.047	0.022	0.083	0.047	0.033	0.064
Leu	0.075	0.023	0.12	0.069	0.049	0.091
Lys	0.028	0.000	0.071	0.028	0.016	0.044
Met	0.017	0.002	0.041	0.015	0.005	0.024
Phe	0.038	0.008	0.061	0.038	0.026	0.052
Pro	0.049	0.009	0.129	0.050	0.031	0.069
Ser	0.089	0.047	0.172	0.088	0.064	0.124
Thr	0.080	0.057	0.132	0.080	0.062	0.108
Trp	0.021	0.000	0.045	0.021	0.011	0.033
Tyr	0.048	0.025	0.147	0.050	0.033	0.067
Val	0.063	0.034	0.102	0.063	0.048	0.080

Table 1B Lower bound (LB), upper bound (UB) and optimal values (F_{OPT}) of protein features

	Exposed residues			**Buried residues**		
	F_{OPT}	**LB**	**UB**	**F_{OPT}**	**LB**	**UB**
pI	4.16	3.63	4.93	5.43	4.11	7.47

Net charge	-16.86	-29.98	-6.15	-4.57	-10.83	0.52
Net charge per length	-0.037	-0.066	-0.016	-0.010	-0.023	0.001
Net positive charge per length	0.040	0.024	0.058	0.025	0.015	0.035
Net negative charge per length	-0.078	-0.102	-0.060	-0.036	-0.048	-0.025
Total charge per length	0.118	0.096	0.151	0.061	0.044	0.078
Gravy score	-0.587	-0.726	-0.472	0.314	0.174	0.449
Aroma score	0.019	0.008	0.029	0.094	0.065	0.115
Aliphatic index	0.150	0.096	0.205	0.584	0.462	0.679
Tiny	0.172	0.131	0.256	0.179	0.138	0.207
Small	0.294	0.252	0.383	0.291	0.245	0.331
Polar	0.277	0.243	0.344	0.170	0.127	0.205
Non-polar	0.161	0.129	0.198	0.388	0.338	0.425
Charged	0.119	0.097	0.154	0.065	0.046	0.086
Acidic	0.076	0.057	0.098	0.033	0.021	0.044
Basic	0.044	0.025	0.065	0.032	0.018	0.047
Aliphatic	0.033	0.018	0.049	0.149	0.119	0.176
Ala	0.033	0.021	0.047	0.047	0.030	0.068
Arg	0.015	0.006	0.027	0.013	0.004	0.022
Asn	0.036	0.024	0.049	0.021	0.010	0.033
Asp	0.044	0.031	0.062	0.020	0.010	0.029
Cys	0.000	0.000	0.005	0.011	0.003	0.022
Gln	0.021	0.012	0.035	0.014	0.006	0.021
Glu	0.032	0.020	0.047	0.013	0.006	0.020
Gly	0.041	0.028	0.064	0.046	0.030	0.063
His	0.006	0.002	0.014	0.011	0.004	0.020
Ile	0.006	0.002	0.013	0.040	0.028	0.057
Leu	0.012	0.004	0.021	0.057	0.037	0.075
Lys	0.021	0.010	0.033	0.007	0.002	0.014
Met	0.002	0.000	0.006	0.012	0.004	0.021
Phe	0.004	0.000	0.009	0.034	0.023	0.046
Pro	0.028	0.016	0.042	0.023	0.011	0.033
Ser	0.052	0.032	0.088	0.034	0.019	0.048
Thr	0.045	0.030	0.069	0.034	0.022	0.049
Trp	0.002	0.000	0.006	0.019	0.008	0.029
Tyr	0.011	0.004	0.019	0.039	0.024	0.054
Val	0.013	0.006	0.023	0.050	0.035	0.066

In Table 1, all features computed from the whole protein sequence are based on the length of the whole protein. All features computed from the mature protein sequence,

the exposed residues and the buried residues, are based on the length of the mature protein.

Preferably, the optimal values and ranges features are selected from Table 2; these features are referred to as the primary features, the other features, i.e. the features in Table 1 that are not in Table 2 are secondary features.

Table 2 Primary features

Feature	Computed from			
	Whole protein	Mature protein	Exposed residues	Buried residues
pI	Y	Y	Y	
Net charge (pH7.2)	Y	Y	Y	
Net charge (pH7.2) per length	Y	Y	Y	
Net positive charge (pH7.2) per length		Y	Y	
Net negative charge (pH7.2) per length			Y	
Total charge (pH7.2) per length		Y	Y	
Gravy score				
Aroma score			Y	
Aliphatic index	Y	Y		Y
Tiny amino acid frequency				Y
Small amino acid frequency	Y	Y		Y
Polar amino acid frequency		Y		
Non-polar amino acid frequency		Y		
Charged amino acid frequency	Y	Y	Y	
Acidic amino acid frequency				Y
Basic amino acid frequency	Y	Y	Y	
Arg	Y	Y	Y	
Gln	Y			
Glu	Y	Y	Y	
Lys	Y	Y	Y	
Met			Y	
Phe		Y	Y	

Thr	Y	Y

"Y": indicates that the feature is a primary feature in the corresponding column of either "whole protein" or "mature protein". All features computed from the whole protein sequence are based on the length of the whole protein. All features computed from the mature protein sequence, the exposed residues and the buried residues in the mature protein are based on the length of the mature protein.

For *K.lactis*, the preferred primary features are depicted in Table 3.

Table 3 Primary features and their values for mature proteins in K.lactis

Feature	Optimal value
Glycosylation sites	6
gravy	-0.40
polar amino acid frequency	0.48
nonpolar amino acid frequency	0.52
charged amino acid frequency	0.25
acidic amino acid frequency	0.11
basic amino acid frequency	0.14
Glu	0.053
Lys	0.081
Thr	0.057

In another embodiment, the secretion of the polypeptide is improved by the following steps:

i) computing protein features for the polypeptide,

ii) determining if one or more protein features of the polypeptide are outside the optimal range or substantially deviate from the optimal value for the eukaryotic host, wherein substantial deviation is defined as a difference of 20%, 30%, 40% or more than 50% from the optimal value,

iii) rationally changing the amino acid sequence of the polypeptide, such that the value of one or more **Fs_AA** of the polypeptide falls within the optimal range or is shifted towards **the optimal value** by a suitable amount, preferably a decrease in

the difference between a protein feature of the polypeptide and the optimal value of the protein feature by 10%, 15%, 20%, or more than 30%.

Preferably, 2, 3, 4 or 5 protein features are modified in combination, more preferably, more than 10, 15 or 20 protein features are modified in combination. Most preferably, more than 25 or 30 protein features are modified in combination.

Preferably, the optimal range is taken from Table 1, more preferably, the optimal range is taken from Table 2. Alternatively, the optimal range is taken from Table 3.

In step iii) above, the amino acid sequence of the polypeptide may be rationally changed by any methods known in the art. For example, this may be achieved by:

(i) retrieving homologous sequences;

(ii) aligning the homologous sequences to the sequence of interest;

(iii) identifying amino acids which are crucial for the proteins functional properties;

(iv) introduce desired amino acid sequence features while retaining functional properties;

(vi) translating the final modified sequence back into the gene using the most optimal codons for the given host;

(vii) cloning and expression of the redesigned polypeptide in the host.

Preferably, at least 5% of the amino acids of the amino acid backbone is modified, more preferably at least 10%, even more preferably at least 15%, even more preferably at least 20% of the amino acids of the amino acid backbone is modified.

Preferably, at least 5 amino acids of the amino acid backbone is modified, more preferably at least 10 amino acids, even more preferably at least 15 amino acids, even more preferably at least 20 amino acids, even more preferably at least 25 amino acids, even more preferably at least 30 amino acids of the amino acid backbone is modified.

Preferably, according to the invention, primary features are improved while the secondary features are kept within a certain boundary. Therefore an overall optimality score F is defined based on D_N values of all *n* primary features and all *m* secondary features:

$$F = \left(\sum_{i=1}^{n} |D_{N,i}|^p + \eta \sum_{j=1}^{m} |D_{N,j}|^p \right)^{1/p}$$

η is a weighing factor between and including 0 and 1 ($0 \leq \eta \leq 1$). Preferably $\eta \leq 0.5$, more preferably $\eta \leq 0.4$, most preferably $\eta = 0.3$. p is between and including 1 and 5 ($1 \leq p \leq 5$), preferably $p = 2$ (F represents then the Euclidean distance). Preferably η =0.3 and p =2. Preferably an improvement in F-score of at least 5% with respect to the wild type reference protein is achieved, more preferably at least 10%, even more preferably at least 15%, even more preferably at least 20% and even more preferably at least 30% improvement is achieved.

Preferably, at least 2, 3, 4, or 5 features are modified, more preferably at least 10, even more preferably at least 15, even more preferably at least 20, even more preferably at least 25, and even more preferably at least 30 features are modified. Preferably, at least 2, 3, 4, or 5 features are improved, more preferably at least 10, even more preferably at least 15, even more preferably at least 20, even more preferably at least 25, and even more preferably at least 30 features are improved, whereas preferably, less than 10, even more preferably less than 5, even more preferably less than 4 features are worsened. Preferably, the features are primary features.

Homologous sequences are preferably retrieved by performing BLAST searches of appropriate sequences databases. The homologous sequences preferably have at least 30%, preferably at least 40%, more preferably at least 50%, 60%, 70%, 80%, 90%, 95,%, 96%, 97%, 98% or 99% identity with the sequence of interest. Most preferably, the homologous sequences preferably have about 50% identity with the sequence of interest. The person skilled in the art will be aware of the fact that several different computer programs are available to align two sequences and determine the homology between two sequences (Kruskal, J. B. (1983) An overview of sequence comparison In D. Sankoff and J. B. Kruskal, (ed.), Time warps, string edits and macromolecules: the theory and practice of sequence comparison, pp. 1-44 Addison Wesley). Any method known in the art may be used for alignment. The percent identity between two amino acid sequences or between two nucleotide sequences may for example be determined using the Needleman and Wunsch algorithm for the alignment of two sequences. (Needleman, S. B. and Wunsch, C. D. (1970) J. Mol. Biol. 48, 443-453).

Methods to identify amino acids crucial for essential functional properties of interest are known in the art. Suitable tools include using a 3D structure or a 3D model of

the protein of interest, mutagenesis studies of the protein of interest or of homologous proteins, the use of site saturated libraries to establish functionally neutral substitutions versus functional substitutions.

When introducing amino acid sequence features, substitutions are preferably chosen in such a way that at the given position the amino acid which is more conform the required amino acid sequence characteristics is selected from the group of amino acids which is observed in homologous sequences. State of the art modeling techniques may be applied to identify allowable substitutions which are not observed in natural homologues. Preferred references for modelling techniques which allow the generation of new sequences adopting a given fold are:

Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003). Design of a novel globular protein fold with atomic-level accuracy *Science* 302, 1364-8.

Baker D (2006). Prediction and design of macromolecular structures and interactions. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 361, 459-63 De Novo protein design: towards fully automated sequence selection' *Journal of Molecular Biology*, *Volume 273, Issue 4*, *7 November 1997*, *Pages 789-796* Bassil I. Dahiyat, Catherine A. Sarisky, Stephen L. Mayo

State of the art computational method allow for the generation of numerous potential sequences which may adopt a given protein fold. By introducing feature optimisation into the scoring functions which are used to filter out the most optimal sequences the most optimal sequences for a given production host might be selected in a computational way.

Protein features which may be modified according to the method of the invention include compositional, physiological and structural features. Suitable examples of such features are the number of amino acids, molecular weight, isoelectric point, net charge at a specific pH, GRAVY score, aliphatic index, instability index, compositional features, atomic composition with respect to C, H, N, O, S atoms, amino acid frequency, dipeptide frequency, tripeptide frequency, acidic amino acid frequency, aliphatic amino acid frequency, aromatic amino acid frequency, basic amino acid frequency, glycosylation pattern and charged amino acid frequency and the features as mentioned in Table 1. A combination of modified features is also encompassed by the present invention. Preferably 2, 3, 4 or 5 protein features are modified in combination. More preferably,

more than 10, 15 or 20 protein features are modified in combination. Most preferably, more than 25 or 30 protein features are modified in combination.

In one embodiment of the invention, one or more glycosylation sites are introduced while other protein features are modified as well. In another embodiment of the invention, the charged amino acid frequency is modified while other protein features are modified as well. In another embodiment of the invention, the polar amino acid frequency is modified while other protein features are modified as well.

The protein feature computed from the entire amino acid or DNA sequence is an average value for the entire protein, which may not reveal local protein properties. For example, a protein could be on average hydrophilic but still contain a large internal hydrophobic region. Local protein properties can be computed from the amino acid or DNA sequence, for example, with the method outlined by Benita et al. (Benita et al., 2006. Molecular and Cellular Proteomics, 5: 1567-1580).

To compute the local property of a certain protein feature, one may compute the protein feature locally in a sliding window of a suitable number of amino acids or nucleotides. The obtained value is then plotted as a curve along the length of the amino acid or DNA sequence of the protein as illustrated in Figure 9.

A number of local features can be defined:

Local feature (LF)	Definition
LF1	Area above the curve and below the lower threshold
LF2	Area below the curve and above the higher threshold
LF3	Largest continuous area above the curve and below the lower threshold
LF4	Largest continuous area below the curve and above the higher threshold
LF5	Fraction of the curve below the lower threshold
LF6	Fraction of the curve above the upper threshold

For example, LF1 corresponds to dark-gray colored area in Fig. 1, while LF2 corresponds to light-gray colored area in Fig. 1. The area can be calculated using the trapezoid method (Benita et al., 2006. Molecular and Cellular Proteomics, 5: 1567-1580).

For computing the local features a suitable upper and lower threshold, as well as the size of the sliding window can be chosen. The sliding window can be of any size. For example, a sliding windows size of 21 amino acids or base pairs can be used. The value of the upper and lower threshold can be chosen to reflect extreme peaks in the curve. For example, a higher upper threshold will take more extreme peaks into account than a lower one. Preferably, upper and lower thresholds are chosen such that the Fischer criteria is maximized for the dataset **S+** and **S-**. The Fischer criteria (J_F) is defined as:

$$J_F = \frac{|\mu_{S^-} - \mu_{S^+}|^2}{\sigma^2_{S^-} + \sigma^2_{S^+}}$$

Where μ_S- and μ_S+ represent means of the local feature values computed from the set **S-** and **S+**, respectively, and σ^2_S- and σ^2_S+ the variance of the local feature values computed from the set **S-** and **S+**, respectively.

Local features defined above can be calculated for any protein features, for example the Gravy score, aroma score and the isoelectric point.

In addition to the features that can be derived from the sequences of successfully secreted protein, it was observed that in particular increasing the hydrophilicity of the solvent accessible surface of target proteins was very successful in increasing the amount of soluble protein which was secreted form the cells. More specifically, not only the expression was increased, but also significantly more protein accumulated in the broth in a soluble form not attached to the biomass or other insoluble material. Given proteins with improved surface hydrophilicity could be recovered at significantly higher secretion. Upon removal of the biomass (by filtration or centrifugation) the major part of the produced protein ends up in the filtrate or the supernatant.

In creasing the hydrophilicity can be done by:

- substituting non-polar amino acids by more polar amino acids
- substituting less polar amino acids by more polar amino acids
- substituting polar amino acids by charged amino acids

As such increasing the hydrophilicity by increasing the number of more polar or charged amino acids will change the amino acid composition and as such can be considered as compositional features which can be adapted in order to increase secretion.

Non-polar amino acids are selected from the group A, V, L, I, C, M, F. Amino acids G, P, Y, W can be considered as non-polar in a polar context and as polar in a non-polar context. More polar residues are selected from the group S,T, N, Q, D, E, H, R, K. Charged residues are selected from the group D, E, H, R, K. Acidic or negatively charged residues are selected from E, D. Basic or positively charged residues are selected from H, K, R. Using a comparative scale for polarity: [A, V, L, I, M, F, C] < [G, P, Y, W] < [S, T] < [N, Q, H] < [D, E, K, R].

It is known that highly hydrophobic surface regions tend to lead to undesired aggregation or undesired sticking to biomass resulting in high production stress in the production host, accumulation of protein in the host, and hampered secretion or no secretion at all. It was observed that substitutions which increase the overall hydrophilicity are very effective in secretion improvement in particular when these residues comprise solvent accessible residues (= protein surface residues). More specifically it was observed that when substituting non-polar residues for more polar residues in accessible surface regions, the fraction of polar residues might even exceed the fraction of polar residues set by the upper boundaries of the compositional features analysis. Non-compatibility of the target protein's sequence features with the host requirements may be compensated for by increasing the hydrophilicity of target protein, more specifically by introducing additional charge distributed in such a way that the positive and negative charge are evenly distributed over the surface preventing negative or positive charge hotspots.

Although some prediction tools are available for predicting which amino acids are likely to be on the surface given a certain amino acid sequence, the performance of these tools is quite poor when it is required to predict solvent accessible non-polar or hydrophobic patches. Therefore to modulate the hydrophilicity of the protein accessible surface a 3D structure or a 3D structural model is required. The 3D structure of protein can be determined by X-ray crystallography and by NMR. In addition comparative

modelling or template based modelling can be applied to construct reliable 3D models for a given sequences based on 3D structures of homologous proteins (http://en.wikipedia.org/wiki/Homology_modeling). Various servers and software packages for comparative modelling be found at: http://en.wikipedia.org/wiki/Protein_structure_prediction_software

For a recent review on protein structure prediction and modelling see Yang Zhang, Current Opinion in Structural Biology 2008, 18:342-348.

Given the atomic coordinates of a 3D structure or 3D model the accessible surface can be calculated by methods known in the art. A well known method is the calculation via a rolling-ball algorithm developed by Frederic Richards (1977, "Areas, volumes, packing and protein structure." Annu Rev Biophys Bioeng, 6:151–176). See also http://en.wikipedia.org/wiki/Accessible_surface_area

For determination of the accessible surface the quaternary structure of the final mature protein should be considered in order to avoid that substitutions will disturb the interaction between the individual polypeptides (the monomers) in the multimer (e.g. dimer, trimer, tetramer etc)

Surface modulation comprises:

- Spotting area's where non-polar residues are accessible from the solvent giving rise to potential sticky patches, which could hamper proper secretion and recovery.
- Exclude those area's that play a functional role e.g. the active site in general and binding pockets for substrate, co-substrates and co-factors more in particular.
- Substitute non-polar for more polar residues which include also charged residues
- Substitution polar residues for more polar residues or charged residues.
- Redistribution of charged residues in order to avoid region with high negative charge or regions with high positive charge
- Instead of replacing hydrophobic surfaces patches, such regions may also be shielded by introducing glycosylation closely to the non-polar regions

In case of the primary structure, increased hydrophilicity is represented by comparing number of polar residues before and after modification e.g.

	wt	variant
polar	84	92
charged	40	44
basic	19	22
acid	21	22
non-polar	118	110

When considering the accessible surface the contribution of various polar amino acids can expressed as the fraction of the accessible surface formed by a particular amino acid or a particular group of amino acids with respect to the total accessible surface. For example, the total accessible surface of the charged residues can be calculated and compared to the total accessible surface area. By taking all the polar residues the polar accessible surface can be calculated. The hydrophilicity of the proteins surface is said to increase when the fraction of polar surface increases at the cost of non-polar surface.

In principle one can also introduce glycosylation and estimate the area which is shielded by the glycosylation. The distribution of charges may be done by any available method, including visual inspection.

In one embodiment, the features to be modified for improved secretion are surface charge (re)distribution, surface polar-non-polar distribution, sequence motifs, such as glycosylation, or a combination of these. The skilled person will understand that modification of one feature, for example an amino acid, will in many instances effect a modification of another feature, for example atomic composition with respect to C, H, N, O, S atoms.

It is to be understood that the methods according to the present invention can conveniently be combined with a state of the art technique to increase levels of protein production or with combinations of one or more of these techniques. These include but are not limited to application of strong promoters, increase of copy number, optimal Kozak sequence, mRNA stabilizing elements and optimized codon usage (WO2008/000632).

Examples

Strains

***A. niger* strains**: WT 1: This *A. niger* strain is used as a wild-type strain. This strain is deposited at the CBS Institute under the deposit number CBS 513.88.

WT 2: This *A. niger* strain is a WT 1 strain comprising a deletion of the gene encoding glucoamylase (*gla*A). WT 2 was constructed by using the "MARKER-GENE FREE" approach as described in EP 0 635 574 B1. In this patent it is extensively described how to delete *gla*A specific DNA sequences in the genome of CBS 513.88. The procedure resulted in a MARKER-GENE FREE Δ*gla*A recombinant *A. niger* CBS 513.88 strain, possessing finally no foreign DNA sequences at all.

WT 3: To disrupt the *pep*A gene encoding the major extracellular aspartic protease PepA in WT 2, *pep*A specific DNA sequences in the genome of WT 2 were deleted, as described by van den Hombergh et al. (van den Hombergh JP, Sollewijn Gelpke MD, van de Vondervoort PJ, Buxton FP, Visser J. (1997) - Disruption of three acid proteases in Aspergillus niger--effects on protease spectrum, intracellular proteolysis, and degradation of target proteins - Eur J Biochem. 247(2): 605-13). The procedure resulted in a MARKER-GENE FREE WT 3 strain, with the *pep*A gene inactivated in the WT 2 strain background.

WT 4: To delete the *hdf*A gene in WT 3, the method as earlier described in detail in WO05/095624 was used to generate *Aspergillus niger* WT 4 (Δ*gla*A, Δ*pep*A, Δ*hdf*A).

WT 5: This *A. niger* strain is a WT 4 strain comprising a deletion which results in an oxalate deficient *A. niger* strain. WT 5 was constructed by using the method as described in EP1157100 and US6,936,438, in which an oxalate deficient strain was obtained by deletion of the *oah*A gene, encoding oxaloacetate hydrolase, Strain WT 5 was selected as a representative strain with the *oah*A gene inactivated in the WT 4 strain background.

WT 6: This *A. niger* strain is a WT 5 strain comprising the deletion of three genes encoding alpha-amylases (*amy*B, *amy*BI and *amy*BII) in three subsequent steps. The construction of deletion vectors and genomic deletion of these three genes has been described in detail in WO2005095624. The vectors pDEL-AMYA, pDEL-AMYBI and pDEL-AMYBII, described in WO2005095624, have been used according the "MARKER-GENE FREE" approach as described in EP 0 635 574 B1. The procedure described above resulted in WT 6, an oxalate deficient, MARKER-GENE FREE Δ*gla*A, Δ*pep*A,

Δ*hdf*A, Δ*amy*A, Δ*amy*BI and Δ*amy*BII amylase-negative recombinant *A. niger* CBS 513.88 strain, possessing finally no foreign DNA sequences at all. As such, strain WT 6 has a low amylase background, has a higher HR/NHR ratio for more efficient targeting of sequences and is more optimized for extracellular protein expression and detection compared to WT 1.

***K.lactis* strains:** To assess the expression of PGE and its variants in *K.lactis* two strains were used. GG799 (New England Biolabs) and a derivative of *K.lactis* CBS 685.97, also called WT 7 herein, that is in more detail describe in the patent US 6,265,186 B1. Strain *K.lactis* WT 7 was derived from CBS 685.97 by means of mutagenesis (classical strain improvement) and genetic engineering.

Chitinase activity assay

The reaction mix contained: 3 mg of chitin-azure (Sigma), 0.5 ml of 0.1 M Na-citrate-phosphate buffer, pH 5.0 and 0.1 ml of sample to be analyzed (culture liquid). The reaction mix was incubated for 24 hours at 37°C with shaking, centrifuged for 10 min at 12000 rpm and the OD590 was measured.

Beta-glucosidase activity using pNP-ß-glucopyranoside as a substrate.

A 3mM pNP-ß-glucopyranoside (Sigma N7006) stock solution was prepared in 50mM sodium acetate buffer pH=4.5. Assay: 250 µl substrate-stock (3mM) + 250 µL diluted enzyme sample was incubated at 40 °C. Reactions were stopped at t = 0, 10, 20 and 30 minutes by mixing 100µl incubate with 100 µl 1M sodiumcarbonate. The extinction was determined at 405 nm using a MTP reader. Activity is expressed in µmol pNP released/ml/min

Beta-glucosidase activity using cellobiose as substrate.

A cellobiose (Sigma C7252) stock solution of 10 mM final concentration was prepared in 50 mM sodium acetate buffer pH=4.5. For the assay 2000 µl substate-stock (10 mM) + 100 µL diluted enzyme sample were mixed and incubated at 40 °C. Reactions were stopped at t = 0, 10, 20 and 30 minutes by mixing 100µl incubate + 100 µl 50 mM sodiumhydroxide. Samples were subjected to ultrafiltration and analyzed using High Performance Anion Exchange Chromatography with Pulsed Amperometric Detection

(HPAEC-PAD), performed on a Dionex DX-500 equipped with an ED 40 pulsed amperometric detector. Activity is expressed in µmol glucose released/ml/min

Endo-glucanase activity using AZO-CM-Cellulose.

The assay is carried out according the Megazyme procedure S-ACMC 04/07 (Megazyme International Ireland Ltd, http://secure.megazyme.com/downloads/en/data/ S-ACMC.pdf). Activity was measured on 2% AZO-CM-Cellulose in 100mM sodium acetate buffer pH 4.6 at 40 °C. For the assay 250 µL substrate stock (2%) + 250 µL diluted enzyme solution were mixed. After 30 minutes 1250 µL of precipitant solution was added. Reactions were stopped by adding precipitant solution: 300 mM sodium acetate buffer pH=5 with 20mM Zn-acetate in ethanol 76 %. De extinction at 590 nm was measured of the supernatant after centrifugation at 1000xg for 10 minutes, using a spectrophotometer. Activity is expressed in µmol dye released/ml/min,

Tributyrine plate assay

The Rhodamine B lipase plate screening assay was done with tributyrin (C4) as a substrate. The Rhodamine B plate assay is commonly used for the screening of lipase activity presence in the samples and was adapted from assay described in literature (G. Kouker, K.E. Jaeger, Appl. and Environ. Microbiol, 1987, 211-213). All chemicals used were analytical grade. An arabic gum emulsion was made by dissolving 17.9 g NaCl and 0.41 g KH_2PO_4 in 400 ml of H_2O and finally 540 ml of glycerol (87%) was added. Six (6.0) g of Arabic gum was slowly added and after dissolving the total volume of 1000 ml was achieved by adding of H_2O.

Rhodamine B solution was prepared by dissolving Rhodamine B at concentration of 20 mg/ml in ethanol. A 4% agarose solution was prepared by dissolving 4 g agarose in 100 ml buffer solution (0.1M Acetate pH=5.5) by heating. The substrate used to screen for lipase activity was tributyrin.

Plate assay procedure: 1 ml of substrate and 1.5 ml Arabic gum emulsion was mixed with 5 ml buffer solution and sonificated using a Soniprep with an amplitude of 20 micron for 2x60 sec or optionally an Ultraturax, set at green for 2 minutes. To this solution 7.5 ml of hot agarose solution was added together with 150 µl of Rhodamine B. The final solution was poured in a Petri dish plate. Plates were stored in the refrigerator until use. Just before use holes of 3 mm diameter were made using a replicator. 10 µl of solution to be checked for lipase activity was pipetted into a hole, after which the plate was

incubated at 37^0 C for 18-24 hours. The fluorescent halo around the hole is indicative for lipase activity.

pNP-butyrate assay

Pre-Gastric Esterase (PGE) activity was determined at 37°C on a final concentration of 1 mM para-nitrophenyl butyrate as substrate against an internal enzyme standard. A substrate solution was prepared by making a 50 mM para-nitrophenyl butyrate stock solution in acetonitril, which was diluted five times in 0.1 M sodium phosphate buffer pH 6.7 containing 0.2% BSA and 2% Triton X-100. 120 µl of 0.1 M sodium phosphate buffer pH 6.7 containing 0.2% BSA, 15 µl of substrate solution was added. After preheating to 37°C, 15 µl of sample in an appropriate dilution was added (dilution in 0.1 M sodium phosphate buffer pH 6.7 containing 0.2% BSA), after which the absorbance increase over 5 minutes of incubation at 37°C was measured spectrophotometrically at 405 nm. Sample responses were corrected for a blank background (incubation of 15 µl of 0.1 M sodium phosphate buffer pH 6.7 containing 0.2% BSA instead of sample) and typically ranged from 0.05 to 0.5 dAbs after blank correction.

The internal standard was calibrated in a titrimetric assay on tributyrin, performed at pH 6.0 and 30°C. Five ml of a PGE sample solution (prepared in milliQ water) were added to 30 mL of a pre-heated tributyrin/Arabic gum emulsion (93 and 57 g/L in water, respectively). Free fatty acid release was measured over 5 minutes by titration with 0.02 N NaOH.

SDS- PAGE electrophoresis

Sample pre-treatment: 30µl sample was added to 35µl water and 25 µl NuPAGE™ LDS sample buffer (4x) Invitrogen and 10 µl NuPAGE™ Sample Reducing agent (10x) Invitrogen. Samples were heated for ten minutes at 70°C in a thermo mixer.

SDS-PAGE was performed in duplicate according to the supplier's instructions (Invitrogen: Gel: 4-12% Bis-Tris gel, Buffer: MES SDS running buffer, Runtime: 35 minutes). One of the two gels was used for blotting, 10 µl of the sample solutions and 1 µl marker M12 (Invitrogen) were applied on the gels (NuPAGE™ BisTris, Invitrogen).

The gels were run at 200V, using the XCELL Surelock, with 600 ml 20 times diluted MES-SDS buffer in the outer buffer chamber and 200ml 20 times diluted MES-SDS buffer, containing 0.5 ml of antioxidant (NuPAGE™ Invitrogen) in the inner buffer chamber. After running, the gels were fixed for one hour with 50% Methanol/ 7% Acetic

acid (50ml), rinsed twice with demineralised water and stained with Sypro Ruby (50ml, Invitrogen) overnight.
Images were made using the Typhoon 9200 (610 BP 30, Green (532 nm), PMT 600V, 100 micron) after washing the gel for ten minutes with demineralised water.

Western blotting

PGE polyclonal antibody

PGE polyclonal antibodies were ordered at Eurogentec (Belgium) using the speedy 28-days program and two synthesized PGE peptides as antigens. The PGE antibody was validated against the commercial Piccantase C (DFS) enzyme preparation (data not shown).
Western blotting was performed according to method of analysis S2300.

membrane : NC 0.45 µm
Runtime : 90 minutes at 25V
Buffer : transfer buffer with methanol

After the transfer to the membrane the following steps were performed:
Block the membrane in 20 ml skim milk (1% skim milk in PBST; 10mM PBS + 0.05% TWEEN20) for two hours.
Antibody 1: SY0716, Rabbit; dissolve 40 µl Antibody in 20 ml PBST) overnight at room temperature (1:500).
Rinse membrane with PBS-T and wash next 3 x 20' with PBST buffer.
Antibody 2: ECL Plex Goat Anti-Rabbit IgG Cy3(GE Healthcare); dissolve 10 µl ECL Plex in 25 ml PBST, keep in dark) 1 hour. (1:2500)
Rinse membrane 4 times and wash next 2 x 10' in PBST
Wash 2 x 10' in PBS
An image was made of the membrane using the Typhoon 9200 (670 BP 30, green (532 nm), PMT 450V, 100 micron).

Molecular biology techniques

In the examples herein, using molecular biology techniques known to the skilled person (see: Sambrook & Russell, *Molecular Cloning: A Laboratory Manual, 3rd Ed.*, CSHL Press, Cold Spring Harbor, NY, 2001), several genes were over expressed and others were down regulated as described below.

All gene replacement vectors described and used, were designed according to known principles and constructed according to routine cloning procedures. In essence, these vectors comprise approximately 1 – 2 kb flanking regions of the respective ORF sequences, to target for homologous recombination at the predestined genomic loci. In addition, they contain the *A. nidulans* bi-directional *amd*S selection marker for transformation, in-between direct repeats. The method applied for gene deletion in all examples herein uses linear DNA, which integrates into the genome at the homologous locus of the flanking sequences by a double cross-over, thus substituting the gene to be deleted by the *amd*S gene. After transformation, the direct repeats allow for the removal of the selection marker by a (second) homologous recombination event. The removal of the *amd*S marker can be done by plating on fluoro-acetamide media, resulting in the selection of marker-gene-free strains. Using this strategy of transformation and subsequent counter-selection, which is also described as the "MARKER-GENE FREE" approach in EP 0 635 574, the *amd*S marker can be used indefinitely in strain modification programs. The general procedure for gene disruption is depicted in Figure 6 of WO2006040312. The general design of deletion vectors was previously described in EP635574B and WO 98/46772 and the use of general cloning vector pGBDEL for constructing deletion vectors and the counter-selection procedure were *a.o.* described in WO06/040312.

Examples of the general design of expression vectors and specifically pGBFIN-expression vectors for gene over expression, transformation, use of markers and selective media can be found in WO199846772, WO199932617, WO2001121779, WO2005095624, EP 635574B and WO2005100573.

Shake flask fermentations

A. niger strains were pre-cultured in 20 ml CSL pre-culture medium (100 ml flask, baffle) as described in the Examples: "Aspergillus niger shake flask fermentations" section of WO 99/32617. After growth for 18-24 hours at 34°C and 170 rpm, 10 ml of this culture is transferred to Fermentation Medium (FM). Fermentation in FM is performed in 500 ml flasks with baffle with 100 ml fermentation broth at 34°C and 170 rpm for the number of days indicated, generally as described in WO99/32617.

The CSL medium consisted of (in amount per litre): 100 g Corn Steep Solids (Roquette), 1 g NaH_2PO_4* H_20, 0.5 g $MgSO_4*7H_2O$, 10 g glucose*H_2O and 0.25 g Basildon (antifoam). The ingredients were dissolved in demi-water and the pH was adjusted to pH

5.8 with NaOH or H_2SO_4; 100 ml flasks with baffle and foam ball were filled with 20 ml fermentation medium and sterilized for 20 minutes at 120°C.
The fermentation medium (FM) consisted of (in amount per liter): 150 g maltose*H_2O, 60 g Soytone (peptone), 1 g NaH_2PO_4*H2O, 15 g $MgSO_4$*7H2O, 0.08 g Tween 80, 0.02 g Basildon (antifoam), 20 g MES, 1 g L-arginine. The ingredients were dissolved in demi-water and the pH was adjusted to pH 6.2 with NaOH or H_2SO4; 500 ml flasks with baffle and foam ball were filled with 100 ml fermentation broth and sterilized for 20 minutes at 120°C.

For *K.lactis* shake flask fermentations, a single colony of a *K.lactis* PGE transformant was inoculated into 100 ml (flask) of YEP(4%)-D/MES medium that contained per liter: 10 g yeast extract, 20g Bacto peptone, 40 g glucose and 100 mM MES pH 6.7. The fermentation was performed at 30 °C in a shake incubator at 280 rpm. Supernatant was collected at day 2 and 3 and further analysed as describe below.

Example 1. Construction of *K. lactis* and *A. niger* expression vectors for wild-type enzymes and enzyme variants according a method of the invention

In this example a number of expression vectors were constructed for variants of the enzymes of the invention. All variants for expression in *Kluyveromyces* were cloned in a pKLPGE- vector very similar to the pKLAC2 expression vector (New England Biolabs). The general layout of all pKLPGE- vectors can be found in Figure 1. All variants for expression in *Aspergillus* were cloned in a pGBFIN-5 or a pGBTOP- expression vector. The construction, general layout and use of these vectors are described in detail in WO199932617.

***K. lactis* constructs**

Calf pregastric esterase (PGE) is an industrially interesting enzyme and its full length cDNA sequence was published by Timmermans *et. al.* (1994, Gene 147: 259-262). For expression of PGE in *Kluyveromyces lactis*, this cDNA sequence was codon pair optimized (SEQ ID No. 1) and prepared synthetically (*e.g.* DNA2.0, USA, GeneArt, Sloning, Germany). An expression construct containing a fusion with the *K. lactis* α-factor pre(pro-) signal sequence and a KREAEA Kex pre(pro-)-sequence processing site

was made. Via *Hind*III and *Not*I restriction sites, the synthetic gene was cloned in the *K. lactis* expression vector, yielding pKLPGE-WT (Figure 1), which also contained an *amd*S selection marker. In addition, several PGE variants were designed with improved protein features according a method of the invention. These mutants differed from the codon pair optimised wild type PGE enzyme (SEQ ID No. 2) with respect to the number of glycosylation sites and/or with respect to hydrophobicity. The PGE mutant enzyme encoding genes were also codon pair optimized and prepared synthetically, as described above. The variants were cloned into the *K. lactis* expression vector as essentially described before using *Xho*I and *Not*I cloning sites. All relevant nucleotide and protein details for PGE constructs can be found in Table 4.

Table 4. Overview of wildtype and mutants of PGE enzymes expressed in *K. lactis*

Name mutant construct	**Nucleotide ref SEQ ID No.**	**Protein ref SEQ ID No.**	**Description of construct and modification compared to wildtype native full PGE sequence (SEQ ID NO. 2).** (Within brackets corresponding positions in the mature PGE sequence)
pKLPGE-WT	1	2	Codon pair optimized PGE as fusion with the *K. lactis* α-factor signal sequence and a KREAEA kex site
pKLPGE-8	3	4	1 extra glycosylation site was added by modifying amino acid K98 [79] to N
pKLPGE-9	5	6	5 extra glycosylation sites were added by modifying amino acids: A70 [51] to S K98 [79] to N R158 [139] to N and R159 [140] to K H318 [289] to N and P320 [301] to S I361 [342] to T
pKLPGE-11	7	8	pI shift of 6.96 to 7.74; number of polar residues was increased from 165 to 181 and number of charged amino acids residues from 80 to 91
pKLPGE-12	9	10	pI shift from 6.96 to 6.7; number of polar residues was increased from 165 to 188 and number of charged amino acids residues from 80 to 103
pKLPGE-10	11	2	PGE variant with native signal sequence fused to α-MAT factor signal pre(pro-)sequence

***A. niger* constructs**

For expression of Calf pregastric esterase PGE in *A. niger*, the cDNA sequence was codon pair optimized (SEQ ID No. 12) and prepared synthetically (*e.g.* DNA2.0, USA, GeneArt, Sloning, Germany). The codon pair optimized PGE encoding gene was prepared synthetically as a fusion to a truncated glucoamylase carrier protein (tAG). The fusion fragment was inserted into a pGBTOP- *A. niger* expression vector as shown for pANPGE-3 in Figure 2.

The wild-type *A. niger* gene An08g09030 encoding a putative chitinase (ZDU, EC 3.2.1.14, Uniprot A5AB48) was identified in the *A. niger* genome sequence (EMBL: AM269948 - AM270415; Pel *et al.*, "Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88". Nat Biotechnol. 2007 Feb; 25 (2):221-231). The cDNA sequence of the wild-type chitinase ZDU can be identified as SEQ ID NO. 17 with the deduced wild-type chitinase ZDU protein sequence as SEQ ID NO. 18. The coding sequence of An08g09030 was codon pair optimized (as detailed in WO2008000632) and the translational initiation sequence of the glucoamylase glaA promoter has been modified into 5'-CACCGTCAAA ATG-3' in all expression constructs generated (as also detailed in WO2006/077258). In addition, an optimal translational termination sequence was used, and therefore the wild-type 5'-TGA-3' translational termination sequence was replaced by 5'-TAAA-3' (as detailed in WO2006/077258) in all expression constructs. The optimized chitinase ZDU construct was synthesized completely as *Pac*I – *Asc*I fragment, subcloned and sequence verified. The *Pac*I – *Asc*I restriction sites at the ends of the synthesized fragments were used to allow cloning in the large vector fragment of a *Pac*I – *Asc*I digested pGBFIN-5 expression vector, generating a pGBFINZDU-WT expression vector (Figure 3).

In addition and in a similar way as for the ZDU chitinase, the *Talaromyces emersonii* beta-glucosidase (ZTB, EC 3.2.1.21, Uniprot Q8X214) and *Phanerochaete chrysosporium* endoglucanase (ZTC, EC 3.2.1.4, Uniprot Q66NB6) were codon pair optimized (as detailed in WO2008000632) and with all appropriate control elements cloned as *Pac*I – *Asc*I fragments in pGBFIN-5, generating pGBFINZTB-WT and pGBFINZTC-WT, respectively

Protein feature optimizations (PFO) according a method of the invention were applied to the calf pregastric esterase, *A. niger* chitinase protein sequence, the *T. emersonii* beta-glucosidase and *P. chrysosporium* endoglucanase protein sequences. The coding

sequences comprising designed variants of the the calf pregastric esterase were synthesized completely as *Eco*RI - *Sna*BI fragments and sequence verified. The synthesized fragments were cloned in a pGBTOP-vector, generating pANPGE-expression constructs. All relevant nucleotide and protein details for *A. niger* PGE constructs can be found in Table 5.

Table 5. Wild-type and variant enzyme expression constructs for *A. niger*, references and their properties

Calf pregastric esterase LipF						
Construct	Info	PFO	F-score	SEQ ID DNA	SEQ ID Protein	Details
pANPGE-3	pI=6.96	N	10.7	12	2	CPO gene tAG fusion with Kex site (KR)
pANPGE-12	pI=4.6	Y	7.4	13	14	pI shift from 6.96 to 4.6, number of polar residues was increased from 165 to 186 and number of charged amino acids residues from 80 to 88
pANPGE-13	pI=4.88	Y	7.2	15	16	pI shift from 6.96 to 4.88, number of polar residues was increased from 165 to 180 and number of charged amino acids residues from 80 to 83

The coding sequences comprising designed variants of the chitinase, the beta-glucosidase and endoglucanase were synthesized completely as *Pac*I – *Asc*I fragments, subcloned and sequence verified. The *Pac*I – *Asc*I restriction sites at the ends of the synthesized fragments were used to allow cloning in the large vector fragment of a *Pac*I – *Asc*I digested pGBFIN-5 expression vector, generating variant pGBFIN- expression vectors. The variant expression constructs were named as described below, and characteristics and reference to respective nucleotide and protein sequences of the pGBFINZDU- constructs can be deduced from Table 6, of the pGBFINZTB- constructs from Table 7 and of the pGBFINZTC- constructs from Table 8.

Table 6. Wild-type and variant enzyme expression constructs for *A. niger*, references and their properties

Chitinase *A. niger*

Construct / Strain	PFO	F-score	SEQ ID DNA	SEQ ID Protein	Example	Activity	Assay	SDS-PAGE
WT6						low	Fig 5	absent
ZDU wt	N	9.7	17	18	4	low	Fig 5	faint
ZDU 6	Y	5.7	19	20	4	improved up to 3-fold	Fig 5	strong
ZDU 7	Y	4.0	21	22	4	improved up to 2-fold	Fig 5	strong

Table 7. Wild-type and variant enzyme expression constructs for *A. niger*, references and their properties

Beta-Glucosidase *Talaromyces emersonii*								
Construct / Strain	PFO	F-score	SEQ ID DNA	SEQ ID Protein	Example	pNP activity	Cellobiose activity	SDS-PAGE
WT6						low	low	absent
ZTB wt	N	11.3	23	24	4	low	low	absent
ZTB 4	Y	8.2	25	26	4	improved up to 20-fold	improved up to 30-fold	strong

Table 8. Wild-type and variant enzyme expression constructs for *A. niger*, references and their properties

Construct / Strain	PFO	F-score	SEQ ID DNA	SEQ ID Protein	Example	AZO-Cellulose activity	SDS-PAGE
WT6						very low	absent
ZTC wt	N	11.3	27	28	4	very low	absent
ZTC 5	Y	5.2	29	30	4	highly improved	strong

Example 2. Expression and secretion analysis of wild-type and protein feature optimized PGE's in *K. lactis*

Strains *K.lactis* GG799 or *K.lactis* WT 7 were transformed with all *K.lactis* pKLPGE-constructs (Table 4) that also contained the *amd*S selection marker. For each of the

transformations, 20 colonies were purified on selective medium containing acetamide. Part of the colony was used to generate a DNA template for a PCR reaction to determine the copy number of the PGE construct in each strain. Per construct, 3 transformants, positive in the PCR screen, were further screened on a plate assay containing tributyrine as an enzymatic substrate. For the wt PGE enzyme, no clear activity halo could be detected using the tributyrine plate assay. Also analysis of the supernatant on SDS-PAGE for PGE production did not show a positive result. Surprisingly, for 4 out of the 5 PGE mutants with optimized protein features a clear activity halo could be observed using the tributyrine plate assay. A number of transformants for wt and mutant PGE's were grown in shake flasks and broth and supernatant were examined for lipase activity using pNP-butyrate as a substrate. A summary of various activity assays for the PGE mutants is shown in Table 9.

Table 9. Activity tests of PGE wt and PFO variants

Sample	Day 2				Day 3			
	pNP assay				pNP assay			
K.lactis	(U/ml)		Plate assay		(U/ml)		Plate assay	
Transformant	Broth	Super-natant	Broth	Super-natant	Broth	Super-natant	Broth	Super-natant
pKLPGE-WT #1	< 0.12	< 0.1	-	-	< 0.12	< 0.1	-	-
pKLPGE-WT #2	< 0.12	< 0.1	-	-	< 0.12	< 0.1	-	-
pKLPGE-WT #3	< 0.12	< 0.1	-	-	< 0.12	< 0.1	-	-
pKLPGE-8 #1	< 0.2	< 0.1	+/-	-	0.24	< 0.1	++	-
pKLPGE-8 #2	< 0.2	< 0.1	+	+/-	0.22	< 0.1	++	+/-
pKLPGE-8 #3	< 0.2	< 0.1	+/-	-	0.31	< 0.1	++	-
pKLPGE-9 #1	0.37	< 0.1	++	++	0.71	0.15	+++	+++
pKLPGE-9 #2	< 0.2	< 0.1	++	+	0.23	< 0.12	++	++
pKLPGE-9 #3	0.44	< 0.1	++	+/-	0.98	0.15	+++	+++
pKLPGE-11 #1	< 0.2	< 0.1	+	-	0.28	< 0.1	+	+/-
pKLPGE-11 #2	0.27	< 0.1	+	-	0.57	< 0.1	+	+/-
pKLPGE-11 #3	0.32	< 0.1	+	-	0.84	< 0.1	+	+/-
pKLPGE-12 #1	1.4	0.28	++	+	1.9	0.41	++	+
pKLPGE-12 #2	4.0	0.67	++	+	6.6	1.2	+	+
pKLPGE-12 #3	8.0	1.6	++	+/-	13	2.8	++	+
pKLPGE-10 #1	< 0.2	< 0.1	-	-	< 0.12	< 0.1	-	-
pKLPGE-10 #2	< 0.2	< 0.1	-	-	< 0.12	< 0.1	-	-
pKLPGE-10 #3	< 0.2	< 0.1	-	-	< 0.12	< 0.1	-	-
GG799 / WT 7	< 0.12	< 0.1	-	-	< 0.12	< 0.1	-	-

For *K.lactis* pKLPGE-WT (PGE CPO) transformants (various copy number) maximum activity of 0.2 U/ml was obtained. By protein feature optimization of PGE, *i.e* as expressed in pKLPGE-12, an increase in activity of more than 50x was observed for this PGE mutant. A number of mutants of the PGE-9, PGE-11 and PGE-12 variants were fermented on a larger scale basis, confirming the improved secretion (data not shown).

In this example it wais shown that by modification of the number of glycosylation sites and by changing the polarity of the hydrophobic enzyme parts exposed to the surface (determined based on PGE modeling) we could dramatically improve the PGE enzyme expression and secretion in *K. lactis*. Furthermore a significant amount of the activity was also found in the supernatant.

Example 3. Expression and secretion analysis of wild-type and protein feature optimized PGE's in *A. niger*

A.niger WT 6 was co-transformed with a pGBAAS construct carrying the *A. nidulans amd*S selection marker and the variant pANPGE- plasmids (Table 5). For each of the transformations, 20 colonies were purified on selective medium containing acetamide and subsequently spore plates were prepared, all as described in WO99/32617. To select *A.niger* transformants that were true co-transformants, e.g. that they contained both PGE and *amd*S cassettes, a PCR check (not shown). The result showed that at least 50% among the 20 selected transformants contained one or more copies of the PGE expressing construct. These PGE containing transformants were continued with. The spores of the PGE contransformants were harvested and shake flask fermentations were performed in FM medium. At day 2 supernatant samples were collected and screened for lipase activity using the tributyrine plate assay.

In samples harvested from the *A.niger* pANPGE-3 transformants very small activity halos could be detected (data not shown). For pANPGE-12 and pANPGE-13 transformants large activity halos could be detected (data not shown). For each construct pANPGE-3, pANPGE-12 and pANPGE-13, transformants (1-3) that showed the largest halo on the tributyrine plate assay were examined for lipase activity using pNP-butyrate as a substrate. A summary of various activity assays for the PGE mutants is shown in Table 10.

Table 10. Wild-type and PFO PGE variants expressed in *A. niger*

Calf pregastric esterase LipF							
Construct	PFO	F-score	SEQ ID DNA	SEQ ID Protein	Activity pNP day 2 & 3 supernatant		Activity Tributyrate plate assay day 2 supernatant
pANPGE-3	N	10.7	12	2	0.6	0.1	+/-
pANPGE-12	Y	7.4	13	14	4.7	5.8	++
pANPGE-13	Y	7.2	15	16	5.6	7.5	++++

++++, +++, ++, +, +/-, - corresponded to very large, large, medium, small, not clear and no halo on the tributyrine plate assay, respectively.

The supernatant samples of WT6 and selected transformants pANPGE-12#16 and pANPGE-13#30 were further analysed on SDS-PAGE gel (Invitrogen) and by western blotting using PGE polyclonal antibodies (see Figure 4). For the *A.niger* PGE PFO variant of pANPGE-12, a band corresponding to the mature PGE could be detected on the SDS-PAGE gel. Using the PGE polyclonal antibody PGE, cross-hybridizing bands could be detected in supernatants of both transformants. The highest molecular weight band (about 55kDa) corresponds probably to the mature PGE mutant and the cross-hybridizing bands of the lower molecular weight could be a result of a proteolytic degradation.

It is concluded that by changing the polarity of the enzyme parts exposed to the surface (determined based on PGE modelling) following the rules of protein feature optimisation we could dramatically improve the PGE enzyme expression in *A.niger*. Furthermore high enzymatic activity was also found in the supernatant.

Example 4. Expression of wild-type and PFO optimized fungal enzymes in *A. niger*

The pGBFINZDU-, pGBFINZTB- and pGBFINZTC- expression constructs, prepared in Example 1 (*super*), were introduced by transformation using *A. niger* as described below. In order to introduce the different pGBFINZDU-, pGBFINZTB- and pGBFINZTC-vectors (Table 6, 7 and 8, respectively) in WT 6, a transformation and subsequent selection of transformants was carried out as described in WO1998/46772 and WO1999/32617. In brief, linear DNA of all the pGBFIN- constructs was isolated and used to transform *A. niger* WT 6. Transformants were selected on acetamide media and colony purified according standard procedures. Colonies were diagnosed for integration at the *gla*A locus and for copy number using PCR. Three independent transformants for each pGBFINZDU-, pGBFINZTB- and pGBFINZTC- construct with similar estimated copy numbers (putative single copy) were selected and named using the number of the transforming plasmid, as for example ZDU-WT-1, ZDU-WT-2, ZDU-WT-3, ZDU-6-1, ZDU-6-2, ZDU-6-3, etc…., respectively.

The selected ZDU-, ZTB- and ZTC- strains and *A. niger* WT6 were used to perform shake flask experiments in 100 ml of the FM medium as described above at 34°C and 170 rpm in an incubator shaker using a 500 ml baffled shake flask. After day 3, day 4

and day 5 of fermentation, samples were taken to determine the amount of extracellular protein produced by gel electrophoresis and the chitinase activity.

The production of chitinase expressed by each of the transformants of the *A. niger* ZDU- transformants containing the different constructs, was measured in the culture supernatant. The measured chitinase activity levels at day 3 are indicated in Figure 5. In addition, the culture supernatants sampled at day 4 were analyzed by SDS gel electrophoresis and staining (Figure 6). From these results, it is clear that an optimized protein features have a positive impact on protein secretion and results in detectable and thus increased protein expression levels and increased activity levels for the chitinase enzyme. Results have been summarized in Table 6.

The production of beta-glucosidase expressed by each of the transformants of the *A. niger* ZTB- transformants containing the different constructs, was measured in the culture supernatant. The culture supernatants sampled at day 4 were analyzed by SDS gel electrophoresis and staining (Figure 7). From these results, it is clear that an optimized protein features have a positive impact on protein secretion and results in detectable and thus increased protein expression levels for the beta-glucosidase enzyme. In addition the activity in the supernatant sampled at day 3 was determined at pH=4.5 and 40ºC using pNP-ß-glucopyranoside as a substrate. The supernatant of the beta-glucosidase which had been subjected to protein feature optimization showed an activity increase of up to 20-fold compared to the parent beta-glucosidase encoded by a codon optimised gene . The background beta-glucosidase activity which is measured for the empty host was two- to four-fold lower than from the parent beta-glucosidase encoded by a codon optimised gene. The activity was also measured using cellobiose as a substrate at pH=4.5 and 40ºC. The measured increase in activity was at least 30-fold compared to the parent beta-glucosidase encoded by a codon optimized gene (empty host strains show three- to ten-fold lower than from the parent beta-glucosidase encoded by a codon optimized gene). Results have been summarized in Table 7.

The production of endo-glucanase expressed by each of the transformants of the *A. niger* ZTC- transformants containing the different constructs, was measured in the culture supernatant. The culture supernatants sampled at day 4 were analyzed by SDS gel electrophoresis and staining (Figure 8). From these results, it is clear that optimized

protein features have a positive impact on protein secretion and results in detectable and thus increased protein expression levels for the endoglucanase enzyme. The endo-glucanase activity in the supernatant sampled at day 3 was determined at pH=4.5 and 40ºC using AZO-CM-cellulose as a substrate. The supernatant of the endo-glucanase which had been subjected to protein feature optimization showed an increase in activity of over 350-fold compared the codon optimized gene expressed in the same host.. It should be noted that due to the very low background activity in the empty strain (undetectable by SDS-PAGE), the increase in activity was expressed in such high figure. For the endo-glucanase encoded by a codon optimized gene the measured activity was about the background activity observed for the empty host strain. . Results have been summarized in Table 8.

Clearly, these examples show how a method of the invention for protein feature optimization can be used for improved secretion and production of proteins and enzymes of interest. Additionally, these results indicate that a method of the invention can be broadly applied to improve protein expression in a host, although the expression construct and host has already several other optimizations, such as for example a strong promoter, an improved translation initiation sequence, an improved translational termination sequence, an optimized codon and codon pair usage and / or an improved host for protein expression.

Applicant's or agent's file reference number 27179-WO-PCT	International application No.

INDICATIONS RELATING TO A DEPOSITED MICROORGANISM

(PCT Rule 13*bis*)

A. The indications made below relate to the microorganism referred to in the description first mentioned on page 7 line 9.

B. **IDENTIFICATION OF DEPOSIT** Further deposits are identified on an additional sheet [X]

Name of depositary institution

CENTRAAL BUREAU VOOR SCHIMMELCULTURES

Address of depositary institution *(including postal code and country)*
Uppsalalaan 8
P.O. Box 85167
NL-3508 AD Utrecht
The Netherlands

Date of deposit 10 August 1988	Accession Number CBS 513.88

C. **ADDITIONAL INDICATIONS** *(leave blank if not applicable)* This information is continued on an additional sheet [ ]

We inform you that the availability of the microorganism identified above, referred to Rule 13bis PCT, shall be effected only by issue of a sample to an expert nominated by the requester until the publication of the mention of grant of the national patent or, where applicable, for twenty years from the date of filing if the application has been refused, withdrawn or deemed to be withdrawn.

D. **DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE** *(if the indications are not for all designated States)*

E. **SEPARATE FURNISHING OF INDICATIONS** *(leave blank if not applicable)*

The indications listed below will be submitted to the International Bureau later *(specify the general nature of the indications e.g., "Accession Number of Deposit")*

For receiving Office use only

[ ] This sheet was received with the international application

Authorized officer

For International Bureau use only

[ ] This sheet was received by the International Bureau on:

Authorized officer

Form PCT/RO/134 (July 1992)

Applicant's or agent's file reference number 27179-WO-PCT	International application No.

INDICATIONS RELATING TO A DEPOSITED MICROORGANISM

(PCT Rule 13*bis*)

A. The indications made below relate to the microorganism referred to in the description first mentioned on page 31 line 8.

B. **IDENTIFICATION OF DEPOSIT** Further deposits are identified on an additional sheet [X]

Name of depositary institution

CENTRAAL BUREAU VOOR SCHIMMELCULTURES

Address of depositary institution *(including postal code and country)*
Uppsalalaan 8
P.O. Box 85167
NL-3508 AD Utrecht
The Netherlands

Date of deposit 11 April 1997	Accession Number CBS 685.97

C. **ADDITIONAL INDICATIONS** *(leave blank if not applicable)* This information is continued on an additional sheet [ ]

We inform you that the availability of the microorganism identified above, referred to Rule 13bis PCT, shall be effected only by issue of a sample to an expert nominated by the requester until the publication of the mention of grant of the national patent or, where applicable, for twenty years from the date of filing if the application has been refused, withdrawn or deemed to be withdrawn.

D. **DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE** *(if the indications are not for all designated States)*

E. **SEPARATE FURNISHING OF INDICATIONS** *(leave blank if not applicable)*

The indications listed below will be submitted to the International Bureau later *(specify the general nature of the indications e.g., "Accession Number of Deposit")*

For receiving Office use only	For International Bureau use only
[ ] This sheet was received with the international application	[ ] This sheet was received by the International Bureau on:
Authorized officer	Authorized officer

Form PCT/RO/134 (July 1992)

CLAIMS

1. Method for improving the secretion of a polypeptide of interest by a eukaryotic host cell, which method comprises modifying the value of a set of relevant protein features in the amino acid backbone of the polypeptide to fall within an optimal range or to become more close to an optimal value for one or more protein features in the eukaryotic host.

2. The method according to claim 1, which method comprises:
(i) determining an optimal range and an optimal value for one or more protein features in the eukaryotic host, and
(ii) determining a set of relevant protein features in the eukaryotic host, which features will improve the secretion of the polypeptide by the eukaryotic host if one or more of these relevant features is modified in the amino acid backbone of the polypeptide, and
(iii) modifying the value of the relevant protein features to fall within the optimal range or more close to the optimal value as determined in (i), wherein (i) and (ii) may be performed in any order.

3. The method according to claim 1 or 2 wherein a relevant set of features is determined by:
a. collecting or creating a dataset **S**, which contains the secretion levels of a suitable amount of proteins in a certain eukaryotic host and the amino acid and DNA sequences of these proteins
b. computing protein features (**F**) for all proteins in the dataset **S**;
c. using a statistical classification method to select a subset of protein features (**Fs**) that gives the best performance of a statistical classifier to distinguish between secreted proteins **S+** and non-secreted proteins **S-** in the dataset S, according to a suitably defined classifier performance criterion.

4. The method according to claim 3, wherein the protein features are calculated from a set of mature proteins.

5. The method according to claims 1-4 wherein an optimal range or an optimal value of protein features for a eukaryotic host are determined by:

a. collecting or creating a dataset **S**, which contains the secretion levels of a suitable amount of proteins in a certain eukaryotic host and the amino acid and DNA sequences of these proteins;
b. computing protein features (**F**) for all proteins in the dataset **S**;
c. determining an optimal value (**F_opt**) for each feature for the eukaryote host by fitting a probability distribution for each protein feature computed from **S+** such that the distribution of the feature values is well described by the chosen probability distribution.
d. determining an optimal range of each feature for the eukaryote host.

6. A method for improving the secretion of the polypeptide by a eukaryotic host, said method comprising:

i) computing protein features for the polypeptide,
ii) determining if one or more protein features of the polypeptide are outside the optimal range or substantially deviate from the optimal value for the eukaryotic host,
iii) rationally changing the amino acid sequence of the polypeptide, such that the value of one or more Fs_AA of the polypeptide falls within the optimal range or is shifted towards the optimal value by a suitable amount, defined by RI or RI_N wherein the change defined by RI or RI_N is preferably more than 10%, 15%, 20%, and most preferably more than 30%.

7. The method according to claims 1-6, wherein the backbone of the polypeptide is modified with respect to one or more of the following features: the number of amino acids, molecular weight, isoelectric point, net charge at a specific pH, GRAVY score, aliphatic index, instability index, compositional features, atomic composition of C, H, N, O, S atoms, amino acid frequency, dipeptide frequency, tripeptide frequency, acidic amino acid frequency, aliphatic amino acid frequency, aromatic amino acid frequency, basic amino acid frequency, local features, localization features, glycosylation pattern and charged amino acid frequency.

8. The method according to claims 1-7, wherein the backbone of the polypeptide is modified with respect to one or more of the following features: basic amino acid frequency, polar amino acid frequency, non-polar amino acid frequency, tiny amino acid frequency, small amino acid frequency, charged amino acid frequency, net charge at pH 7.2, isoelectric point, frequency of Asn, Arg, Ile, Cys, His, Gln, Val, Lys, Gly, Thr and Leu, respectively, localization features, turn as calculated by Garnier, PEST motif as calculated by EPESTFIND, LF values for pI, LF values for Gravy score, LF values for aroma score, sulphur (S) composition.

9. The method according to claim 1-7, wherein the backbone of the polypeptide is modified with respect to one or more features selected from the group consisting of: pI, net charge, net charge per length, net positive charge per length, net negative charge per length, total charge per length, gravy score, aroma score, aliphatic index, tiny amino acid frequency, small amino acid frequency, polar amino acid frequency, non-polar amino acid frequency, charged amino acid frequency, acidic amino acid frequency, basic amino acid frequency, aliphatic amino acid frequency, frequency of Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr and Val, respectively.

10. The method according to claim 1-9, wherein the backbone of the polypeptide is modified with respect to one or more features selected from the group consisting of: pI, net charge (pH7.2), net charge (pH7.2) per length, net positive charge (pH7.2) per length, total charge (pH7.2) per length, aliphatic index, small amino acid frequency, polar amino acid frequency, non-polar amino acid frequency, charged amino acid frequency, amino acid frequency, frequency of Arg, Gln, Glu, Lys, Phe and Thr, respectively.

11. The method according to claim 1-9, wherein the backbone of the polypeptide is modified with respect to one or more features selected from the group consisting of: glycosylation sites, gravy score, polar amino acid frequency, non-polar amino acid frequency, charged amino acid frequency, acidic amino acid frequency, basic amino acid frequency, frequency of Glu, Lys and Thr, respectively.

12. The method according to claim 1-11, wherein at least 5% of the amino acids of the amino acid backbone is modified, more preferably at least 10%, even more preferably at least 15%, even more preferably at least 20% of the amino acids of the amino acid backbone is modified.

13. The method according to claim 1-11, wherein at least 5 amino acids of the amino acid backbone is modified, more preferably at least 10 amino acids, even more preferably at least 15 amino acids, even more preferably at least 20 amino acids, even more preferably at least 25 amino acids, even more preferably at least 30 amino acids of the amino acid backbone is modified.

14. The method according to claim 1-13, wherein an improvement in F-score of at least 5% with respect to the wild type reference protein is achieved, more preferably at least 10%, even more preferably at least 15%, even more preferably at least 20% and even more preferably at least 30% improvement is achieved, wherein the F-score is calculated according to the formula:

$$F = \left(\sum_{i=1}^{n} \left| D_{N,i} \right|^{p} + \eta \sum_{j=1}^{m} \left| D_{N,j} \right|^{p} \right)^{1/p}$$,wherein η is a weighing factor between and including 0 and 1 ($0 \leq \eta \leq 1$) and preferably $\eta \leq 0.5$, more preferably $\eta \leq 0.4$, most preferably $\eta = 0.3$, and wherein p is between and including 1 and 5 ($1 \leq p \leq 5$) and preferably $p = 2$.

15. The method according to claim 1-14, wherein at least 2, 3, 4, or 5 features are modified, more preferably at least 10, even more preferably at least 15, even more preferably at least 20, even more preferably at least 25, and even more preferably at least 30 features are modified.

16. The method according to claim 1-15, wherein at least 2, 3, 4, or 5 features are improved, more preferably at least 10, even more preferably at least 15, even more preferably at least 20, even more preferably at least 25, and even more preferably at least 30 features are improved, whereas preferably less than 10, even more preferably less than 5, even more preferably less than 4 features are worsened.

17. The method according to claims 1-16, wherein the features are primary features.

18. The method according to claims 1 or 17, wherein the backbone of the polypeptide is modified with respect to one or more other protein features.

19. The method according to claims 1-18, wherein the backbone of the mature polypeptide is modified.

20. The method according to claims 1-19, wherein the eukaryotic cell is a yeast cell or a filamentous fungal cell.

21. The method according to claims 1-20, wherein the polypeptide is a mammalian or a bacterial polypeptide.

22. The method according to claims 1-21, wherein the specificity of the modified polypeptide substantially remains the same as before improvement of the secretion.

23. The method according to claims 1-22, wherein the specific activity of the modified polypeptide substantially remains the same as before improvement of the secretion.

24. The method according to claims 1-23, wherein the improvement of secretion is measured by increase in the activity and wherein the activity in the extracellular medium is increased by at least 5%.

25. The method according to claims 1-24, wherein the polypeptide is an enzyme, a membrane protein, a hormone or a receptor.

26. Method for the production of a polypeptide of interest comprising, applying the method according to anyone of claims 1-25 to the polypeptide of interest and producing the modified polypeptide by recombinant technology.

27. A polypeptide obtained by the method according to claim 26.

28. A polypeptide obtainable by the method according to claim 26.

29. The modified polypeptide obtained according to claims 1-25.

30. A modified polypeptide obtainable according to claim 1-25.

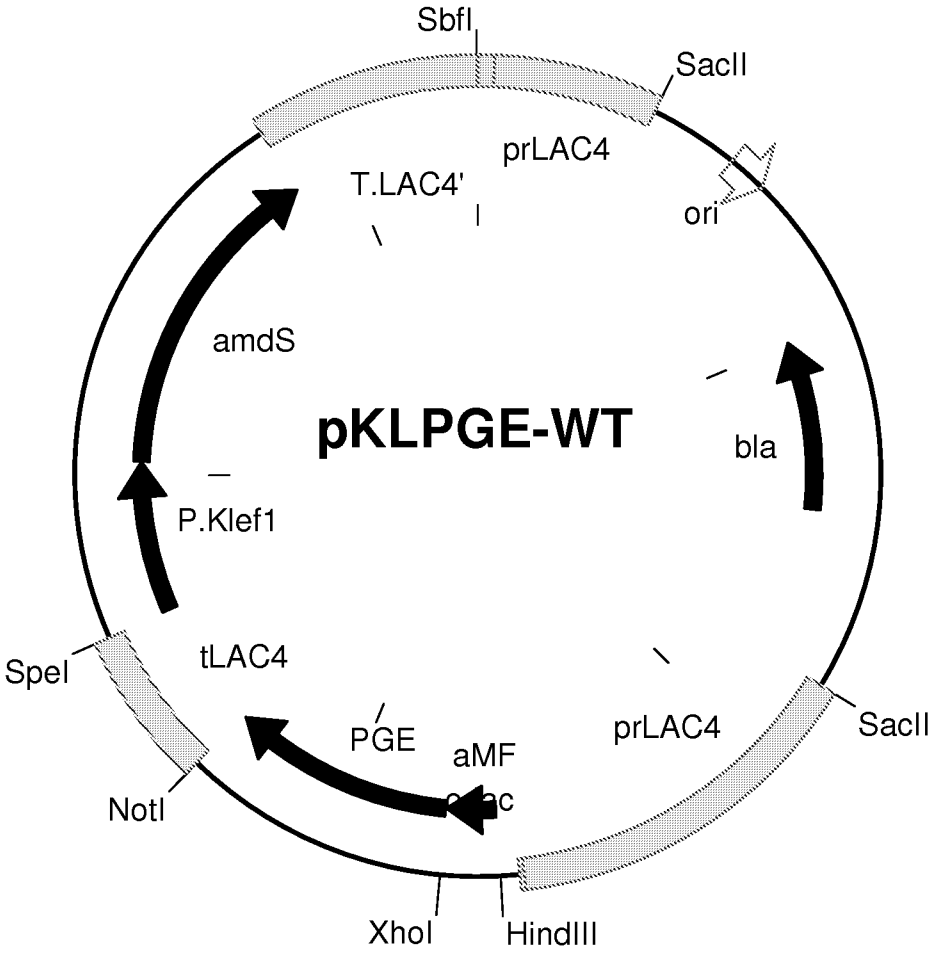
SbfI
SacII
prLAC4
T.LAC4'
ori
amdS
pKLPGE-WT
bla
P.Klef1
SpeI
tLAC4
PGE
aMF
prLAC4
SacII
NotI
XhoI
HindIII

Figure 1

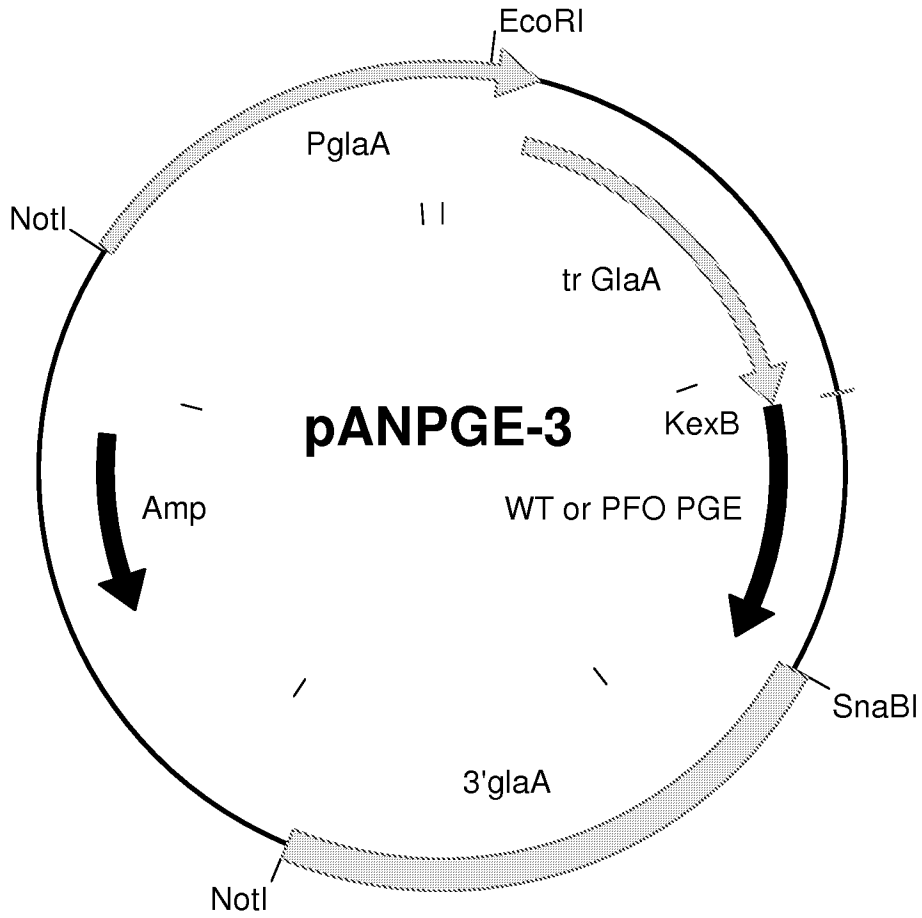
EcoRI
PglaA
NotI
tr GlaA
pANPGE-3
KexB
Amp
WT or PFO PGE
SnaBI
3'glaA
NotI

Figure 2

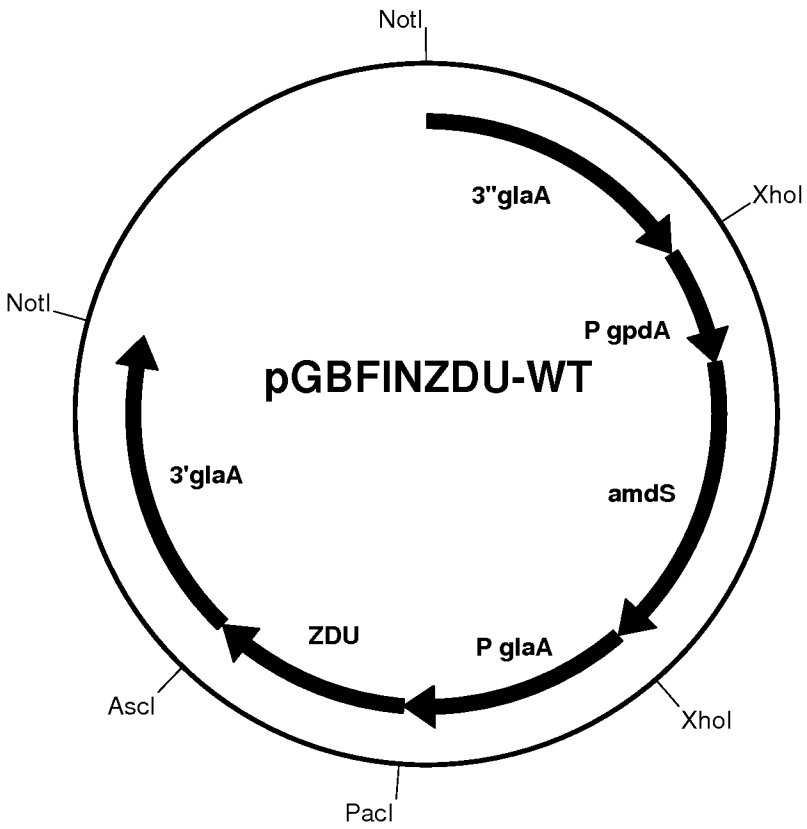
NotI
3''glaA
XhoI
NotI
P gpdA
pGBFINZDU-WT
3'glaA
amdS
ZDU
P glaA
AscI
XhoI
PacI

Figure 3

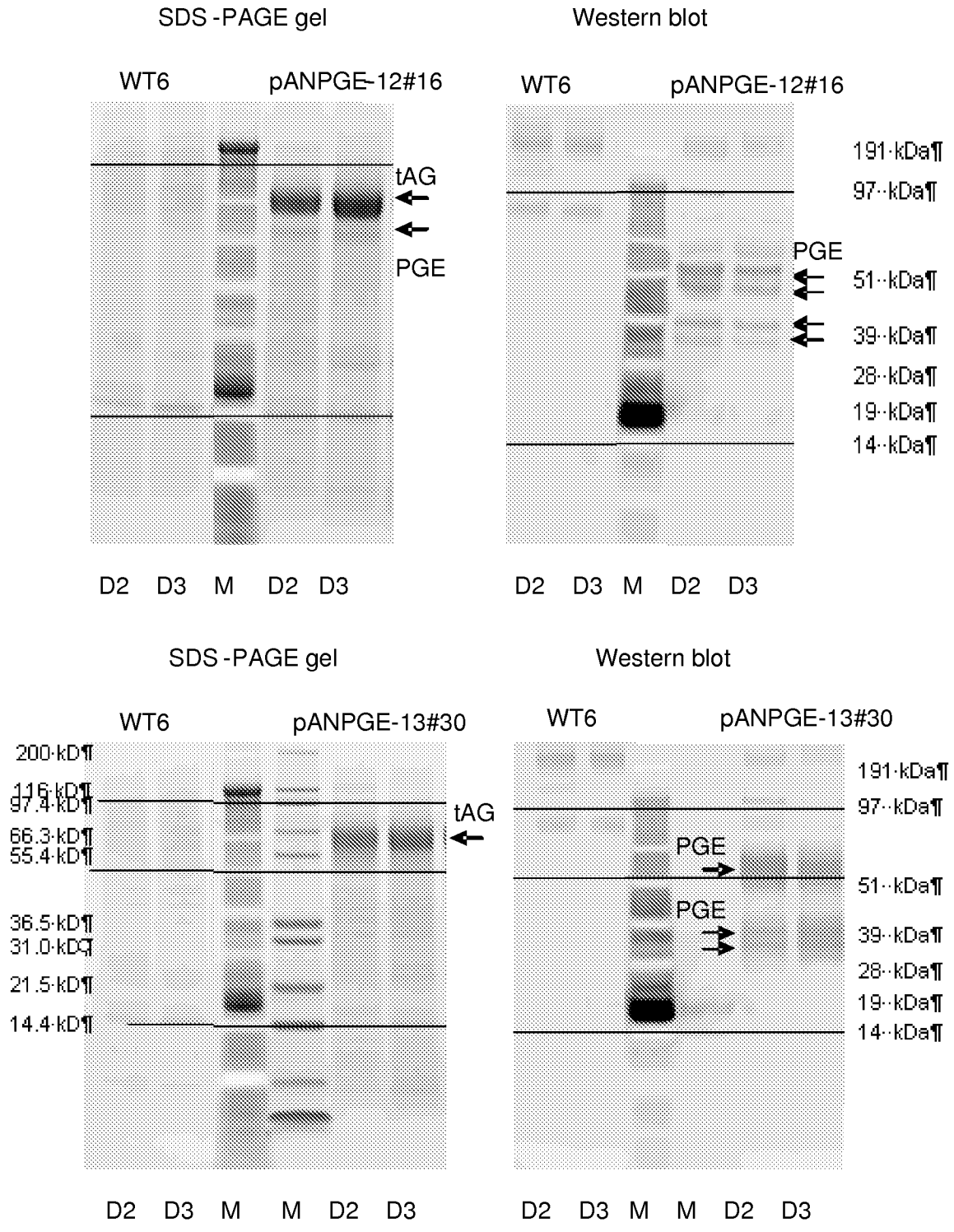
SDS -PAGE gel
Western blot
WT6
pANPGE-12#16
WT6
pANPGE-12#16
tAG
PGE
PGE
191·kDa¶
97··kDa¶
51··kDa¶
39··kDa¶
28··kDa¶
19··kDa¶
14··kDa¶
D2 D3 M D2 D3
D2 D3 M D2 D3
SDS -PAGE gel
Western blot
WT6
pANPGE-13#30
WT6
pANPGE-13#30
200·kD¶
116·kD¶
97.4·kD¶
66.3·kD¶
55.4·kD¶
36.5·kD¶
31.0·kD¶
21.5·kD¶
14.4·kD¶
tAG
PGE
PGE
191·kDa¶
97··kDa¶
51··kDa¶
39··kDa¶
28··kDa¶
19··kDa¶
14··kDa¶
D2 D3 M M D2 D3
D2 D3 M M D2 D3

Figure 4

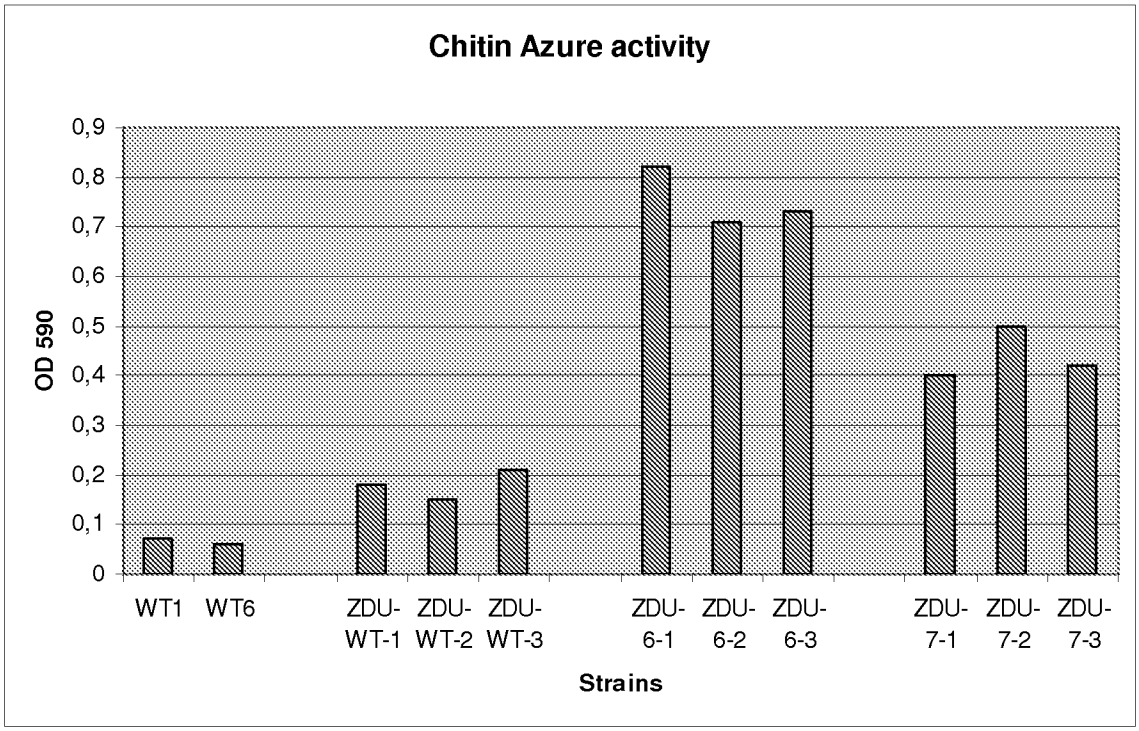
Chitin Azure activity
OD 590
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
WT1
WT6
ZDU-WT-1
ZDU-WT-2
ZDU-WT-3
ZDU-6-1
ZDU-6-2
ZDU-6-3
ZDU-7-1
ZDU-7-2
ZDU-7-3
Strains

Figure 5

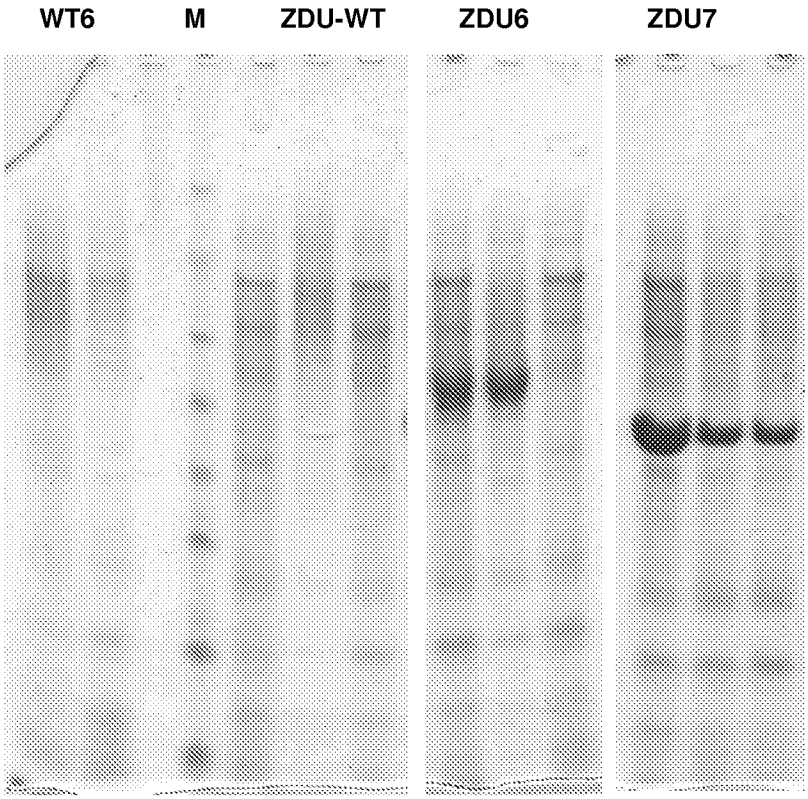

Figure 6

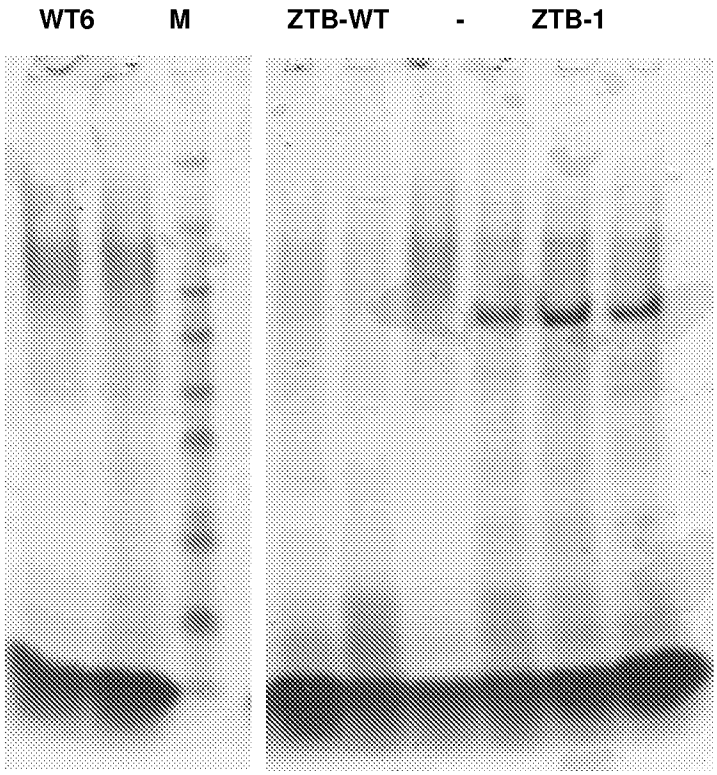

Figure 7

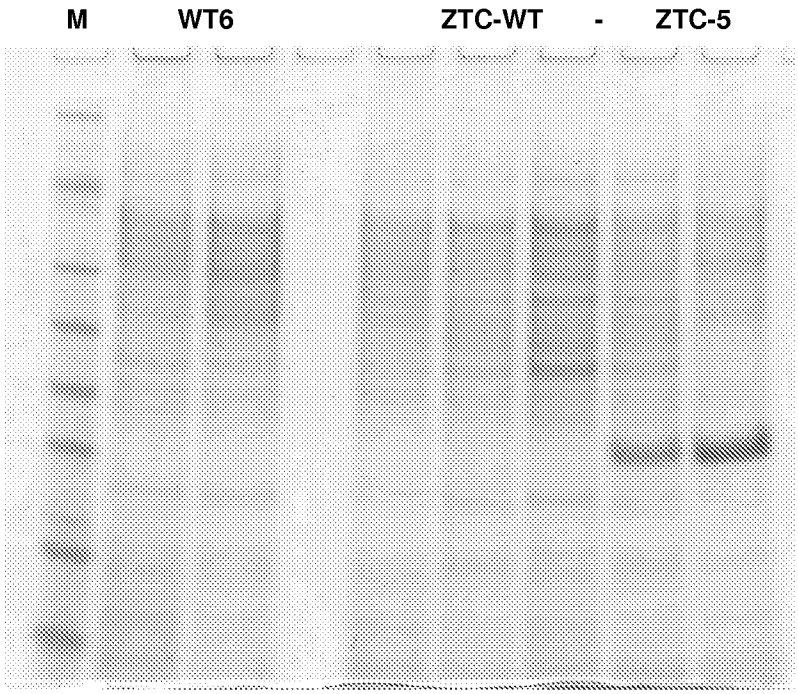

Figure 8

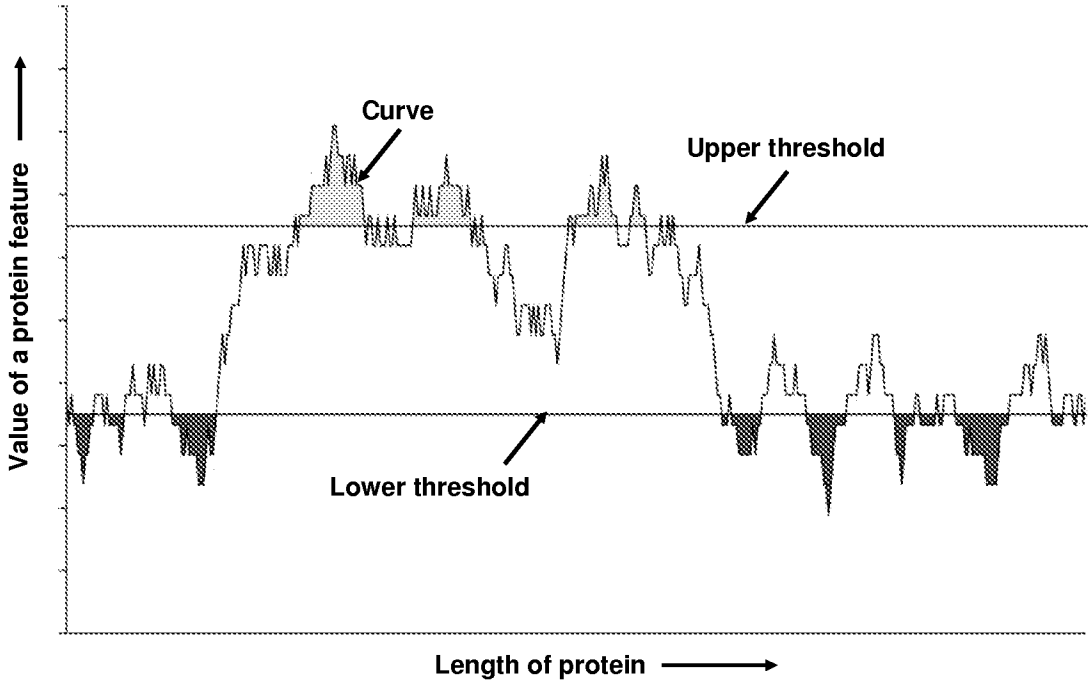
Curve
Upper threshold
Value of a protein feature
Lower threshold
Length of protein

Figure 9

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2010/052918

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12P21/02 C12N15/67
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12P C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, BIOSIS, EMBASE, WPI Data, Sequence Search

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 02/055717 A2 (GENENCOR INT [US]; KOLKMAN MARC [NL]) 18 July 2002 (2002-07-18) * abstract page 2, line 32 - page 3, line 26	27-30
X	WO 96/05228 A1 (CREATIVE BIOMOLECULES INC [US]; US HEALTH [US]; JOST CAROLINA R [US];) 22 February 1996 (1996-02-22) * abstract page 4, line 11 - line 21	27-30
A	WO 2008/000632 A1 (DSM IP ASSETS BV [NL]; ROUBOS JOHANNES ANDRIES [NL]; PEIJ VAN NOEL NIC) 3 January 2008 (2008-01-03) cited in the application the whole document	1-26

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 June 2010	19/07/2010

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Sonnerat, Isabelle

2

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2010/052918

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GOUKA R J ET AL: "Efficient production of secreted proteins by Aspergillus: progress, limitations and prospects." APPLIED MICROBIOLOGY AND BIOTECHNOLOGY JAN 1997 LNKD- PUBMED:9035405, vol. 47, no. 1, January 1997 (1997-01), pages 1-11, XP002540667 ISSN: 0175-7598 page 8, right-hand column - page 9, right-hand column	1-26

2

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2010/052918

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 02055717 A2	18-07-2002	AU 2002243205 A1	24-07-2002
		EP 1354056 A2	22-10-2003
WO 9605228 A1	22-02-1996	AU 700475 B2	07-01-1999
		AU 3244695 A	07-03-1996
		CA 2197232 A1	22-02-1996
		EP 0779897 A1	25-06-1997
		JP 10504458 T	06-05-1998
		US 5888773 A	30-03-1999
WO 2008000632 A1	03-01-2008	AU 2007263880 A1	03-01-2008
		CA 2657975 A1	03-01-2008
		CN 101490262 A	22-07-2009
		EA 200900096 A1	30-06-2009
		EP 2035561 A1	18-03-2009
		JP 2009540845 T	26-11-2009
		US 2009286280 A1	19-11-2009