Abstract: Provided are polynucleotide sequences and synthetic genes encoding celllobiohydrolase enzymes for expression in a host organism with improved and/or refined translational kinetics, and methods of making same. The resultant celllobiohydrolase-encoding nucleotide is predicted to be translated rapidly along its entire length. Expression of the resultant celllobiohydrolase-encoding nucleotide is predicted to result in improved protein expression levels in cases where inappropriate or excessive translation pauses reduce protein expression. In addition, expression of the resultant celllobiohydrolase-encoding nucleotide is predicted to result in improved levels of active and/or natively folded and functional polypeptide expression in cases where inappropriate or excessive translational pauses causes expression of inactive, insoluble, aggregated or somehow dysfunctional or minimally active celllobiohydrolase.
Declaration under Rule 4.17:
— of inventorship (Rule 4.17(iv))
Published:
— with international search report
— before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments
— with sequence listing part of description published separately in electronic form and available upon request from the International Bureau
FIELD OF THE INVENTION

[0001] The present invention relates to refining the translational kinetics of an mRNA into polypeptide, and polypeptide-encoding nucleotide sequences which have refined translational properties.

DESCRIPTION OF THE RELATED ART

[0002] Recent innovations have shown that enzymes can be useful for industrial applications. However, production of large amounts of functional enzyme is often limited. Despite the burgeoning knowledge of expression systems and recombinant DNA, significant obstacles remain when one attempts to express a foreign or synthetic gene in a non-native host organism. Often, a synthetic gene, even when coupled with a strong promoter, is inefficiently translated and can produce a low yield of protein, a faulty protein, or in many cases, low yields of an inactive protein. The same is frequently true of exogenous genes foreign to the expression organism. Even when the gene is translated in a sufficiently efficient manner that recoverable quantities of the translation product are produced, the protein is often inactive, insoluble, aggregated, or otherwise different in structural and activity properties from the native protein expressed in the native organism.

[0003] One family of enzymes that hold industrial applications is cellulases. Cellulases are enzymes used in the treatment of textiles and in the production of cellulosic ethanol that are capable of hydrolysis of the β-D-glucosidic linkages in cellulose. Cellulolytic enzymes have been traditionally divided into three major classes: endoglucanases, exoglucanases or cellobiohydrolases and β-glucosidases (Knowles, J. et al., (1987), Trends Biotech. 5, 255-261); and are known to be produced by a large number of bacteria, yeasts and fungi. An endoglucanase hydrolyses β-1,4-glycosidic linkages randomly. It does not attack cellobiose but hydrolyzes celloolactones, phosphoric-acid-swollen cellulose and substituted celluloses like CMC (carboxymethyl cellulose) and HEC (hydroxyethyl cellulose). Some endoglucanases also act on crystalline cellulose. By contrast, a cellobiohydrolase acts on cellulose and, in particular, splits off cellobiose.
units from the non-reducing end of the chain. Cellobiohydrolase hydrolyses cellodextrins but not cellobiose. β-Glucosidase hydrolyses celllobiose and cellooligosaccharides to glucose, but does not attack cellulose or higher cellodextrins.

[0004] The filamentous fungus Trichoderma reesei produces a complete set of cellulosic enzymes needed for efficient solubilization of native cellulose. Its two cellobiohydrolases, cellobiohydrolase-I (CBH-I) and cellobiohydrolase-II (CBH-II) are key enzymes in the breakdown of crystalline cellulose. CBH-II is an exoglucanase releasing predominantly cellobiose from the ends of the polymeric glucose chains. In large-scale ethanol production, a steady supply of CBH-II can play a valuable role.

[0005] Despite knowledge in the art related to expression of a foreign or synthetic gene in a host organism, T. reesei CBH-II (TrCBH-II) does not express well in host organisms such as Escherichia coli or Saccharomyces cerevisiae. As a result, large-scale production is limited. Therefore, there is a continued need for improved expression of cellobiohydrolase enzymes.

SUMMARY

[0006] Some translational pauses are resultant from the presence of particular codon pairs in the nucleotide sequence encoding the polypeptide to be translated. As provided herein, inappropriate or excessive translation pauses can reduce protein expression considerably. Further, the translational pausing properties of codon pairs vary from organism to organism. As a result, exogenous expression of genes foreign to the expression organism can lead to inefficient translation and poor expression. Even when the gene is translated in a sufficiently efficient manner that recoverable quantities of the translation product are produced, the protein is often inactive, insoluble, aggregated, or otherwise different in properties from the native protein. Thus, removing inappropriate or excessive translation pause structures coded for by specific di-codon nucleotide sequences in the open reading frame (ORF) can improve protein expression.

[0007] In accordance with the above, provided herein are cellobiohydrolase-encoding nucleotide sequences with refined translational kinetics and methods of designing and synthesizing the same. In one embodiment is provided a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has amino acid sequence identity with an original cellobiohydrolase polypeptide, and wherein predicted translation pauses in the expression organism have been removed or reduced by replacing original codon pairs with different codon pairs encoding identical amino acids.
or conservative amino acid substitutions thereof. The resultant cellobiohydrolase-encoding nucleotide is predicted to be translated rapidly along its entire length. Expression of the resultant cellobiohydrolase-encoding nucleotide is predicted to result in improved protein expression levels in cases where inappropriate or excessive translation pauses reduce protein expression. In addition, expression of the resultant cellobiohydrolase-encoding nucleotide is predicted to result in improved levels of active and/or natively folded polypeptide expression products in cases where inappropriate or excessive translation pauses cause expression of inactive, insoluble or aggregated cellobiohydrolase.

[0008] Also provided herein are cellobiohydrolase-encoding DNA sequences, wherein the encoded sequence has amino acid sequence identity with an original cellobiohydrolase-encoding DNA sequence and is adapted for expression in a heterologous host organism, wherein at least 1, 2, or 3 codon pairs of the original sequence that are predicted to cause a translational pause in the host organism have been replaced with codon pairs that are predicted to be less likely to cause a translational pause therein. In selected embodiments, the at least three codon pairs of the original sequence that are predicted to cause a translational pause in the host organism are highly-overrepresented codon pairs therein and have been replaced with codon pairs that are not highly-overrepresented therein. In some embodiments, the host organism is not human, *E. coli* or *S. cerevisiae*.

[0009] In some embodiments are provided a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 3 codon pairs to be replaced are selected from the following: CCCTCT (nucleotides 463-468); GGCCAA (nucleotides 94-99); CAGTTT (nucleotides 565-570); GATATC (nucleotides 703-708); GTGGAA (nucleotides 691-696); GGATTT (nucleotides 1192-1197); GGTATT (nucleotides 1198-1203). In some such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some aspects of the above embodiments, at least 3 of the following codon pair replacements have been made: CCCTCT (nucleotides 463-468) replaced with CCTTCT; GGCCAA (nucleotides 94-99) replaced with GGTCAA;
CAGTTT (nucleotides 565-570) replaced with CAATTT; GATATC (nucleotides 703-708) replaced with GACATT; GTGGAA (nucleotides 691-696) replaced with GTTGAA; GGATTT (nucleotides 1192-1197) replaced with GGTTC; GGTATT (nucleotides 1198-1203) replaced with GGAATT. In certain aspects, the DNA sequence is optimized for expression in *S. cerevisiae*.

[0010] In some embodiments are provided a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 3 codon pairs to be replaced are selected from the following: CTCGCTG (nucleotides 760-765); ATTGCC (nucleotides 631-636); GACAGC (nucleotides 1285-1290); GTCTGG (nucleotides 88-93); GTCTGG (nucleotides 1246-1251); TTGCTG (nucleotides 1231-1236); GTGGTG (nucleotides 571-576); ACGCTG (nucleotides 22-27); ACGCTG (nucleotides 31-36); GACTGG (nucleotides 1168-1173); GCCGGA (nucleotides 559-564); CTGGTG (nucleotides 748-753). In some such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some aspects of the above embodiments, at least 3 of the following codon pair replacements have been made: CTCGCTG (nucleotides 760-765) replaced with CTGGGT; ATTGCC (nucleotides 631-636) replaced with ATTGCG; GACAGC (nucleotides 1285-1290) replaced with GACTCT; GTCTGG (nucleotides 88-93) replaced with GTTTGG; GTCTGG (nucleotides 1246-1251) replaced with GTTTTG; TTGCTG (nucleotides 1231-1236) replaced with CTGGCTG; GTGGTG (nucleotides 571-576) replaced with GTTTGT; ACGCTG (nucleotides 22-27) replaced with ACCCTC; ACGCTG (nucleotides 31-36) replaced with ACCCTG; GACTGG (nucleotides 1168-1173) replaced with GATTTG; GCCGGA (nucleotides 559-564) replaced with GCCGGG; CTGGTG (nucleotides 748-753) replaced with CTGGGT. In certain aspects, the DNA sequence is optimized for expression in *E. coli*.

[0011] In some embodiments are provided a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In
some embodiments, the at least 3 codon pairs to be replaced are selected from the following: CAGTTT (nucleotides 565-570); TTTGAC (nucleotides 1303-1308); TCGT TT (nucleotides 1240-1245); GGCCAA (nucleotides 94-99); AAGAAT (nucleotides 541-546); AAGAAT (nucleotides 934-939); GCCAAA (nucleotides 649-654); GTCAAG (nucleotides 1252-1257); GGTATT (nucleotides 1198-1203); ATCAAC (nucleotides 808-813); GGCCAT (nucleotides 865-870); CTTCCA (nucleotides 835-840); GATATC (nucleotides 703-708); TCGTTG (nucleotides 1228-1233). In some such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some aspects of the above embodiments, at least 3 of the following codon pair replacements have been made: CAGTTT (nucleotides 565-570) replaced with CAATTT; TTTGAC (nucleotides 1303-1308) replaced with TTTGAT; TCGTTT (nucleotides 1240-1245) replaced with TCTTTT; GGCCAA (nucleotides 94-99) replaced with GGACAA; AAGAAT (nucleotides 541-546) replaced with AAAAAT; AAGAAT (nucleotides 934-939) replaced with AAAAAA; GCCAAA (nucleotides 649-654) replaced with GCTAAA; GTCAAG (nucleotides 1252-1257) replaced with GTTAAA; GGTATT (nucleotides 1198-1203) replaced with GGAATC; ATCAAC (nucleotides 808-813) replaced with ATTAAT; GGCCAT (nucleotides 865-870) replaced with GGACAC; CTTCCA (nucleotides 835-840) replaced with TTGCCT; GATATC (nucleotides 703-708) replaced with GATATA; TCGTTG (nucleotides 1228-1233) replaced with TCATTG. In certain aspects, the DNA sequence is optimized for expression in *P. pastoris*.

[0012] In some embodiments are provided a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 3 codon pairs to be replaced are selected from the following: GGCCAA (nucleotides 94-99); CAGTTT (nucleotides 565-570); GATATC (nucleotides 703-708); TATT TG (nucleotides 853-858); GGCCAT (nucleotides 865-870); TCGTTG (nucleotides 1228-1233); TTTGTC (nucleotides 1243-1248); TTCCAA (nucleotides 1363-1368). In some such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some aspects of
the above embodiments, at least 3 of the following codon pair replacements have been made: GGCCAA (nucleotides 94-99) replaced with GGTCAA; CAGTTT (nucleotides 565-570) replaced with CAATTC; GATATC (nucleotides 703-708) replaced with GACATT; TATTTT (nucleotides 853-858) replaced with TATTTA; GGCCAT (nucleotides 865-870) replaced with GGACAT; TCGTTG (nucleotides 1228-1233) replaced with TCTTTA; TTTGTC (nucleotides 1243-1248) replaced with TTGTTT; TTCCAA (nucleotides 1363-1368) replaced with TTCCAG. In certain aspects, the DNA sequence is optimized for expression in K. lactis.

[0013] In some embodiments are provided a cellulbiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellulbiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 3 codon pairs to be replaced are selected from the following: GTGCCT (nucleotides 55-60); GCCAAT (nucleotides 370-375); GCTATT (nucleotides 406-411); GCGGA (nucleotides 559-564); GCCAAT (nucleotides 778-783); TTGGCA (nucleotides 967-972); AAGCTG (nucleotides 1051-1056); GCTATT (nucleotides 1066-1071); GCCAAT (nucleotides 1084-1089); ACCGGA (nucleotides 1147-1152); ACCGGA (nucleotides 1189-1194); GGTATT (nucleotides 1198 - 1203); GACAGC (nucleotides 1285-1290); GATGCC (nucleotides 1327-1332); GCCTTG (nucleotides 1330-1335); CAGCTT (nucleotides 1381-1386). In some such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some aspects of the above embodiments, at least 3 of the following codon pair replacements have been made: GTGCCT (nucleotides 55-60) replaced with GTTCCG; GCCAAT (nucleotides 370-375) replaced with GCTAAT; GCTATT (nucleotides 406-411) replaced with GCCATT; GCGGA (nucleotides 559-564) replaced with GCTGGT; GCCAAT (nucleotides 778-783) replaced with GCGAAT; TTGGCA (nucleotides 967-972) replaced with TTGGCT; AAGCTG (nucleotides 1051-1056) replaced with AAATTG; GCTATT (nucleotides 1066-1071) replaced with GCCATT; GCCAAT (nucleotides 1084-1089) replaced with GCTAAT; ACCGGA (nucleotides 1147-1152) replaced with ACCGGT; ACCGGA (nucleotides 1189-1194) replaced with ACAGGT; GGTATT (nucleotides 1198 - 1203) replaced with GGAATC; GACAGC (nucleotides 1285-1290) replaced with GATTCT; GATGCC (nucleotides
1327-1332) replaced with GACGCC; GCCTTG (nucleotides 1330-1335) replaced with GCCCTT; CAGCTT (nucleotides 1381-1386) replaced with CAGTTG. In certain aspects, the DNA sequence is optimized for expression in \textit{Z. mobilis}.

\[0014\] Also provided herein is a cellulbiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75\% amino acid sequence identity with amino acids 27-471 of wild-type cellulbiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, and wherein at least three codon pairs of the wild-type sequence that are predicted to cause a translational pause in the host organism have been replaced with codon pairs that are predicted to be less likely to cause a translational pause therein. In selected embodiments, the codon pairs of the wild-type sequence that are predicted to cause a translational pause in the host organism are highly-overrepresented codon pairs therein and have been replaced with codon pairs that are not highly-overrepresented therein, wherein a highly-overrepresented codon pair is a codon pair that has a translational kinetics value greater than 5, or 3, or 2.5, or 2 times the standard deviation of translational kinetics values for the host organism. In certain embodiments, the host organism is not human, \textit{E. coli} or \textit{S. cerevisiae}.

\[0015\] Also provided herein is a cellulbiohydrolase-encoding DNA sequence, having at least a 75\% amino acid sequence identity with amino acids 27-471 of wild-type cellulbiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least three codon pairs of the wild-type sequence that are predicted to cause a translational pause in the host organism have been replaced with codon pairs that are predicted to be less likely to cause a translational pause therein, and wherein the host organism is selected from the following: \textit{Pichia pastoris}; \textit{Oryctolagus cuniculus} (rabbit); \textit{Macaca fascicularis} (Long-tailed monkey); \textit{M. mulatta} (Monkey); \textit{E. coli} K12 W3110; \textit{E. coli} UTI89; \textit{E. coli} O157:H7 EDL933; \textit{E. coli} O157:H7 str. Sakai; \textit{Bombyx mori}; \textit{Spodoptera frugipera}; \textit{Drosophila melanogaster}; \textit{Kluyveromyces lactis}, \textit{Zymomonas mobilis} and \textit{Schizosaccharomyces pombe}.

\[0016\] Also provided herein is an expression system, comprising an expression vector in a host organism, wherein the expression vector includes the DNA sequence of the embodiments provided herein, operably linked to an expression control sequence.

\[0017\] In some embodiments, provided herein is a system for degrading cellulose, comprising one or more host organisms that collectively include DNA sequences operably encoding the following enzymes: endo-1,4- \(\beta\)-glucanase, exo-1,4- \(\beta\)-D-
glucanase, and β-D-glucosidase; wherein the enzymes are heterologous to the one or more host organisms, and wherein transcriptional kinetics of each of the DNA sequences encoding the enzymes has been modified to replace at least three codon pairs present in the original sequence for each enzyme, wherein the at least three replaced codon pairs are predicted to cause a translational pause in the host organism, and wherein said modification results in silent permutation or conservative amino acid substitution of said at least three codon pairs. In some aspects, the one or more host organisms are selected from the group consisting of: Saccharomyces cerevisiae, Pichia pastoris, Escherichia coli, Bombyx mori, Spodoptera frugiperda, Drosophila melanogaster, Kluyveromyces lactis, Zymomonas mobilis and Schizosaccharomyces pombe. In some aspects, each encoded enzyme has at least a 75% amino acid sequence identity with the original sequence of the enzyme. In some aspects the exo-1,4- β-D-glucanase retains at least 75% of the enzymatic activity of wild-type TrCBH-II (SEQ ID NO: 2) under normal physiological conditions.

[0018] In some embodiments are provided a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least 1, 2 or 3 codon pairs present in SEQ ID NO:1 and which encode amino acids 27-62 of SEQ ID NO: 2 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof, and wherein at least one replacement codon pair is predicted to be less likely to cause a translational pause in the heterologous host organism relative to the respective wild type codon pair when expressed in the heterologous host organism. In certain aspects, the z score of at least one replacement codon pair when expressed in the heterologous host organism is no more than 150% of the z score for the wild type codon pair when expressed in the native organism. In certain aspects, no replacement codon encoding amino acids 27-62 of SEQ ID NO: 2 has a z score for expression in the heterologous host that is more than 400%, or 300%, or 200%, or 150% or 100% of the mean or median of the five highest z scores of the wild type codon pairs encoding amino acids 27-62 when expressed in the native organism. In certain aspects, no replacement codon encoding amino acids 27-62 of SEQ ID NO: 2 has a z score for expression in the heterologous host that is more than 400%, or 300%, or 200%, or 150% or 100% of the wild type codon pair TCCAAC when expressed in the native organism.
[0019] In some embodiments are provided a cellobiohydrolase-en coding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least 1, 2 or 3 codon pairs present in SEQ ID NO:1 and which encode amino acids 107-471 of SEQ ID NO: 2 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof, and wherein at least one replacement codon pair is predicted to be less likely to cause a translational pause in the heterologous host organism relative to the respective wild type codon pair when expressed in the heterologous host organism. In certain aspects, the z score of at least one replacement codon pair when expressed in the heterologous host organism is no more than 150% of the z score for the wild type codon pair when expressed in the native organism. In certain aspects, no replacement codon encoding amino acids 107-471 of SEQ ID NO: 2 has a z score for expression in the heterologous host that is more than 400%, or 300%, or 200%, or 150% or 100% of the mean or median of the five highest z scores of the wild type codon pairs encoding amino acids 107-471 when expressed in the native organism. In certain aspects, no replacement codon encoding amino acids 107-471 of SEQ ID NO: 2 has a z score for expression in the heterologous host that is more than 400%, or 300%, or 200%, or 150% or 100% of the wild type codon pair GCAAAAG when expressed in the native organism.

[0020] In some embodiments are provided a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least 1, 2 or 3 codon pairs present in SEQ ID NO:1 and which encode amino acids 62-107 of SEQ ID NO: 2 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof, and wherein at least one replacement codon pair is predicted to be equally or more likely to cause a translational pause in the heterologous host organism relative to the respective wild type codon pair when expressed in the heterologous host organism. In certain aspects, the z score of at least one replacement codon pair when expressed in the heterologous host organism is at least 75% of the z score for the wild type codon pair when expressed in the native organism. In certain aspects, at least one replacement codon encoding amino acids 62-107 of SEQ ID NO: 2 has a z score for expression in the heterologous that is more than 200%, or 100%,
or 75%, or 50% or 40% of the mean or median of the five highest z scores of the wild type codon pairs encoding amino acids 62-107 when expressed in the native organism. In certain aspects, at least one replacement codon encoding amino acids 62-107 of SEQ ID NO: 2 has a z score for expression in the heterologous host that is more than 200%, or 100%, or 75%, or 50% or 40% of the wild type codon pair TCTACT when expressed in the native organism.

[0021] Also provided are polynucleotides comprising any of the DNA sequences provided herein. Also provided herein are isolated polynucleotides comprising the DNA sequence of SEQ ID NOs:3, 5, 7, 9, 11, 13, 15, 17, 19, 21 or 23. In some embodiments, such a polynucleotide is a DNA polynucleotide, while also contemplated herein, such a polynucleotide can be an RNA polynucleotide comprising the RNA-equivalent of said DNA sequence. Also provided are cells comprising such a polynucleotide. In some such cells, the cell expresses the polypeptide encoded by the polynucleotide. Also provided are methods of introducing a polynucleotide into a host cell comprising providing a host cell; and contacting said host cell with any of the polynucleotides provided herein under conditions that permit the polynucleotide to be introduced into the host cell. Also provided are methods of expressing a polypeptide comprising providing a cell comprising any of the polynucleotides provided herein; and placing the cell under conditions that permit the cell to express the polypeptide encoded by the DNA sequence, whereby said encoded polypeptide is expressed by said cell. Also provided are methods of hydrolyzing a carbohydrate comprising providing a carbohydrate comprising at least one glycosidic bond; providing a polypeptide encoded by any of the polynucleotides provided herein; and contacting said carbohydrate with said polypeptide under conditions that permit said polypeptide to hydrolyze at least one glycosidic bond of said carbohydrate; whereby at least one glycosidic bond of said carbohydrate is hydrolyzed. In some such embodiments, the carbohydrate is cellulose. In some such embodiments, the carbohydrate comprises two or more β-1,4-linked glucose units. Also provided herein are isolated polypeptides encoded by any of the DNA sequences provided herein, provided that the amino acid sequence of said polypeptide is not SEQ ID NO: 2.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] Figure 1 depicts a graphical display of z scores of translational kinetics values for codon pair utilization in T. Reesei of nucleic acid sequences encoding the
cellulbiohydrolase-II enzyme of *T. Reesei* (TrCBH-II), plotted as a function of codon pair position.

[0023] Figures 2-6 depicts effects of Translational Engineering™ on protein expression levels. Each of Figures 2-6 depict graphical displays of z scores of translational kinetics values for codon pair utilization of nucleic acid sequences encoding TrCBH-II, plotted as a function of codon pair position.

[0024] Figure 2A depicts a graphical display of the *S. cerevisiae* expression of the native nucleic acid sequence encoding the TrCBH-II protein. Figure 2B depicts a graphical display of the *S. cerevisiae* expression of a nucleic acid sequence encoding the TrCBH-II which has been modified to eliminate codon pairs that are predicted to cause a translational pause in *S. cerevisiae*.

[0025] Figure 3A depicts a graphical display of the *E. coli* expression of the native nucleic acid sequence encoding the TrCBH-II protein. Figure 3B depicts a graphical display of the *E. coli* expression of a nucleic acid sequence encoding the TrCBH-II which has been modified to eliminate codon pairs that are predicted to cause a translational pause in *E. coli*.

[0026] Figure 4A depicts a graphical display of the *P. pastoris* expression of the native nucleic acid sequence encoding the TrCBH-II protein. Figure 4B depicts a graphical display of the *P. pastoris* expression of a nucleic acid sequence encoding the TrCBH-II which has been modified to eliminate codon pairs that are predicted to cause a translational pause in *P. pastoris*.

[0027] Figure 5A depicts a graphical display of the *K. lactis* expression of the native nucleic acid sequence encoding the TrCBH-II protein. Figure 5B depicts a graphical display of the *K. lactis* expression of a nucleic acid sequence encoding the TrCBH-II which has been modified to eliminate codon pairs that are predicted to cause a translational pause in *K. lactis*.

[0028] Figure 6A depicts a graphical display of the *Z. mobilis* expression of the native nucleic acid sequence encoding the TrCBH-II protein. Figure 6B depicts a graphical display of the *Z. mobilis* expression of a nucleic acid sequence encoding the TrCBH-II which has been modified to eliminate codon pairs that are predicted to cause a translational pause in *Z. mobilis*. 
DETAILED DESCRIPTION

[0029] Biomass is the earth's most attractive alternative among fuel sources and most sustainable energy resource and is reproduced by the bioconversion of carbon dioxide. Ethanol produced from biomass is today the most widely used biofuel when blended with gasoline. As the carbon dioxide released by combustion is recycled into biomass, the use of biofuels can significantly reduce the accumulation of greenhouse gas. Ethanol is just one example of the uses of biomass harvesting using industrial enzymes. The technologies associated with biomass harvesting are similarly applicable in the production of other biofuels, fine chemicals as well as other diverse applications.

[0030] Cellulose is the major polysaccharide of plants, where it plays a predominantly structural role. In recent years, it has been proposed that waste cellulosic biomass could be used as a cheap and readily available sugar to replace starchy materials in fermentation. Many researchers have previously tried to develop an efficient and inexpensive process for ethanol and other biofuels production from such waste by using recombinant bacteria and yeast (e.g., *Saccharomyces cerevisiae* (Aristidou and Penttila (2000) Curr. Opin. Biotechnol. 11:187-198; Bothast et al. (1999) Biotechnol. Prog. 15:867-875; Ingram et al. (1998) Biotechnol. Bioeng. 58:204-214), but so far with limited success. A process of this kind can address environmental problems such as global warming and lessen dependence on fossil fuels.

[0031] A variety of highly specialized microorganisms have evolved to produce enzymes that either synergistically or in complexes can carry out the complete hydrolysis of cellulose. The anaerobic bacteria *Clostridium thermocellum* and *Clostridium cellulovorans* and the filamentous fungus *Trichoderma reesei* are known as cellulolytic and xylanolytic microorganisms. The bacteria *C. thermocellum* and *C. cellulovorans* produce a cellulose complex consisting of cellulase and hemicellulase organized on the cell surface (Doi and Tamaru (2001) Chem. Rec. 1:24-32; Shoham et al. (1999) Trends Microbiol. 7:275-281). In an exemplary, well-characterized organism, *T. reesei*, three types of cellulolytic enzyme are extracellularly secreted, including five endoglucanases (EG [EC 3.2.1.4]) (Okada et al (1998) Appl. Environ. Microbiol. 64:555-563), two cellobiohydrolases (CBH [EC 3.2.1.91]) (Henrissat et al. (1985) Bio/Technology 3:722-726; Teeri et al. (1987) Gene 51:43-52), and two β-glucosidases (BGL [EC 3.2.1.21]) (Chen et al. (1992) Biochim. Biophys. Acta 1121:54-60). Endoglucanases act randomly against the amorphous region of the cellulose chain to produce reducing and nonreducing ends for cellobiohydrolases, which produce cellobiose.
from reducing or nonreducing ends of crystalline cellulose. Exoglucanase enzymes, including CBH-I and CBH-II, liberate the disaccharide D-cellobiose from 1,4-β-glucans. Cellulose chains are thus efficiently degraded to soluble cellobiose and cellobiohydrolases, by the endo-exo synergism of EG and CBH (Henrissat et al. (1985) Bio/Technology 3:722-726). In the last step of enzymatic cellulose degradation, cellobiohydrolases are hydrolyzed to glucose by β-glucosidase. In addition to endo-exo synergism, exo-exo synergism between the two cellbiohydrolases has also been reported (Teeri, T. T. (1997) Trends Biotechnol. 15:160-167).

The predominant polysaccharide in the primary cell wall of biomass is cellulose, the second most abundant is hemi-cellulose, and the third is pectin. The secondary cell wall, produced after the cell has stopped growing, also contains polysaccharides and is strengthened through polymeric lignin covalently cross-linked to hemicellulose. Cellulose is a homopolymer of anhydrocellobiose and thus a linear β-(1-4)-D-glucan, while hemicelluloses include a variety of compounds, such as xylans, xylglucans, arabinoxylans, and mannans in complex branched structures with a spectrum of substituents. Although generally polymorphous, cellulose is found in plant tissue primarily as an insoluble crystalline matrix of parallel glucan chains. Hemicelluloses usually hydrogen bond to cellulose, as well as to other hemicelluloses, which helps stabilize the cell wall matrix.

DNA constructs encoding cellulase enzymes, including cellbiohydrolases, are known in the art. For example, U.S. Patent No. 5,686,593 relates to cellulose- or hemicellulose-degrading enzymes that are derivable from a fungus other than Trichoderma or Phanerochaete, and which comprise a carbohydrate binding domain homologous to a terminal A region of T. reesei cellulases.


It is known that when using other recombinant ethanologenic bacteria or yeast to ferment cellulose, addition of commercial cellulase is necessary for ethanol
production. For example, when *T. reesei* endoglucanase II and CBH-II, and *Aspergillus aculeatus* \(\beta\)-glucosidase 1, were simultaneously co-displayed on the cell surface of a yeast strain, the yeast strain was able to directly produce ethanol from cellulose, (whereas a yeast strain co-displaying only \(\beta\)-glucosidase 1 and endoglucanase II could not) indicating the key role of CBH-II in the industrial conversion of cellulose to ethanol (Fujita et al. (2004) Appl Environ Microbiol. 70:1207-1212).

[0036] Despite knowledge in the art related to expression of a foreign or synthetic gene in a host organism, *T. reesei* CBH-II does not express well in host organisms such as *E. coli* or *S. cerevisiae*. Accordingly, provided herein are cellobiohydrolase-encoding nucleotide sequences and methods of making the same for improved expression of cellobiohydrolase enzymes.

[0037] Some translational pauses are resultant from the presence of particular codon pairs in the nucleotide sequence encoding the polypeptide to be translated. As provided herein, inappropriate or excessive translation pauses can reduce protein expression considerably. Further, the translational pausing properties of codon pairs vary from organism to organism. As a result, exogenous expression of genes foreign to the expression organism can lead to inefficient translation. Even when the gene is translated in a sufficiently efficient manner that recoverable quantities of the translation product are produced, the protein is often inactive, insoluble, aggregated, or otherwise different in properties from the native protein. Thus, removing inappropriate or excessive translation pauses can improve protein expression.

[0038] Methods of determining patterns of codon pair utilization are known in the art, as exemplified by U.S. Patent Number 5,082,767 (which is incorporated by reference herein in its entirety), which describes analysis of patterns of nonrandom codon pair usage. The information obtained from codon pair utilization analysis can be used to construct and express altered or synthetic genes having desired levels of translational efficiency, to introduce translational pause sites into heterologous genes, and to ascertain relationship or ancestral origin of nucleotide sequences in accordance with the methods provided herein and the knowledge in the art.

[0039] A translational pause can serve to slow translation of the nascent amino acid chain. In some instances when such translational pauses arise in translation in native genes in the native organism, the pause(s) can serve to facilitate proper polypeptide folding, post-translational modification, re-organization/folding at protein domain boundaries, or other steps toward arriving at the native, active wild type protein.
Accordingly, in some embodiments provided herein, one or more pauses that are predicted to be present in native translation of cellobiohydrolase is/are preserved in a modified cellobiohydrolase-encoding polynucleotide provided in accordance with the teachings herein. For example, a codon pair in the modified cellobiohydrolase-encoding polynucleotide can be selected to have a predicted translational kinetics value that is at least 50%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, or 99% that of the native codon pair whose predicted pause is to be preserved; further, the codon pair in the modified cellobiohydrolase-encoding polynucleotide can be selected to be located within 30, 20, 10, 9, 8, 7, 6, 5, 4, 3, 2 or 1 codons of the native codon pair whose predicted pause is to be preserved.

[0040] Accordingly, as used herein, Translation Engineering™ refers to a process used to modify the translational kinetics of a polypeptide-encoding nucleic sequence. For example, Translation Engineering™ can be applied to modify the translational kinetics of a polypeptide-encoding nucleic sequence when expressed in its native organism. In another example, Translation Engineering™ can be applied to modify the translational kinetics of a polypeptide-encoding nucleic sequence when expressed in its native organism. In some embodiments, this process alters the polypeptide-encoding nucleic sequence to optimize codon usage and codon pair optimization in the organism in which the polypeptide-encoding nucleic sequence is expressed. For example, sequence modifications can be made to place or prevent restriction sites in the sequence, eliminate strong RNA secondary structures and avoid inadvertent Shine-Delgarno sequences. Additionally, Translation Engineering™ involves modifying the translational kinetics of a polypeptide-encoding nucleic sequence by removing, preserving, and/or inserting translational pauses into the polypeptide-encoding nucleic sequence.

[0041] In accordance with the above, provided herein are cellobiohydrolase-encoding nucleotide sequences with refined translational kinetics and methods of making same. In one embodiment, provided is a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has amino acid sequence identity with wild-type cellobiohydrolase, and wherein predicted translation pauses in the expression organism have been removed or reduced by replacing input-sequence codon pairs with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the resultant cellobiohydrolase-encoding nucleotide is predicted to be translated rapidly along its entire length. Expression of the resultant
cellobiohydrolase-encoding nucleotide is predicted to result in improved protein expression levels in cases where inappropriate or excessive translation pauses reduce protein expression. In addition, expression of the resultant cellobiohydrolase-encoding nucleotide is predicted to result in improved levels of active and/or natively folded polypeptide expression in cases where inappropriate or excessive translation pauses causes expression of inactive, insoluble or aggregated cellobiohydrolase. In some embodiments, expression of the resultant cellobiohydrolase-encoding nucleotide is predicted to result in improved levels of active and/or natively folded polypeptide expression in cases where one or more predicted pauses are preserved from the native expression profile or are added to preserve expression of active and/or soluble cellobiohydrolase. Thus, the cellobiohydrolase-encoding nucleotide sequences provided herein allow for one or more of the following results: higher expression levels; higher enzymatic activity; greater protein stability and resistance to degradation; and increased solubility.

[0042] Accordingly, nucleic acid sequences encoding the cellobiohydrolase-II enzyme of T. Reesei (TrCBH-II) are provided. The nucleotide sequences provided herein include the native sequence from T. Reesei shown in the sequence listing (SEQ ID NO: 1) which encodes the TrCBH-II amino acid sequence (SEQ ID NO: 2).

[0043] Further, provided herein are nucleic acid sequences encoding TrCBH-II with refined translational kinetics for expression in S. cerevisiae (SEQ ID NO: 3), E. coli (SEQ ID NO: 9), P. pastoris (SEQ ID NO: 15), K. lactis (SEQ ID NO: 21) and Z. mobilis (SEQ ID NO: 23). Also provided herein are sequences where additional sequence has been added to the 3' or 5' ends, or both. As will be understood by one of skill in the art, nucleotide sequences may be added 3' or 5' of any nucleic acid, for example, to facilitate hybridization of PCR primers, to add cloning restriction sites or other sites that facilitate cloning and/or expression. Accordingly, provided in the sequence listing are nucleic acid sequences with additional 5' and 3' cloning and/or PCR sequences, and which encode TrCBH-II with refined translational kinetics for expression in S. cerevisiae (SEQ ID NOS: 5 and 7), E. coli (SEQ ID NOS: 11 and 13) and P. pastoris (SEQ ID NOS: 17 and 19).

[0044] Further, provided in the sequence listing are TrCBH-II amino acid sequences encoded by the nucleotide sequences with refined translational kinetics described herein. Thus, TrCBH-II nucleic acid sequences with refined translational kinetics for expression in S. cerevisiae (SEQ ID NOS: 3, 5 and 7) encode the amino acid
sequences shown in the sequence listing (SEQ ID NOS: 4, 6 and 8). TrCBH-II nucleic acid sequences with refined translational kinetics for expression in *E. coli* (SEQ ID NOS: 9, 11 and 13) encode the amino acid sequences shown in the sequence listing (SEQ ID NOS: 10, 12 and 14). TrCBH-II nucleic acid sequences with refined translational kinetics for expression in *P. pastoris* (SEQ ID NOS: 15, 17 and 19) encode the amino acid sequences shown in the sequence listing (SEQ ID NOS: 16, 18 and 20). TrCBH-II nucleic acid sequences with refined translational kinetics for expression in *K. lactis* (SEQ ID NO: 21) encode the amino acid sequences shown in the sequence listing (SEQ ID NO: 22). TrCBH-II nucleic acid sequences with refined translational kinetics for expression in *Z. mobilis* (SEQ ID NO: 23) encode the amino acid sequences shown in the sequence listing (SEQ ID NO: 24).

[0045] Also provided herein are cellobiohydrolase-encoding DNA sequences, and polynucleotides comprising such DNA sequences, wherein the encoded sequence has amino acid sequence identity with an original cellobiohydrolase polypeptide and is adapted for expression in a heterologous host organism, wherein at least three codon pairs of the original sequence that are predicted to cause a translational pause in the host organism have been replaced with codon pairs that are predicted to be less likely to cause a translational pause therein. In selected embodiments, the at least three codon pairs of the original sequence that are predicted to cause a translational pause in the host organism are highly-overrepresented codon pairs therein and have been replaced with codon pairs that are not highly-overrepresented therein. In some embodiments, the host organism is not human, *E. coli* or *S. cerevisiae*.

[0046] As used herein, a cellobiohydrolase polynucleotide encodes a polypeptide having cellobiohydrolase activity. Cellobiohydrolase, exoglucanase, exo-1,4-β-D-glucanase and like terms refers to the enzymatic hydrolysis of a glucoside bond in a polysaccharide or an oligosaccharide containing D-glucose subunits bonded through β-1,4 bonds, to release cellobiose, a disaccharide in which D-glucose is bonded through a β-1,4 bond. A method for measuring the cellobiohydrolase activity is exemplified by a known method in which an enzymatic reaction is carried out using phosphoric acid-swollen cellulose as a substrate and the existence of cellobiose in the reaction is confirmed by thin-layer silica gel chromatography, as described in U.S. Patent No. 6,566,113, hereby incorporated by reference in its entirety.

[0047] The polynucleotides provided herein encode polypeptides that have a cellobiohydrolase activity. Thus, a cellobiohydrolase-encoding polynucleotide
comprising any of the DNA sequences provided herein can be transcribed and the resulting RNA translated to produce a polypeptide with cellobiohydrolase activity.

[0048] In some embodiments, the cellobiohydrolase-encoding DNA sequence is adapted for expression in a heterologous host organism. As used herein, a DNA sequence that has been adapted for expression is a DNA sequence that has been inserted into an expression vector or otherwise modified to contain regulatory elements necessary for expression of the DNA in the host cell, positioned in such a manner as to permit expression of the DNA in the host cell. Such regulatory elements required for expression include promoter sequences, transcription initiation sequences and, optionally, enhancer sequences. For example, a DNA sequence may be inserted into a plasmid vector adapted for expression in a bacterial cell, such as *E. coli* or *Z. mobilis*, a eukaryotic cell, such as *S. cerevisiae*, *K. lactis* or other yeast, or any other host organism.

[0049] A heterologous host organism is an organism used to express DNA, RNA or protein that is foreign to the host organism. In certain aspects, the host organism is not human, *E. coli* or *S. cerevisiae*.

[0050] In some embodiments, polynucleotides provided herein also encode polypeptides that have other glycosidase activities such as an endoglucanase activity and a β-D-glucosidase activity.

**Changes to translational kinetics**

[0051] The methods and sequences provided herein permit modification of the translational kinetics of an mRNA into a TrCBH-II-encoding polypeptide. Translational kinetics of an mRNA into polypeptide can be changed in order to achieve any of a variety of expression profiles. For example, translational kinetics of an mRNA into polypeptide can be changed in order to remove some or all translational pauses. In another example, translational kinetics of an mRNA into polypeptide can be changed in order to replace some or all translational pauses predicted to occur within an autonomous folding unit of a nascent protein. In another example, translational kinetics of an mRNA into polypeptide can be changed in order to replace some or all over-represented codon pairs.

[0052] It is proposed herein that the presence of a pause or translation slowing codon pair can queue ribosomes back to the beginning of the coding sequence, thereby inhibiting further ribosome attachment to the message which can result in down-regulation of protein expression levels as the rate of translation initiation readily saturates and the slowest translation step time becomes rate limiting. It is also proposed herein that
the presence of a pause or translational slowing codon pair can stall or detach a ribosome. It is also proposed herein that the presence of a pause or translational slowing codon pair can expose naked mRNA, which is then subject to message degradation. It is also proposed herein that the presence of a pause or translational slowing codon pair can decouple translation from transcription, leading to protein expression failure. For these reasons and more, methods for analyzing, designing and producing gene sequences and polynucleotides to remove or decrease in number, or selectively preserve or insert, pauses, or to replace or modify translational slowing codon pairs, have great utility.

[0053] Organism-specific codon usage and codon pair usage, and the presence of organism-specific pause sites, result in gene translation that is highly adapted to the original host organism. For example, ribosomal pausing sites that may be functional in a human cell will typically be scrambled, random, or not appropriate or not recognized in the proper context in a bacterium or other non-native host. A heterologous cDNA or synthetic polynucleotide has a random but high probability of inadvertently encoding a pause site somewhere, often leading to protein expression and/or activity failure.

[0054] Differences between codon pair (pause signal) coding among bacteria or among vertebrates are sufficient to make cross-family gene expression unpredictable. For example, in various organisms such as bacteria, a significant pause or translational slowing can result in premature transcription termination and/or messenger degradation. Even in eukaryotes there is a coupling between export of mRNA from the nucleus and translation; thus a different, but still effective system of clearing untranslated mRNA exists in eukaryotes.

[0055] Methods for refining translational kinetics of an mRNA into polypeptide can be performed according to any method known in the art, as exemplified in U.S. Patent Publication No. 2008/0046192, published on February 21, 2008, which is incorporated by reference herein in its entirety. For example, a polypeptide-encoding nucleotide can be designed to be predicted to be translated rapidly along its entire length. Thus, some polypeptide-encoding nucleotides provided herein are those that have been engineered to remove all predicted pauses. Expression of such a polypeptide-encoding nucleotide can result in improved protein expression levels and improved levels of active and/or natively folded polypeptide expression.

[0056] Further methods of refining translational kinetic values are contemplated herein, as can be seen in U.S. Patent Publication No. 2007/0298503, published on December 27, 2007, and U.S. Patent Publication No. 2007/0275399,
published on November 29, 2007, each of which is incorporated by reference herein in its entirety.

[0057] As provided herein, a test of translation pausing or slowing as a result of codon pair usage can be performed by comparing a series of genes that have random pauses with modified genes where codon pairs predicted to cause translational pauses are replaced. Unmodified genes moved from their source organism and expressed in a heterologous host can have an altered set of codon pairs predicted to cause a translational pause or ribosomal slowing (e.g., an altered set of over-represented codon pairs), resulting in altered configuration and location of presumed pause sites. Creation of synthetic codon-pair-optimized genes can have a dramatic effect on expression: expression of difficult-to-express genes can be seen for the first time, or improved at least 2-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, 10-fold, 12-fold, 15-fold, 20-fold, 25-fold, 30-fold, or more, relative to unmodified polypeptide-encoding nucleic acid sequences.

[0058] In some embodiments, translational kinetics of an mRNA into TrCBH-II-encoding polypeptide can be changed in order to remove some or all translational pauses or replace other codon pairs that cause translational slowing, message instability and degradation, and poor protein translation, expression, and functional properties. While not intending to be limited to the following, it is believed that, for at least some proteins, reduction or elimination of translational pauses can serve to increase the expression level and/or quality and characteristics of the protein. Accordingly, by removing some or all translational pauses or replacing other codon pairs that cause translational slowing, the expression levels and/or quality of an expressed protein can be increased.

[0059] For example, the cellobiohydrolase-encoding nucleotide sequences provided herein allow for one or more of the following results: higher expression levels, higher enzymatic activity, greater protein stability, resistance to degradation, and increased solubility compared to the original native gene when expressed in a heterologous host.

[0060] Thus, also provided herein are cellobiohydrolase-encoding nucleotide sequences that have been modified to have one or more transcriptional pauses or slowing sites removed by modifying one or more codon pairs to a corresponding codon pair that is less likely to cause a translational pause or slowing. While in some embodiments it is preferred to replace all codon pairs predicted to cause a translational pause or slowing, in
other embodiments, it is sufficient to replace a subset of codon pairs predicted to cause a translational pause or slowing. For example, expression levels can be increased by replacing at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more codon pairs predicted to cause a translational pause or slowing. In another example, at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98% or 99% of codon pairs predicted to cause a translational pause or slowing are replaced by, for example, substituting different codon pairs that encode the same amino acids.

[0061] In some embodiments, translational kinetics of an mRNA into polypeptide can be changed in order to remove some or all translational pauses predicted to occur within an autonomous folding unit of a protein. As used herein, an autonomous folding unit of a protein refers to an element of the overall protein structure that is self-stabilizing and often folds independently of the rest of the protein chain. Such autonomous folding units typically correspond to a protein domain. As provided herein, expression of a gene in a heterologous host organism can result in translational pauses located in regions that inhibit protein expression and/or protein folding. Since the presence of codon pairs predicted to cause a translational pause or slowing in protein-encoding regions separating regions encoding different autonomous folding units of the protein can serve to pause or slow translation, it is also contemplated that removal of translational pauses predicted to occur within an autonomous folding unit of a protein, particularly for heterologously-expressed proteins, can result in improved expression levels and/or folding of expressed proteins. Accordingly, provided herein are methods of changing translational kinetics of an mRNA into polypeptide by removing some or all translational pauses predicted to occur within an autonomous folding unit of a protein, thereby increasing expression levels and/or improving the folding of the expressed protein.

[0062] It is further contemplated that preserving or inserting a translational pause in a region predicted to separate autonomous folding units of a protein, particularly for heterologously-expressed proteins, can result in improved folding and/or solubility of expressed proteins. Accordingly, provided herein are methods of changing translational kinetics of an mRNA into polypeptide by preserving, relative to native, or inserting one or more translational pauses in one or more regions predicted to separate autonomous folding units of a protein, thereby increasing improving the folding and/or solubility of the expressed protein.
In the methods provided herein that include changing translational kinetics of an mRNA into polypeptide by modifying codon pairs with regard to their location within or outside of autonomous folding units of proteins, one step can include identifying predicted autonomous folding units of a protein. Methods for identifying predicted autonomous folding units of a protein or protein domains are known in the art, and include alignment of amino acid sequences with protein sequences having known structures, and threading amino acid sequences against template protein domain databases. Such methods can employ any of a variety of software algorithms in searching any of a variety of databases known in the art for predicting the location of protein domains. The results of such methods will typically include an identification of the amino acids predicted to be present in a particular domain, and also can include an identification of the domain itself, and an identification of the secondary structural element, if any, in which each amino acid sequence of a domain is located.

In some instances, it is not possible to modify the polypeptide-encoding nucleotide sequence to remove a translational pause not present in the expression profile of the polypeptide in the native host organism. For example, there may be no codon pairs that are not predicted to cause a translational pause or slowing and that encode a corresponding pair of amino acids. In such instances, several options are available: the codon pair that is least likely to cause a translational pause or slowing can be selected; an amino acid insertion, deletion or mutation can be introduced to yield a codon pair that is not predicted to cause a translational pause or slowing; or no change is made. One option in a computational method is to request human input in order to resolve the issue. The computational method may, for example, involve the use of a computer that is programmed to request human input. Alternatively, the computer may be programmed to make a selection, or combination of selections, such that multiple genes, or Ordered Gene Sets or small permutation libraries are designed and synthetically produced for use in expression analysis. In methods in which an amino acid insertion, deletion or mutation is made in order to change translational kinetics, it is preferable to select a change that is predicted not to substantially influence the final three-dimensional structure of the protein and/or the activity of the protein. Such an amino acid insertion, deletion or mutation can include, for example, a conservative amino acid substitution such as the conservative substitutions shown in Table 1. The substitutions shown are based on amino acid physical-chemical properties, and as such, are independent of
organism. In some embodiments, the conservative amino acid substitution is a substitution listed under the heading of exemplary substitutions.

<table>
<thead>
<tr>
<th>Original Residue</th>
<th>Conservative Substitutions</th>
<th>Exemplary Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala (A)</td>
<td>val; leu; ile</td>
<td>val</td>
</tr>
<tr>
<td>Arg (R)</td>
<td>lys; gin; asn</td>
<td>lys</td>
</tr>
<tr>
<td>Asn (N)</td>
<td>gin; his; lys; arg</td>
<td>gin</td>
</tr>
<tr>
<td>Asp (D)</td>
<td>glu</td>
<td>glu</td>
</tr>
<tr>
<td>Cys (C)</td>
<td>ser</td>
<td>ser</td>
</tr>
<tr>
<td>Gln (Q)</td>
<td>asn</td>
<td>asn</td>
</tr>
<tr>
<td>Glu (E)</td>
<td>asp</td>
<td>asp</td>
</tr>
<tr>
<td>Gly (G)</td>
<td>pro; ala</td>
<td>ala</td>
</tr>
<tr>
<td>His (H)</td>
<td>asn; gin; lys; arg</td>
<td>arg</td>
</tr>
<tr>
<td>He (I)</td>
<td>leu; val; met; ala; phe</td>
<td>leu</td>
</tr>
<tr>
<td>Leu (L)</td>
<td>ile; val; met; ala; phe</td>
<td>ile</td>
</tr>
<tr>
<td>Lys (K)</td>
<td>arg; gin; asn</td>
<td>arg</td>
</tr>
<tr>
<td>Met (M)</td>
<td>leu; phe; ile</td>
<td>leu</td>
</tr>
<tr>
<td>Phe (F)</td>
<td>leu; val; ile; ala; tyr</td>
<td>leu</td>
</tr>
<tr>
<td>Pro (P)</td>
<td>ala</td>
<td>ala</td>
</tr>
<tr>
<td>Ser (S)</td>
<td>thr</td>
<td>thr</td>
</tr>
<tr>
<td>Thr (T)</td>
<td>ser</td>
<td>ser</td>
</tr>
<tr>
<td>Trp (W)</td>
<td>tyr; phe</td>
<td>tyr</td>
</tr>
<tr>
<td>Tyr (Y)</td>
<td>trp; phe; thr; ser</td>
<td>phe</td>
</tr>
<tr>
<td>Val (V)</td>
<td>ile; leu; met; phe; ala</td>
<td>leu</td>
</tr>
</tbody>
</table>

[0065] While in some embodiments, all codon pairs predicted to cause a translational pause or slowing are treated equally, in other embodiments, one or more different threshold levels can be established for differential treatment of codon pairs, where codon pairs above a highest threshold are the codon pairs most likely to cause a translational pause or slowing, and succeedingly lower codon pair threshold-based groups
correspond to successively lower likelihoods of the respective codon pairs causing a translational pause or slowing. Based on the codon pair groupings, different numbers or percentages of codon pairs can be replaced for each of these different threshold-based groups. For example, 95% or more codon pairs above a highest threshold level can be replaced, while 90% or less of all codon pairs between that level and an intermediate threshold level are replaced. As contemplated herein, codon pairs likely to cause a translational pause or slowing can be segregated into two or more different threshold-based groups, three or more different threshold-based groups, four or more different threshold-based groups, five or more different threshold-based groups, six or more different threshold-based groups, or more. Discussion of specific thresholds are provided elsewhere herein; however, typically the higher the threshold, the higher the likelihood of a translational pause or slowing caused by a codon pair with a translational kinetics value greater than the threshold. In embodiments in which codon pairs likely to cause a translational pause or slowing can be segregated into two or more different threshold-based groups, different numbers or percentages of codon pairs can be replaced for each codon pair group. For example, in one embodiment, at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98% or 99% of codon pairs above a highest threshold are replaced, while the same or a lower percentage of codon pairs are replaced from codon pair groups corresponding to one or more lower thresholds. Typically, for each successively lower threshold group, the same or a lower percentage of codon pairs are replaced. In one example, all codon pairs above a highest threshold are replaced, while a codon pair above an intermediate threshold is replaced only if the codon pair is located within an autonomous folding unit. In another example, all codon pairs above a highest threshold are replaced, while a codon pair above an intermediate threshold is replaced only if the codon pair can be replaced without requiring a change in the encoded polypeptide sequence. In another example, all codon pairs above a highest threshold are replaced, while a codon pair above a first higher intermediate threshold is replaced only if the codon pair can be replaced without changing the encoded polypeptide sequence or with only a conservative change to the encoded polypeptide sequence, while a codon pair above a second lower intermediate threshold is replaced only if the codon pair can be replaced without requiring any change in the encoded polypeptide sequence. While the above discussion has been applied to the use of a plurality of threshold levels, it will be readily apparent to one skilled in the art that, in the place of using threshold levels, an evaluation method can be used that determines the degree to which a codon pair should
be replaced according to the translational kinetics value of the codon pair, where the
degree to which the codon pair should be replaced can be counterbalanced by any of a
variety of user-determined factors such as, for example, presence of the codon pair within
or between autonomous folding units, and degree of change to the encoded polypeptide
sequence.

[0066] In accordance with the methods and sequences provided herein, a
translational kinetics value of a codon pair is a representation of the degree to which it is
expected that a codon pair is associated with a translational pause. Methods of
determining the translational kinetics value of a codon pair are discussed elsewhere
herein. Such translational kinetics values can be normalized to facilitate comparison of
translational kinetics values between species. In some embodiments, the translational
value can be the degree of over-representation of a codon pair. An over-represented
codon pair is a codon pair which is present in a protein-encoding sequence in higher
abundance than would be expected if all codon pairs were statistically randomly
abundant. When translational kinetics values of codon pairs are determined, a codon pair
predicted to cause a translational pause or slowing is a codon pair whose likelihood of
causing a translational pause or slowing is at least one standard deviation above the mean
translational kinetics value, where a particular translational kinetics value above the mean
translational kinetics value in this context refers to a translational kinetics value indicative
of a greater likelihood of causing translational pausing or slowing, relative to a mean
translational kinetics value, and is not strictly limited to a particular mathematical
relationship (e.g., greater than the mean) since the depiction of propensity to cause a
translational pause by a translational kinetics value can be selected to be negative or
positive, based on the selected implementation by one skilled in the art. For example,
over-represented codon pairs may be graphically displayed as a positive function in a
SpeedPlot™, as depicted in Figure 1, where a positive deflection or peak above a selected
threshold describes a translational pause or slowing at the exact nucleotide location as
defined by the abscissa. In the methods provided herein, a threshold for the translational
kinetics value of codon pairs that are predicted to cause a translational pause or slowing
can be set in accordance with the method and level of stringency desired by one skilled in
the art. For example, when it is desired to identify only a small number of the codon pairs
most likely to cause a translational pause or slowing, a threshold value can be set to 5, or
3, or 2, or 1.5 standard deviations or more above the mean. Typical threshold values can
be at least 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 3, 3.5, 4, 4.5 and 5 or more standard deviations

-25-
above the mean. As provided herein, a plurality of thresholds can be applied in the herein-provided methods in segregating codon pairs into a plurality of groups. Each threshold of such a plurality can be a different value selected from 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 3, 3.5, 4, 4.5 and 5 or more standard deviations above the mean.

[0067] In some embodiments, translational kinetics of an mRNA into polypeptide can be changed to add or retain one or more translational pauses predicted to occur before, after or within an autonomous folding unit of a protein, or between autonomous folding units. While not intending to be limited to the following, it is proposed that translational pauses are present in wild type genes in order to slow translation of a nascent polypeptide subsequent to translation of a protein domain, thus providing time for acquisition of secondary and at least partial tertiary structure in the domain prior to further downstream translation and reorganization or reconfiguration of the growing polypeptide or domain. By modifying the translational kinetics of complex multi-domain proteins it may be possible to experimentally alter the time each domain has available to organize. Folding of a heterologously-expressed gene having two or more independent domains can be altered by the presence of pause sites between the domains. Refolding studies indicate that the time it takes for a protein to settle into its final configuration may take longer than the translation of the protein. Pausing may allow each domain to partially organize and commit to a particular, independent fold. Other co-translational events, such as those associated with co-factors, protein subunits, protein complexes, membranes, chaperones, secretion, or proteolysis complexes, also can depend on the kinetics of the emerging nascent polypeptide. Pauses can be introduced by engineering one codon pair predicted to cause a translational pause or slowing, or two or more such codon pairs into the sequence to facilitate these co-translational interactions.

[0068] As such, provided herein is the recognition that the presence of codon pairs predicted to cause a translational pause or slowing in protein-encoding regions separating regions encoding different autonomous folding units of the protein can serve to pause translation and facilitate folding of the nascent translated protein, where autonomous folding units can be secondary structural elements such as an alpha helix, or can be tertiary structural elements such as a protein domain. Accordingly, provided herein are methods of changing translational kinetics of an mRNA into polypeptide by including or preserving one or more translational pauses predicted to occur before, after, or between autonomous folding units of a protein, thereby increasing the likelihood that the translated protein will be properly folded. In such embodiments, typically a
translational pause is preserved, which refers to maintaining the same codon pair for a polypeptide-encoding nucleotide sequence that is expressed in the native host organism, or, when the polypeptide-encoding nucleotide sequence is heterologously expressed, changing the codon pair as appropriate to have a translational kinetics value comparable to or closest to the translational kinetics value of the native codon pair in the native host organism.

[0069] In some instances, it is not possible to modify the polypeptide-encoding nucleotide sequence to preserve or insert a translational pause without causing a change to the encoded amino acid sequence. For example, there may be no codon pairs that are predicted to cause a translational pause or slowing and that encode the same pair of amino acids as encoded in the original sequence. In such instances, several options are available. First, proximal codon pairs can be selected to be replaced in order to introduce a translational pause or slowing. For example, one of the 1, 2, 3, 4 or 5 most proximal codon pairs upstream (5' of the desired pause site) or one of the 1, 2, 3, 4 or 5 most proximal codon pairs downstream (3' of the desired pause site) can be chosen for replacement to introduce the translational pause or slowing. Typically in such instances, the selected codon pair for replacement to introduce the translational pause or slowing is the codon pair closest to the originally desired codon pair location of the translational pause or slowing, provided the desired translational pause or slowing can be attained (e.g., 1 codon pair upstream or downstream is typically selected instead of 2 codon pairs upstream or downstream, provided the desired translational pause or slowing can be attained). Alternatively, a translational pause or slowing can be introduced by selecting a replacement codon pair encoding a conservative amino acid substitution, such as the conservative substitutions shown in Table 1. In some embodiments, replacement of a proximal codon pair to introduce a translational pause or slowing is preferred over replacement of a codon pair resulting in a change in the encoded amino acid sequence.


[0071] Further, provided herein is the recognition that predicted pause sites may be conserved across different proteins in the same species, or in related proteins across two or more species. In some embodiments, graphical displays of translational
kinetics values of one or more proteins can be used to provide information to assist in the
selection of a translational pause or slowing to preserve or insert in a redesigned polypeptide-encoding nucleotide sequence. In particular, graphical displays of translational kinetics values can permit, for example, alignment of homologous proteins from different species and an identification, based on this alignment, of predicted translational pause or slowing sites that are conserved in the aligned proteins. Such predicted translational pause or slowing sites can be preserved or inserted in a redesigned polypeptide-encoding nucleotide sequence. In another example, regions between autonomous folding units in one or more proteins within a particular species can be graphically examined for the presence or absence of predicted pause sites. Such graphical display methods can result in an identification of a region between autonomous folding units in which a translational pause or slowing is desirably preserved in a redesigned polypeptide-encoding sequence.

[0072] Methods for identifying and selecting conserved translational pauses can be performed according to any method known in the art, as exemplified in U.S. Patent Publication No. 2007/0298503, published on December 27, 2007, and U.S. Patent Publication No. 2007/0275399, published on November 29, 2007. For example, the codon pair translation kinetics values can be compared with a database of related gene sequences and conserved pause sites can be identified. Additionally, a synthetic gene can be designed wherein at least one conserved pause site is maintained to provide a synthetic gene with modified translation kinetics.

Redesign of polypeptide-encoding nucleotide sequence

[0073] As provided herein, codon pairs are associated with translational pauses, and can thereby influence translational kinetics of an mRNA into polypeptide. Thus, the methods of changing translational kinetics provided herein will typically be performed by modifying or designing one or more nucleotide sequences encoding a polypeptide to be expressed. Accordingly, provided herein are methods of modifying a gene or designing a synthetic nucleotide sequence encoding the polypeptide encoded by the gene, collectively referred to herein as redesigning a polypeptide-encoding gene sequence or redesigning a polypeptide-encoding nucleotide sequence. Also included in the various embodiments provided herein are redesigned gene sequences encoding polypeptides that are not identical to the original gene.
In some embodiments are provided a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 50%, 60%, 70%, 75%, 80%, 85%, and more typically at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to the wild type cellobiohydrolase polypeptide sequence as set forth in SEQ ID NO: 1.

In certain embodiments, at least 1, 2 or 3 codon pairs of a polynucleotide sequence encoding the cellobiohydrolase (SEQ ID NO: 2) have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 1, 2 or 3 codon pairs to be replaced are selected from the following: CCCTCT (nucleotides 463-468); GGCCAA (nucleotides 94-99); CAGTTT (nucleotides 565-570); GATATC (nucleotides 703-708); GTGGAA (nucleotides 691-696); GGATTT (nucleotides 1192-1197); GGTATT (nucleotides 1198-1203), or any other codon pair that can suitably be substituted. In some such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

In some aspects of the above embodiments, at least 1, 2, or 3 of the following codon pair replacements have been made: CCCTCT (nucleotides 463-468) replaced with CCTTCT; GGCCAA (nucleotides 94-99) replaced with GGTCAA; CAGTTT (nucleotides 565-570) replaced with CAATTT; GATATC (nucleotides 703-708) replaced with GACATT; GTGGAA (nucleotides 691-696) replaced with GTTGAA; GGATTT (nucleotides 1192-1197) replaced with GGTTC; GGTATT (nucleotides 1198-1203) replaced with GGAATT or any other codon pair replacement that can suitably be made.

In certain embodiments, at least 1, 2 or 3 codon pairs of a polynucleotide sequence encoding the cellobiohydrolase (SEQ ID NO: 2) have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 1, 2 or 3 codon pairs to be replaced are selected from the following: CTCGGT (nucleotides 760-765); ATTGCC (nucleotides 631-636); GACAGC (nucleotides 1285-1290); GTCTGG (nucleotides 88-93); GTCTGG (nucleotides 1246-1251); TTGCTG (nucleotides 1231-1236); GTGGTG (nucleotides 571-576); ACGCTG (nucleotides 22-27); ACGCTG (nucleotides 31-36); GACTGG (nucleotides 1168-1173); GCCGGA (nucleotides 559-564); CTGGTG (nucleotides 748-753), or any other codon pair that can suitably be substituted. In some
such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

[0078] In some aspects of the above embodiments, at least 1, 2, or 3 of the following codon pair replacements have been made: CTCGGT (nucleotides 760-765) replaced with CTGGGT; ATTGCC (nucleotides 631-636) replaced with ATTGCG; GACAGC (nucleotides 1285-1290) replaced with GACTCT; GTCTGG (nucleotides 88-93) replaced with GTTTGG; GTCTGG (nucleotides 1246-1251) replaced with GTTTGG; TTGCTG (nucleotides 1231-1236) replaced with CTGCTG; GTGGTG (nucleotides 571-576) replaced with GTTGTG; ACGCTG (nucleotides 22-27) replaced with ACCCTC; ACGCTG (nucleotides 31-36) replaced with ACCCTG; GACTGG (nucleotides 1168-1173) replaced with GATTTG; GCCGGA (nucleotides 559-564) replaced with GCGGGC; CTGGTG (nucleotides 748-753) replaced with CTGGTT or any other codon pair replacement that can suitably be made.

[0079] In certain embodiments, at least 1, 2 or 3 codon pairs of a polynucleotide sequence encoding the cellulbiohydrolase (SEQ ID NO: 2) have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 1, 2 or 3 codon pairs to be replaced are selected from the following: CAGTTT (nucleotides 565-570); TTTGAC (nucleotides 1303-1308); TCGTTT (nucleotides 1240-1245); GGCCAA (nucleotides 94-99); AAGAAT (nucleotides 541-546); AAGAAT (nucleotides 934-939); GCCAAA (nucleotides 649-654); GTCAAG (nucleotides 1252-1257); GGTATT (nucleotides 1198-1203); ATCAAC (nucleotides 808-813); GGCCAT (nucleotides 865-870); CTTCCA (nucleotides 835-840); GATATC (nucleotides 703-708); TCGTTG (nucleotides 1228-1233), or any other codon pair that can suitably be substituted. In some such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

[0080] In some aspects of the above embodiments, at least 1, 2, or 3 of the following codon pair replacements have been made: CAGTTT (nucleotides 565-570) replaced with CAATT; TTTGAC (nucleotides 1303-1308) replaced with TTTGAT; TCGTTT (nucleotides 1240-1245) replaced with TCTTTT; GGCCAA (nucleotides 94-99) replaced with GGACAA; AAGAAT (nucleotides 541-546) replaced with AAAAAT; AAGAAT (nucleotides 934-939) replaced with AAAAAC; GCCAAA (nucleotides 649-
654) replaced with GCTAAA; GTCAAG (nucleotides 1252-1257) replaced with GTTAAA; GGTATT (nucleotides 1198-1203) replaced with GGAATC; ATCAAC (nucleotides 808-813) replaced with ATTAAT; GCCCAT (nucleotides 865-870) replaced with GGACAC; CTTCCA (nucleotides 835-840) replaced with TTG CCT; GATATC (nucleotides 703-708) replaced with GATATA; TCGTTG (nucleotides 1228-1233) replaced with TCGTTT or any other codon pair replacement that can suitably be made.

[0081] In certain embodiments, at least 1, 2 or 3 codon pairs of a polynucleotide sequence encoding the cellulbiohydrolase (SEQ ID NO: 2) have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 1, 2 or 3 codon pairs to be replaced are selected from the following: GGCCCA (nucleotides 94-99); CAGTTT (nucleotides 565-570); GATATC (nucleotides 703-708); TATTTT (nucleotides 853-858); GGCCAT (nucleotides 865-870); TCGTTG (nucleotides 1228-1233); TTTGTC (nucleotides 1243-1248); TTCCCA (nucleotides 1363-1368), or any other codon pair that can suitably be substituted. In some such polynucleotides, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

[0082] In some aspects of the above embodiments, at least 1, 2, or 3 of the following codon pair replacements have been made: GGCCCA (nucleotides 94-99) replaced with GGTCGA; CAGTTT (nucleotides 565-570) replaced with CAATTC; GATATC (nucleotides 703-708) replaced with GACATT; TATTTT (nucleotides 853-858) replaced with TATTTA; GGCCAT (nucleotides 865-870) replaced with GGACAT; TCGTTG (nucleotides 1228-1233) replaced with TCTTTA; TTTGTC (nucleotides 1243-1248) replaced with TTCGTT; TTCCCA (nucleotides 1363-1368) replaced with TTCCAG or any other codon pair replacement that can suitably be made.

[0083] In certain embodiments, at least 1, 2 or 3 codon pairs of a polynucleotide sequence encoding the cellulbiohydrolase (SEQ ID NO: 2) have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. In some embodiments, the at least 1, 2 or 3 codon pairs to be replaced are selected from the following: GTG CCT (nucleotides 55-60); GCCCAAT (nucleotides 370-375); GCTATT (nucleotides 406-411); GCCGGA (nucleotides 559-564); GCCCAAT (nucleotides 778-783); TTGGCA (nucleotides 967-972); AAGCTG (nucleotides 1051-1056); GCTATT (nucleotides 1066-1071); GCCAAT (nucleotides 1084-1089); ACCGGA (nucleotides 1147-1152); ACCGGA (nucleotides 1189-1194);
GGTATT (nucleotides 1198-1203); GACAGC (nucleotides 1285-1290); GATGCC
(nucleotides 1327-1332); GCCTTG (nucleotides 1330-1335); CAGCTT (nucleotides
1381-1386), or any other codon pair that can suitably be substituted. In some such
gene sequences, at least 3, or 4, or 5, or 6 or more of the specified codon pairs have
been replaced with different codon pairs encoding identical amino acids or conservative
amino acid substitutions thereof.

[0084] In some aspects of the above embodiments, at least 1, 2, or 3 of the
following codon pair replacements have been made: GGTATT (nucleotides 1198-1203)
replaced with GTGCCT (nucleotides 55-60) replaced with GTTCCG; GCCAAT
(nucleotides 370-375) replaced with GCTAAC; GCTATT (nucleotides 406-411) replaced
with GCCATT; GCCCGA (nucleotides 559-564) replaced with GCTGGA; GCCAAT
(nucleotides 778-783) replaced with GCAGAT; TTGGCA (nucleotides 967-972)
replaced with TTGGCT; AAGCTG (nucleotides 1051-1056) replaced with AAATTG;
GCTATT (nucleotides 1066-1071) replaced with GCCATT; GCCAAT (nucleotides
1084-1089) replaced with GCTAAC; ACCGGA (nucleotides 1147-1152) replaced with
ACCGGT; ACCGGA (nucleotides 1189-1194) replaced with ACAGGT; GGTATT
(nucleotides 1198-1203) replaced with GGAATC; GACAGC (nucleotides 1285-1290)
replaced with GATTCT; GATGCC (nucleotides 1327-1332) replaced with GACGCC;
GCCTTG (nucleotides 1330-1335) replaced with GCCCTT; CAGCTT (nucleotides
1381-1386) replaced with CAGTTG or any other codon pair replacement that can
suitably be made.

[0085] In some embodiments, provided is a cellubiohydrolase-encoding DNA
sequence adapted for expression in a heterologous host organism, wherein at least 1, 2, 3,
4, 5, 6, 7, 8, 9, 10, or more codon pairs present in wild-type nucleotide sequence and
which encode the cellulose binding domain of the cellubiohydrolase, have been replaced
with different codon pairs encoding identical amino acids or conservative amino acid
substitutions thereof. The conserved amino acid sequence pattern and domain boundaries
for cellulose binding domains are known in the art. In the case of the cellubiohydrolase
of SEQ ID NO: 2, the cellulose binding domain includes at least amino acids 35-58, 30-61
or 27-62.

[0086] Typically in such embodiments, the replacement codon pairs are
predicted to be less likely to cause a translational pause in the heterologous host organism
relative to the respective wild type codon pair when expressed in the heterologous host
organism. That is, the embodiments in which one or more codon pairs encoding amino
acids of the cellulose binding domain have been replaced include embodiments in which
the nucleotide sequence encoding the cellulose binding domain is changed to increase the
predicted translational kinetics of translation of the cellulose binding domain. As
provided herein, incomplete translation, improper folding, or other protein expression
shortcomings can result from the presence of one or more translational pauses in a
heterologously-expressed polypeptide. In some embodiments, removal of one or more of
these pauses can increase the speed of translation of the cellulose binding domain, and
thereby increase the quantity of protein produced and/or increase the amount of stable,
properly folded, active, and/or soluble protein produced.

[0087] In such embodiments, the replacement codons, i.e., the codons added
as replacements for the wild type codons, are typically predicted to be less likely to cause
a translational pause. For example, the replacement codon can have a translational
kinetics value in the heterologous host organism that is 95%, 90%, 85%, 80%, 75%, 70%,
or less, than the translational kinetics value of the wild type codon pair when expressed in
the heterologous host organism. In some embodiments, the replacement codon is selected
to have a translational kinetics value similar to the translational kinetics value of the wild
type codon pair in the native organism. For example, the z score of at least one
replacement codon pair when expressed in the heterologous host organism can be no
more than 250%, 200%, 150%, 125% or 100% of the z score for the wild type codon pair
when expressed in the native organism.

[0088] In some embodiments, provided is a cellobiohydrolase-encoding DNA
sequence adapted for expression in a heterologous host organism, wherein at least 1, 2, 3,
4, 5, 6, 7, 8, 9, 10, or more codon pairs present in wild-type nucleotide sequence and
which encode the glycosyl hydrolase domain of the cellobiohydrolase, have been replaced
with different codon pairs encoding identical amino acids or conservative amino acid
substitutions thereof. The conserved amino acid sequence pattern and domain boundaries
for glycosyl hydrolase domains are known in the art. In the case of the cellobiohydrolase
of SEQ ID NO: 2, the glycosyl hydrolase domain includes at least amino acids 124-437,
115-450 or 107-471.

[0089] Typically in such embodiments, the replacement codon pairs are
predicted to be less likely to cause a translational pause in the heterologous host organism
relative to the respective wild type codon pair when expressed in the heterologous host
organism. That is, the embodiments in which one or more codon pairs encoding amino
acids of the glycosyl hydrolase domain have been replaced include embodiments in which
the nucleotide sequence encoding the glycosyl hydrolase domain is changed to increase the predicted translational kinetics of translation of the glycosyl hydrolase domain. As provided herein, incomplete translation, improper folding, or other protein expression shortcomings can result from the presence of one or more translational pauses in a heterologously-expressed polypeptide. In some embodiments, removal one or more of these pauses can increase the speed of translation of the glycosyl hydrolase domain, and thereby increase the quantity of protein produced and/or increase the amount of stable, properly folded, active, and/or soluble protein produced.

[0090] In such embodiments, the replacement codons, i.e., the codons added as replacements for the wild type codons, are typically predicted to be less likely to cause a translational pause. For example, the replacement codon can have a translational kinetics value in the heterologous host organism that is 95%, 90%, 85%, 80%, 75%, 70%, or less, than the translational kinetics value of the wild type codon pair when expressed in the heterologous host organism. In some embodiments, the replacement codon is selected to have a translational kinetics value similar to the translational kinetics value of the wild type codon pair in the native organism. For example, the z score of at least one replacement codon pair when expressed in the heterologous host organism can be no more than 250%, 200%, 150%, 125% or 100% of the z score for the wild type codon pair when expressed in the native organism.

[0091] In some embodiments, provided is a cellobiohydrolase-encoding DNA sequence adapted for expression in a heterologous host organism, wherein at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more codon pairs present in wild-type nucleotide sequence and which encode the region between the cellulose binding domain and the glycosyl hydrolase domain of the cellobiohydrolase, have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof. The conserved amino acid sequence pattern and domain boundaries for the cellulose binding domain and glycosyl hydrolase domain are described hereinafter.

[0092] Typically in such embodiments, the replacement codon pairs are predicted to be more likely to cause a translational pause in the heterologous host organism relative to the respective wild type codon pair when expressed in the heterologous host organism. That is, the embodiments in which one or more codon pairs encoding amino acids in the region between the cellulose binding domain and the glycosyl hydrolase domain have been replaced include embodiments in which the nucleotide sequence encoding the region between the cellulose binding domain and the
glycosyl hydrolase domain is changed to decrease the predicted translational kinetics of translation of the region between the cellulose binding domain and the glycosyl hydrolase domain. As provided herein, incomplete translation, improper folding, or other protein expression shortcomings can result from the absence of one or more translational pauses in a heterologously-expressed polypeptide. In some embodiments, adding one or more of these pauses can increase the speed of translation of the glycosyl hydrolase domain, and thereby increase the quantity of protein produced and/or increase the amount of stable, properly folded, active, and/or soluble protein produced.

[0093] In such embodiments, the replacement codons, i.e., the codons added as replacements for the wild type codons, are typically predicted to be more likely to cause a translational pause. For example, the replacement codon can have a translational kinetics value in the heterologous host organism that is 105%, 110%, 115%, 120%, 125%, 130%, or more, than the translational kinetics value of the wild type codon pair when expressed in the heterologous host organism. In some embodiments, the replacement codon is selected to have a translational kinetics value similar to the translational kinetics value of the wild type codon pair in the native organism. For example, the z score of at least one replacement codon pair when expressed in the heterologous host organism can be at least 75%, 80%, 85%, 90%, 95% or 100% of the z score for the wild type codon pair when expressed in the native organism.

[0094] Thus, provided herein are methods for redesigning the polypeptide-encoding nucleotide sequence provided herein to modify the translational kinetics of the polypeptide-encoding nucleotide sequence, where the polypeptide-encoding nucleotide sequence is altered such that one or more codon pairs have a decreased likelihood of causing a translational pause or slowing relative to the unaltered polypeptide-encoding nucleotide sequence. For example, one or more nucleotides of a polypeptide-encoding nucleotide sequence can be changed such that a codon pair containing the changed nucleotides has a translational kinetics value indicative of a decreased likelihood of causing a translational pause or slowing relative to the unchanged polypeptide-encoding nucleotide sequence.

[0095] While it will be understood by those of skill in the art that a redesigned polypeptide-encoding nucleotide sequence need not possess a high degree of identity to the polypeptide-encoding nucleotide sequence of the original gene, in some embodiments, the redesigned polypeptide-encoding nucleotide sequence will have at least 50%, 60%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99%
nucleotide identity with the polypeptide-encoding nucleotide sequence of the original gene. As used herein an original gene refers to a gene for which codon pair refinement is to be performed; such original genes can be, for example, wild type genes, native genes, naturally occurring mutant genes, other mutant genes such as site-directed mutant genes or engineered or completely synthetic genes. In other embodiments, the polynucleotide sequence will be completely synthetic, and will bear much lower identity with the original gene, e.g., no more than 90%, 80%, 70%, 60%, 50%, 40%, or lower.

[0096] Because of the redundancy of the triplet genetic code it is possible to preserve amino acid sequence coding while redesigning the polypeptide-encoding gene nucleotide sequence. Polypeptide-encoding nucleotide sequences can be redesigned to be convenient to work with and specifically tailored to a particular host and vector system of choice. The resulting sequence can be designed to: (1) reduce or eliminate translational problems caused by inappropriate ribosome pausing, such as those caused by over-represented codon pairs or other codon pairs with translational values predictive of a translational pause; (2) have codon usage refined to avoid over-reliance on rare codons; (3) reduce in number or remove particular restriction sites, splice sites, internal Shine-Dalgarno sequences, or other sites that may cause problems in cloning or in interactions with the host organism; or (4) have controlled RNA secondary structure to avoid detrimental translational termination effects, translation initiation effects, or RNA processing, which can arise from, for example, RNA self-hybridization. When a synthetic polypeptide-encoding nucleotide sequence is to be used, this sequence also can be designed to avoid oligonucleotides that mis-hybridize, resulting in genes that can be assembled from refined oligonucleotides that by thermodynamic necessity only pair up in the desired manner, using methods known in the art, as exemplified in U.S. Patent Publication No. 2005/0106590, which is hereby incorporated by reference in its entirety.

[0097] In some instances, it is not possible to modify the polypeptide-encoding nucleotide sequence to suitably modify the translational kinetics of the mRNA into polypeptide without modifying the amino acid sequence of the encoded polypeptide. In such instances, an amino acid insertion, deletion or mutation can be introduced to yield a codon pair that is not predicted to cause a translational pause or slowing; or no change is made. In methods in which an amino acid insertion, deletion or mutation is made in order to change translational kinetics, the change is preferably predicted to not substantially influence the final three-dimensional structure of the protein and/or the activity of the protein. Such non-identical polypeptides can vary by containing one or
more insertions, deletions and/or mutations. Although the nature and degree of change to the polypeptide sequence can vary according to the purpose of the change, typically such a change results in a polypeptide that is at least 50%, 60%, 70%, 75%, 80%, 85%, and more typically at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identical to the wild type polypeptide sequence.

[0098] In some embodiments, the sequence of the polynucleotide can be generated, optionally in conjunction with optimization of a plurality of parameters where one such parameter can be codon pair usage, where the resultant polynucleotide can be prepared by assembly of a plurality of oligonucleotides sufficiently small to be synthesized by known oligonucleotide synthetic methods. Methods known in the art for optimizing multiple parameters in synthetic nucleotide sequences can be applied to optimizing the parameters recited in the present claims. Such methods may advantageously include those exemplified in U.S. Patent App. Publication No. 2005/0106590, U.S. Patent App. Publication No. 2007/0009928, and R. H. Lathrop et al. "Multi-Queue Branch-and-Bound Algorithm for Anytime Optimal Search with Biological Applications" in Proc. Intl. Conf. on Genome Informatics, Tokyo, Dec. 17-19, 2001 pp. 73-82; in Genome Informatics 2001 (Genome Informatics Series No. 12), Universal Academy Press, which are incorporated herein by reference in their entireties. Briefly, in addition to optimizing the various parameters, an exemplary method for generating a sequence can also include dividing the desired sequence into a plurality of partially overlapping segments; optimizing the melting temperatures of the overlapping regions of each segment to disfavor hybridization to the overlapping segments which are non-adjacent in the desired sequence; allowing the overlapping regions of single stranded segments which are adjacent to one another in the desired sequence to hybridize to one another under conditions which disfavor hybridization of non-adjacent segments; and filling in, ligating, or repairing the gaps between the overlapping regions, thereby forming a double-stranded DNA with the desired sequence. This process can be performed manually or can be automated, e.g., in a general purpose digital computer. In one embodiment, the search of possible codon assignments is mapped into an anytime branch and bound computerized algorithm developed for biological applications.

[0099] Accordingly, provided herein are methods of designing a synthetic nucleotide sequence for the polynucleotides provided herein, where the synthetic nucleotide sequence also is typically designed to have desirable translational kinetics properties, such as the removal of some or all codon pairs predicted to result in a
translational pause or slowing. Such design methods include determining a set of partially overlapping segments with optimized melting temperatures, and determining the translational kinetics of the synthetic sequence, where if it is desired to change the translational kinetics of the synthetic gene, the sequences of the overlapping segments are modified and refined in order to approximate the desired translational kinetics while still possessing acceptable hybridization properties. In some embodiments, this process is performed iteratively. In some embodiments, a criterion is established for selecting codon pairs having high translational kinetics values to be replaced with codon pairs having lower the translational kinetics values unless a codon pair of this group is the site of a planned pause. For example, the top 1%, 1.5%, 2%, 2.5%, 3%, 3.5%, 4%, 4.5%, 5%, 5.5%, 6%, 6.5%, 7%, 7.5%, 8%, 8.5%, 9%, 9.5%, or 10% of codon pairs ranked by translational kinetics values can be replaced by codon pairs having lower translational kinetics values, such as translational kinetics value below a user defined level that can be, for example, a translational kinetics value equal to or below the translational kinetics values of codon pairs not in the top selected percentage, unless a codon pair of this group is the site of a planned pause (in which case it is not necessarily replaced). In another example all codon pairs above a user-selected translational kinetics value, such as more than 5, 4.5, 4, 3.5, 3, 2.5 or 2 standard deviations above the mean translational kinetics value can be replaced by codon pairs having lower translational kinetics values, such as translational kinetics value below a user defined level that can be, for example, a translational kinetics value that is 4, 3.5, 3, 2.5, 2, 1.5 or 1 standard deviations less than the mean translational kinetics value, unless a codon pair of this group is the site of a planned pause (in which case it is not necessarily replaced). Further synthetic nucleotide sequence refinement methods can be employed where additional properties of the synthetic nucleotide sequence can be refined in addition to hybridization and codon pair usage properties, where such properties can include, for example, codon usage, reduced number of restriction sites or Shine-Dalgarno sequences, or reduced detrimental RNA secondary structure, as described above.

[0100] Those skilled in the art will recognize that various optimization methods can be used, e.g., simulated annealing, genetic algorithms, branch and bound techniques, hill-climbing, Monte Carlo methods, other search strategies, and the like. Thus, the methods provided herein for designing the polynucleotide sequences provided herein, that include optimization of a plurality of parameters, where one such parameter is codon pair usage, can be implemented in by applying those parameters to art-recognized
algorithms or techniques. Advantageously, sequence design is performed using an optimization method that designs a synthetic nucleotide sequence encoding the polypeptide to be expressed.

[0101] The polynucleotide sequences design methods provided herein can be employed where a plurality of properties of the polynucleotide sequences can be refined in addition to codon pair usage properties, where such properties can include, but are not limited to, melting temperature gap between oligonucleotides of synthetic gene, average codon usage, average codon pair chi-squared (e.g., z score), worst codon usage, worst codon pair (e.g., z score), maximum usage in adjacent codons, Shine-Dalgarno sequence (for E. coli expression), occurrences of 5 consecutive G's or 5 consecutive Cs, occurrences of 6 consecutive A's or 6 consecutive Ts, long exactly repeated subsequences, cloning restriction sites, user-prohibited sequences (e.g., other restriction sites), codon usage of a specific codon above user-specified limit, and out-of-frame stop codons (framecatchers). In embodiments that include expression in a eukaryotic host organism, additional properties that can be considered in a process of designing a polynucleotide sequence include, but are not limited to, occurrences of RNA splice sites, occurrences of polyA sites, and occurrence of ribosome binding sequence. For example, a process of designing a poly nucleotide sequence can include constraints including, but not limited to, minimum melting temperature gap between oligonucleotides of synthetic gene, minimum average codon usage, maximum average codon pair chi-squared (z score), minimum absolute codon usage, maximum absolute codon pair (z score), minimum maximum usage in adjacent codons, no Shine-Dalgarno sequence (for E. coli expression), no occurrences of 5 consecutive G's or 5 consecutive Cs, no occurrences of 6 consecutive A's or 6 consecutive T's no long exactly repeated subsequences, no cloning restriction sites, no user-prohibited sequences (e.g., other restriction sites), and optionally no codon usage of a specific codon above user-specified limit. In embodiments that include expression in a eukaryotic host organism, additional constraints can include, but are not limited to, minimum occurrences of RNA splice sites, minimum occurrences of polyA sites, and occurrence of ribosome binding sequence. A process of designing a polynucleotide sequence can include preferences including, but not limited to, prefer high average codon usage, prefer low average codon pair chi-squared, prefer larger melting temperature gap, prefer more out of frame stop codons (framecatchers), and optionally prefer evenly distributed codon usage. Any of a variety of nucleotide sequence refinement/optimization methods known in the art can be used to refine the
polynucleotide sequence according to the codon pair usage properties, and according to any of the additional properties specifically described above, or other properties that are refined in nucleotide sequence redesign methods known in the art. In some embodiments, a branch and bound method is employed to refine the polynucleotide sequence according to codon pair usage properties and at least one additional property, such as codon usage.

[0102] In some embodiments, the methods provided herein can further include analyzing at least a portion of the candidate polynucleotide sequence in frame shift, and selecting codons for the candidate polynucleotide sequence such that stop codons are added to at least one said frame shift. In additional embodiments, the generating step further includes analyzing at least a portion of the candidate polynucleotide sequence in frame shift, and selecting codons for the candidate polynucleotide sequence such that one or more stop codons in one, two or three reading frames are added downstream of polypeptide-encoding region of the nucleotide sequence.

[0103] In some embodiments, methods are provided for redesigning a polypeptide-encoding gene for expression in a host organism, by providing a data set representative of codon pair translational kinetics for the host organism which includes translational kinetics values of the codon pairs utilized by the host organism, providing a desired polypeptide sequence for expression in the host organism, and generating a polynucleotide sequence encoding the polypeptide sequence by analyzing candidate nucleotides to select, where possible, codon pairs that are predicted not to cause a translational pause in the host organism, with reference to the data set, thereby providing a candidate polynucleotide sequence encoding the desired polypeptide.

[0104] Also provided herein are methods for redesigning a polypeptide-encoding gene for expression in a host organism, by providing a first data set representative of codon pair translational kinetics for the host organism which includes translational kinetics values of the codon pairs utilized by the host organism, providing a second data set representative of at least one additional desired property of the synthetic gene, providing a desired polypeptide sequence for expression in the host organism, and generating a polynucleotide sequence encoding the polypeptide sequence by analyzing candidate nucleotides to select, where possible, both (i) codon pairs that are predicted not to cause a translational pause in the host organism, with reference to the first data set, and (ii) nucleotides that provide a desired property, with reference to the second data set, thereby providing a candidate polynucleotide sequence encoding the desired polypeptide. In some embodiments, a branch and bound method is employed to refine the polypeptide-
encoding nucleotide sequence according to codon pair usage properties of the first data set and according to the properties of the second data set. In some embodiments, the second data set contains codon preferences representative of codon usage by the host organism, including the most common codons used by the host organism for a given amino acid.

[0105] Accordingly, provided herein is a cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 50%, 60%, 70%, 75%, 80%, 85%, and more typically at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to the wild type cellobiohydrolase polypeptide sequence as set forth in SEQ ID NO: 2. In certain aspects of the above embodiments, the polynucleotide provided herein is adapted for expression in a heterologous host organism. A heterologous host organism is an organism used to express DNA, RNA or protein that is foreign to the host organism. In certain aspects, the host organism is not human, *E. coli* or *S. cerevisiae*.

[0106] In certain aspects of the above embodiments, at least 1, 2 or 3 codon pairs of the original sequence that are predicted to cause a translational pause in the host organism have been replaced with codon pairs that are predicted to be less likely to cause a translational pause therein. In selected embodiments, the at least three codon pairs of the original sequence that are predicted to cause a translational pause in the host organism are highly-overrepresented codon pairs therein and have been replaced with codon pairs that are not highly-overrepresented therein. As described further below, a highly-overrepresented codon pair is a codon pair that has a translational kinetics value greater than a designated threshold, wherein a threshold value can be at least 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 3, 3.5, 4, 4.5 or 5 or more standard deviations above the mean translational kinetics value.

[0107] Also provided herein is a cellobiohydrolase-encoding DNA sequence, having at least a 75% sequence identity with an original cellobiohydrolase polypeptide sequence as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least three codon pairs of the original sequence that are predicted to cause a translational pause in the host organism have been replaced with codon pairs that are predicted to be less likely to cause a translational pause therein, and wherein the host organisms are selected from the following: *Pichia pastoris; Orzytolagus cuniculus* (rabbit); *Macaca fascicularis* (Long-tailed monkey); *M. mulatta* (Monkey); *E. coli* K12 W3110; *E. coli* UTI89; *E. coli* O157:H7 EDL933; *E. coli* O157:H7 str. Sakai;
Bombyx mori; Spodoptera frugiperda; Drosophila melanogaster, Kluyveromyces lactis, Zymomonas mobilis and Schizosaccharomyces pombe.

[0108] Thus, the methods provided herein can include analyzing the candidate polynucleotide sequence to confirm that no codon pairs are predicted to cause a translational pause in the host organism by more than a designated threshold. As described elsewhere herein, the likelihood that a particular codon pair will cause translational pausing or slowing in an organism (or the relative predicted magnitude thereof) can be represented by a translational kinetics value. The translational kinetics value can be expressed in any of a variety of manners in accordance with the guidance provided herein. In one example, a translational kinetics value can be expressed in terms of the mean translational kinetics value and the corresponding standard deviation for all codon pairs in an organism. For example, the translational kinetics value for a particular codon pair can be expressed in terms of the number of standard deviations that separate the translational kinetics value of the codon pair from the mean translational kinetics value. In methods that include analyzing the candidate polynucleotide sequence to confirm that no codon pairs are predicted to cause a translational pause in the host organism by more than a designated threshold, a threshold value can be at least 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 3, 3.5, 4, 4.5 or 5 or more standard deviations above the mean translational kinetics value. Although such a method is described in terms of a binary scoring of a codon pair as either at least or less than the threshold value, one skilled in the art, in view of the teachings herein, will recognize that multiple thresholds can be used, or methods can be used that weight a codon pair along a continuum according to the translational kinetics value, based on the teachings provided herein and the general knowledge in the art.

[0109] In some embodiments, in addition to generating a candidate nucleotide sequence according to codon pair usage properties, the methods provided herein also include generating a candidate nucleotide sequence according to codon usage. As is known in the art, different organisms can have different preference for the three-nucleotide codon sequence encoding a particular amino acid. As a result, translation can often be improved by using the most common three-nucleotide codon sequence encoding a particular amino acid. Thus, some methods provided herein also include generating a candidate nucleotide sequence such that codon utilization is non-randomly biased in favor of codons most commonly used by the host organism. Codon usage preferences are
known in the art for a variety of organisms and methods for selecting the more commonly used codons are well known in the art.

[0110] In some embodiments, the methods of redesigning a polypeptide-encoding nucleotide sequence are based on a plurality of properties, where a conflict in the preferred nucleotide sequence arising from the plurality of properties is determined in order to optimize the predicted translational kinetics. That is, when the plurality of properties being optimized would lead to more than one possible nucleotide sequence depending on which property is to be accorded more weight, typically, the conflict is resolved by selecting the nucleotide sequence predicted to be translated more rapidly, for example, due to fewer predicted translational pauses. In some embodiments, the methods of redesigning a polypeptide-encoding nucleotide sequence are based on a plurality of properties, where a conflict in the preferred nucleotide sequence arising from the plurality of properties is determined in order to optimize codon pair usage preferences. That is, when the plurality of properties being optimized would lead to more than one possible nucleotide sequence depending on which property is to be accorded more weight, typically, codon pair usage will be accorded more weight in order to resolve the conflict between the more than one possible nucleotide sequences. In one example, the methods provided herein can include identifying at least one instance of a conflict between selecting common codons and avoiding codon pairs predicted to cause a translational pause; in such instances, the conflict is resolved in favor of avoiding codon pairs predicted to cause a translational pause.

[0111] Some embodiments provided herein include generating a candidate polynucleotide sequence encoding the polypeptide sequence, the candidate polynucleotide sequence having a non-random codon pair usage, such that the codon pairs encoding any particular pair of amino acids have the lowest translational kinetics values. In some embodiments, the candidate polynucleotide sequence encoding the polypeptide sequence is generated and/or altered such that the encoded amino acid sequence is not altered. In some embodiments, the candidate polynucleotide sequence encoding the polypeptide sequence is generated and/or altered such that the three dimensional structure of the encoded polypeptide is not substantially altered. In some embodiments, the candidate polynucleotide sequence encoding the polypeptide sequence is generated and/or altered such that no more than conservative amino acid changes are made to the encoded polypeptide.
The methods provided herein can further include a step of refining or altering the candidate polynucleotide sequence in accordance with a second nucleotide sequence property to be refined. For example, in embodiments in which codon usage is also refined, the methods further include generating or refining a candidate polynucleotide sequence encoding a polypeptide sequence such that the candidate polynucleotide sequence has a non-random codon usage, where the most common codons used by the host organism are over-represented in the candidate polynucleotide sequence. The methods can include refining or altering the candidate polynucleotide sequence in accordance with any of a variety of additional properties provided herein, including but not limited to, melting temperature gap between oligonucleotides of synthetic gene, Shine-Dalgarno sequence, occurrences of 5 consecutive G’s or 5 consecutive Cs, occurrences of 6 consecutive A’s or 6 consecutive T’s long exactly repeated subsequences, cloning restriction sites, or any other user-prohibited sequences. Further, any of a variety of combinations of these properties can be additionally included in the nucleotide sequence refinement methods provided herein.

The method provided herein can further include an evaluation step in which after the candidate polynucleotide sequence is altered, the sequence is compared with at least a portion of a data set of a property against which the sequence was refined. In such methods, it is possible to compare the candidate sequence to the data set in order to determine whether or not the candidate sequence possesses the desired or acceptable properties with respect to the data set. For example, subsequent to a round of nucleotide sequence refinement, it can be evaluated whether or not the codon pairs of the candidate sequence have acceptable translational kinetics values. If the values are deemed to be acceptable or desired, no further sequence alteration is required with respect to the property. In view of the methods provided herein which can be directed to the refinement or optimization of a plurality of properties, the candidate nucleotide sequence can be compared to each property considered in the refinement, and, if the values for all properties are deemed to be acceptable or desired, no further sequence alteration is required. If the values for fewer than all properties are deemed to be acceptable or desired, the candidate nucleotide sequence can be subjected to further sequence alteration and evaluation.

Thus, it is contemplated herein that the sequence alteration steps of methods provided herein can be performed iteratively. That is, one or more steps of altering the nucleotide sequence can be performed, and the candidate nucleotide sequence
can be evaluated to determine whether or not further sequence alteration is necessary and/or desirable. These steps can be repeated until values for all properties are deemed to be acceptable or desired, or until no further improvement can be achieved.

Determination of translational kinetics values for codon pairs

[0115] The methods and sequences provided herein include determination and use of translational kinetics values for codon pairs. As provided herein, such a translational kinetics value can be calculated and/or empirically measured, and the final translational kinetics value used in graphical displays and methods of predicting translational kinetics can be a refined value resultant from two or more types of codon pair translational kinetics information. The various types of codon pair translational kinetics information that can be used in refining or replacing a translational kinetics value for a codon pair include, for example, values of observed versus expected codon pair frequencies in a particular organism, normalized values of observed versus expected codon pair frequencies in a particular organism, the degree to which observed versus expected codon pair frequency values are conserved in related proteins across two or more species, the degree to which observed versus expected codon pair frequency values are conserved at predicted pause sites such as boundaries between autonomous folding units in related proteins across two or more species, the degree to which codon pairs are conserved at predicted pause sites across different proteins in the same species, and empirical measurement of translational kinetics for a codon pair.

[0116] The values of observed versus expected codon pair frequencies in a host organism can be determined by any of a variety of methods known in the art for statistically evaluating observed occurrences relative to expected occurrences. Regardless of the statistical method used, this typically involves obtaining codon sequence data for the organism, for example, on a gene-by-gene basis. In some embodiments, the analysis is focused only on the coding regions of the genome. Because the analysis is a statistical one, a large database is preferred. Initially, the total number of codons is determined and the number of times each of the 61 non-terminating codons appears is determined. From this information, the expected frequency of each of the 3721 (61^2) possible non-terminating codon pairs is calculated, typically by multiplying together the frequencies with which each of the component codons appears. This frequency analysis can be carried out on a global basis, analyzing all of the sequences in the database together; however, it is typically done on a local basis, analyzing each sequence individually. This
will tend to minimize the statistical effect of an unusually high proportion of rare codons in a sequence. After the frequency data is obtained, for each sequence in the database, the expected number of occurrences of each codon pair is calculated by, for example, multiplying the expected frequency by the number of pairs in the sequence. This information can then be added to a global table, and each next succeeding sequence can be analyzed in like manner. This analysis results in a table of expected and observed values for each of the 3271 non-terminating codon pairs. The statistical significance of the variation between the expected and observed values can then be calculated, and the resulting information can be used in further practice of the various examples and embodiments provided herein.

[0117] In some embodiments, the values of observed versus expected codon pair frequencies are chi-squared values, such as chi-squared 2 (chisq2) values or chi-squared 3 (chisq3) values. Methods for calculating chi-squared values can be performed according to any method known in the art, as exemplified in U.S. Patent No. 5,082,767, which is incorporated by reference herein in its entirety. The result of chi-squared calculations is a list of 3,721 non-terminating codon pairs, each with an expected and observed value, together with a value for chi-squared (chisq1):

\[ \text{chisq1} = (\text{observed-expected})^2 / \text{expected} \]

[0118] In order to remove the contribution to chi-squared of non-randomness in amino acid pairs, a new value chi-squared 2 (chisq2) can be calculated as follows. For each group of codon pairs encoding the same amino acid pair (i.e., 400 groups), the sums of the expected and observed values are tallied; any non-randomness in amino acid pairs is reflected in the difference between these two values. Therefore, each of the expected values within the group is multiplied by the factor [sum observed/sum expected], so that the sums of the expected and observed values with the group are equal. The new chi-squared, chisq2, is evaluated using these new expected values. Calculation methods for removing the contribution to chi-squared of non-randomness in amino acid pairs are known in the art, as exemplified in Gutman and Hatfield, Proc. Natl. Acad. Sci. USA, (1989) 86:3699-3703.

[0119] Further, in order to remove the contribution to chi-squared of non-randomness in dinucleotides, a new value chi-squared 3 (chisq3) can be calculated. Correction is made only for those dinucleotides formed between adjacent codon pairs; any bias of dinucleotides within codons (codon triplet positions I-II and II-III) will directly affect codon usage and is, therefore, automatically taken into account in the
underlying calculations. For each dinucleotide pair formed between adjacent codon pairs (i.e., 16 pairs), the sums of the expected and observed values are tallied; any non-randomness in dinucleotide pairs is reflected in the difference between these two values. Therefore, each of the expected values within the group is multiplied by the factor [sum observed/sum expected], so that the sums of the expected and observed values with the group are equal. The new chi-squared, chisq3, is evaluated using these new expected values.

[0120] As provided herein, and as will be readily apparent to those skilled in the statistical art, that further values chi-squared N (chisqN) could be calculated similarly by removing one or more other variables in like fashion.

[0121] Analyses of the E. coli, S. cerevisiae, and human databases illustrate two important features. First, there is a highly significant codon pair bias in all three species, even after the amino acid nearest neighbor bias (chisq2) and the dinucleotide bias (chisq3) are discounted. Second, the effect associated with dinucleotide bias, i.e., the difference between chisq2 and chisq3, is much more pronounced in eukaryotes than in E. coli. It is by far the predominant effect in mammals, representing two thirds of the amount of chisq2 in excess of its expectation in human. Mouse and rat data exhibit a very similar pattern. Dinucleotide bias represents a smaller effect in yeast, and only a very minor one in E. coli. Although the predominant dinucleotide bias in human is the well-known CpG deficit, other dinucleotides are also very highly biased. For example, there is a deficit of TA, as well as an excess of TG, CA and CT. Overall, the deficit of CpG contributes only 35% of the total dinucleotide bias in the human database, and 17% in yeast.

[0122] As provided herein, the values of observed versus expected codon pair frequencies in a host organism herein can be normalized. Normalization permits different sets of values of observed versus expected codon pair frequencies to be compared by placing these values on the same numerical scale. For example, normalized codon pair frequency values can be compared between different organisms, or can be compared for different codon pair frequency value calculations within a particular organism (e.g., different calculations based on input sequence information or based on different calculations such as chisql or chisq2 or chisq3). Typically, normalization results in codon pair frequency values that are described in terms of their mean and standard deviation from the mean.
An exemplary method for normalizing codon pair frequency values is the calculation of z scores. The z score for an item indicates how far and in what direction that item deviates from its distribution’s mean, expressed in units of its distribution’s standard deviation. The mathematics of the z score transformation are such that if every item in a distribution is converted to its z score, the transformed scores will have a mean of zero and a standard deviation of one. The z scores transformation can be especially useful when seeking to compare the relative standings of items from distributions with different means and/or different standard deviations, z scores are especially informative when the distribution to which they refer is normal. In a normal distribution, the distance between the mean and a given z score cuts off a fixed proportion of the total area under the curve.

An exemplary method for determining z scores for codon pair chi-squared values is as follows: First, a list of all 3721 possible non-terminating codon pairs is generated. Second, for the ith codon pair, the ith chi-squared value is calculated, where the ith chi-squared value is denoted c. The chi-squared value, Σi, is given the sign of (observed - expected), so that over-represented codon pairs are assigned a positive c, and under-represented codon pairs are assigned a negative Σi. The formula for c, is:

\[ Σ_i = \text{sgn}(\text{obs}, - \text{exp.}) \ast (\text{obs}, - \text{exp.})^2 / \text{exp}, \]

Third, the mean chi-squared value is calculated where the mean is denoted m. The formula for the mean is:

\[ m = (Σ^1_i \Sigma) / 3721 \]

where Σ means sum over i. Fourth, the standard deviation of the chi-squared values is calculated, where the standard deviation is denoted s. The formula for the standard deviation is:

\[ s = \sqrt{(Σ^1_i (Σ) - m)^2 / 3721} \]

where \( \sqrt{ } \) means square root. Fifth, for the ith chi-squared value Σi, a z score is calculated by subtracting the mean then dividing by the standard deviation, wherein the ith z score is denoted Zi. The formula for the z score is:

\[ z_i = (c_i - m) / s \]

The above-described values of observed codon pair frequency versus expected codon pair frequency can be used as first approximations of translational kinetics of a polypeptide-encoding nucleotide sequence. However, such values are not true predictors of translational kinetics, and refinement of such values to more accurately predict translational kinetics can be performed according to the methods provided herein.
Thus, provided herein are methods of refining the predictive capability of a translational kinetics value of a codon pair in a host organism by providing an initial translational kinetics value based on the value of observed codon pair frequency versus expected codon pair frequency for a codon pair in a host organism, providing additional translational kinetics data for the codon pair in the host organism, and modifying the initial translational kinetics value according to the additional codon pair translational kinetics data to generate a refined translational kinetics value for the codon pair in the host organism. The translational kinetics data that can be used to refine translational kinetics values and methods of modifying translational kinetics values according to such additional translational kinetics data to generate a refined translational kinetics value for a codon pair in a host organism are provided below.

[0127] In one embodiment, translational kinetics data that can be used to refine translational kinetics values are based on recurrence of a codon pair and/or recurrence of a predicted translational kinetics value associated with a codon pair. Recurrence-based refinement of translational kinetics values is based on the investigation of multiple polypeptide-encoding nucleotide sequences to determine whether or not there are multiple occurrences of either codon pairs or predicted translational kinetics values in those sequences. Recurrence-based refinement of translational kinetics can be performed using any of a variety of known sequence comparison methods consistent with the examples provided herein. For purposes of exemplification, and not for limitation, the following example of recurrence-based refinement of translational kinetics is provided.

[0128] In one exemplary embodiment, the predicted translational kinetics value for a codon pair can be refined according to the degree to which observed versus expected codon pair frequency values are conserved in related proteins across two or more species. As provided herein, related proteins are proteins having homologous amino acid sequences and/or similar three dimensional structures. Related proteins having homologous amino acid sequences will typically have at least about 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% sequence identity. Related proteins having similar three dimensional structures will typically share similar secondary structure topology and similar relative positioning of secondary structural elements; exemplary related proteins having three dimensional structures are members of the same SCOP-classified Family (see, e.g., Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol. Biol.* 247, 536-540.).
The observed versus expected codon pair frequency values for any given codon pair can vary from species to species. However, as provided herein, evolutionarily related proteins in different species will typically conserve some or all translational pause or slowing sites. Based on this, an observed conservation of one or more predicted translational pause or slowing sites in evolutionarily related proteins of different species can confirm or increase the likelihood that a translational pause or slowing site is a functional translational kinetics signal. The codon pair located at the position on a protein that is confirmed as, or considered to have an increased likelihood of, containing an actual translational pause or slowing can itself be confirmed as being, or considered to have an increased likelihood of being, a functional translational kinetics signal. Similarly, a codon pair located at a position on a protein that is confirmed as not containing, or considered to have a decreased likelihood of containing, an actual translational pause or slowing, can itself be confirmed as not acting, or considered to have an decreased likelihood of acting, as a functional translational kinetics signal. Accordingly, initially predicted translational kinetics data, e.g., data based on values of observed codon pair frequency versus expected codon pair frequency, can be modified according to conserved codon pair frequency values across two or more species, which can lead to the codon pair being confirmed as: being a functional translational kinetics signal; being considered to have an increased likelihood of being a functional translational kinetics signal; being confirmed as not acting as an actual translational pause codon pair; or being considered to have a decreased likelihood of being a functional translational kinetics signal.

In another embodiment, the predicted translational kinetics value for a codon pair can be refined according to the presence of the codon pair at a location predicted by methods other than codon pair frequency methods to contain a translational pause or slowing site. One example of such a predicted location is a boundary location between autonomous folding units of a protein. While not intending to be limited to the following, it is proposed that translational pauses are present in wild type genes in order to slow translation of a nascent polypeptide subsequent to translation of a secondary structural element of a protein and/or a protein domain, thus providing time for acquisition of secondary and at least partial tertiary structure by the nascent protein prior to further downstream translation, and thereby allowing each domain to partially organize and commit to a particular, independent fold. As such, it is proposed herein that codon pairs can be associated with translational pauses between autonomous folding units of a
protein, where autonomous folding units can be secondary structural elements such as an alpha helix, or can be tertiary structural elements such as a protein domain. Thus, the presence of a codon pair at a boundary location between autonomous folding units of a protein can confirm or increase the likelihood that the codon pair acts to pause or slow translation. Accordingly, predicted translational kinetics data, e.g., data based on values of observed codon pair frequency versus expected codon pair frequency, can be modified according to the presence of the codon pair at a boundary location between autonomous folding units of a protein, which can increase the likelihood of the codon pair acts to pause or slow translation. For example, an over-represented codon pair that is present at a boundary location between autonomous folding units of a protein can be confirmed as acting as a translational pause or slowing codon pair.

[0131] In the above embodiment, a single observation of the codon pair at a boundary location between autonomous folding units of a protein can confirm or increase the likely translational pause or slowing properties of a codon pair. However, typically a plurality of observations will be used to more accurately estimate the translational pause or slowing properties of a codon pair. Thus, methods of using, for example, predicted boundary locations can be combined with methods that are based on recurrence of a codon pair and/or recurrence of a predicted translational kinetics value associated with a codon pair in methods of refining a predicted translational kinetics value for a codon pair. For example, a protein present in two or more species can have conserved boundary locations between autonomous folding units of the protein, and recurrent presence of an over-represented codon pair at the boundary locations can confirm the likelihood of an actual translational pause at that boundary location, leading to confirmation, or increased likelihood, that the corresponding codon pair for the respective species acts as a translational pause or slowing codon pair. In another example, two or more proteins of the same species can have boundary locations between autonomous folding units, and recurrent presence of an over-represented codon pair at the boundary locations can confirm or indicate the likelihood of an actual translational pause at that boundary location, leading to confirmation or indication of increased likelihood that the corresponding codon pair acts as a translational pause or slowing codon pair.

[0132] Such recurrence-based methods also can be used to confirm or indicate increased likelihood that a non-over-represented codon pair (e.g., an under-represented codon pair or a represented-as-expected codon pair) acts as a translational pause or slowing codon pair. For example, two or more proteins of the same species can have
boundary locations between autonomous folding units, and recurrent presence of a non-over-represented codon pair at the boundary locations, particularly if no over-represented codon pair is present, can confirm or indicate the likelihood of an actual translational pause at that boundary location, leading to confirmation or indication of increased likelihood that the corresponding codon pair acts as a translational pause or slowing codon pair.

[0133] Such recurrence-based methods also can be used to confirm or indicate the likelihood that a codon pair, such as an over-represented codon pair, does not act as a translational pause or slowing codon pair. For example, two or more proteins of the same species can have boundary locations between autonomous folding units, and consistent absence of a non-over-represented codon pair at the boundary locations can confirm or indicate increased likelihood that the codon pair does not act as a translational pause or slowing codon pair.

[0134] In another embodiment, the predicted translational kinetics value for a codon pair can be refined according to empirical measurement of translational kinetics for a codon pair. The influence of a codon pair on translational kinetics can be experimentally measured, and these experimental measurements can be used to refine or replace the predicted translational kinetics values for a codon pair. Several methods of experimentally measuring the translational kinetics of a codon pair are known in the art, and can be used herein, as exemplified in Irwin et al, J. Biol. Chem., (1995) 270:22801. One such exemplary assay is based on the observation that a ribosome pausing at a site near the beginning of an mRNA coding sequence can inhibit translation initiation by physically interfering with the attachment of a new ribosome to the message, and, thus, the codon pair to be assayed can be placed at the beginning of a polypeptide-encoding nucleotide sequence and the effect of the codon pair on translational initiation can be measured as an indication of the ability of the codon pair to cause a translational pause.

Another such exemplary assay is based on the fact that the transit time of a ribosome through the leader polypeptide coding region of the leader RNA of the trp operon sets the basal level of transcription through the trp attenuator, and, thus, the codon pair to be assayed can be placed into a trpLep leader polypeptide codon region, and level of expression can be inversely indicative of the translational pause properties of the codon pair, due to a faster translation causing formation of a stem-loop attenuator in the leader RNA, which results in transcriptional attenuation.
As will be apparent to one skilled in the art, the methods provided herein for calculation of translational kinetics values can be applied to the native organism of the polypeptide of SEQ ID NO: 2, and also can be applied to a selected organism in which the polypeptide of SEQ ID NO: 2, or a modification thereof, is to be heterologously expressed. For example, the nucleotide sequence information of an organism can be used to calculate chi-squared values in accordance with the methods provided herein, and the translational kinetics values can be based on these chi-squared values as well as on additional translational kinetics information provided herein, including, but not limited to, codon pairs conserved in domain boundaries and empirically measured translational kinetics for a codon pair. Exemplary organisms for which translational kinetics values can be calculated and used to prepare a nucleotide sequence encoding a cellobiohydrolase protein provided herein include Pichia pastoris; Oryctolagus cuniculus (rabbit); Macaca fascicularis (Long-tailed monkey); M. mulatta (Monkey); E. coli K12 W3110; E. coli UTI89; E. coli O157:H7 EDL933; E. coli O157:H7 str. Sakai; Bombyx mori; Spodoptera frugiperda; Drosophila melanogaster, Kluyveromyces lactis, Zymomonas mobilis and Schizosaccharomyces pombe.

Calculation methods of modifying translational kinetics values based on additional translational kinetics data

The translational kinetics data described herein can be combined in such a manner as to provide a refined translational kinetics value for a codon pair in a host organism. Methods of combining predictive data to arrive at a refined predictive value are known in the art and can be used herein.

Estimates for translational kinetics values are informed by a number of knowledge sources known to those skilled in the art, including but not limited to experimental measurement, conservation at protein structural boundaries and across homologous families, statistical inference from genomic sequence data, and the like as provided elsewhere herein. All these disparate knowledge sources must be integrated into an overall estimate for purposes of gene design and engineering. The general problem of integrating diverse and disparate knowledge sources is ubiquitous and well-studied in many different engineering fields, e.g., distributed sensor fusion in remote sensing, bagging classifiers in machine learning, heterogeneous database integration in data warehouses, or perceptual integration in artificial intelligence. Many useful and applicable approaches are known to the art.

Operationally, this means to choose H so as to maximize the probability of H given D, written P(H|D). By Bayes's rule, this may be rewritten as P(H|D) = P(D|H) * P(H) / P(D). This is equivalent to maximizing P(D|H) * P(H) because P(D) is constant for all H. The term P(H) is identified with the degree of belief in hypothesis H before the data was observed. The term P(D|H), read "the probability of D given H," is identified with how well hypothesis H predicts the observed data D. Thus, the Bayesian approach seeks to find an hypothesis that is a priori likely and also explains the data well.

In this example, an hypothesis H is that a given sequence feature, e.g., a given codon pair, has utility for translational kinetics engineering, e.g., creates a translational pause site. The observed data D may have several observations, e.g., D = D1 & D2 & D3 & D4, where D1 = an experimental measurement, D2 = conserved at protein structural domain boundaries, D3 = conserved across homologous protein families, and D4 = indicated as over-represented by statistical analysis that yields a high chisq3 value. In this case, the term P(D|H) = P(D1 & D2 & D3 & D4 | H), which indicates to choose an hypothesis that explains each of the observed datum. Of course, different data sources have different rates and magnitudes of observational error. This falls naturally into the Bayesian approach because the probability framework extends naturally to encompass the probability of observational error, as P(D|H) = P(D|H) * P(D is correct) + P(not D|H) * P(D is not correct). For example, an experimental measurement D1 that has been confirmed by replicate testing would have a very low probability of error, and therefore it would dominate the estimate if available.

In the general case, where no experimental measurement is available, several Bayesian approaches are commonly employed. The simplest, which often works
well, is named "Naive Bayes" because it assumes conditional independence among the individual observed data items. In this case, \( P(D|H) = P(D_1 & D_2 & D_3 & D_4 | H) = P(D_1 | H) * P(D_2 | H) * P(D_3 | H) * P(D_4 | H) \), where each of the individual terms is further expanded as \( P(D_i | H) = P(D_i | H) * P(D_i \text{ is correct}) + P(\text{not } D_i | H) * P(D_i \text{ is not correct}) \) as indicated above. The terms \( P(D_i \text{ is correct}) \) and \( P(D_i \text{ is not correct}) \) can be estimated a priori by the correlation of \( D_i \) with previous experimental measurements. The terms \( P(D_i | H) \) and \( P(\text{not } D_i | H) \) are obtained by observing whether or not hypothesis \( H \) is consistent with observed data item \( D_i \). More complex and powerful Bayesian approaches are also well known to the art. The fully general approach rewrites \( P(D|H) = P(D_1 & D_2 & D_3 & D_4 | H) = P(D_4 | D_3 & D_2 & D_1 & H) * P(D_3 | D_2 & D_1 & H) * P(D_2 | D_1 & H) * P(D_1 | H) \). Many other approaches, both Bayesian and others, are well known to the art.

[0142] By way of example, the translational kinetics values for a codon pair can be refined by consideration of, for example, chi-squared value of observed versus expected codon pair frequency and the degree to which codon pairs are conserved at predicted pause sites across different proteins in the same species, for example, at protein structure domain boundaries. An over-represented codon pair which is present with above-random frequency at boundary locations between autonomous folding units of proteins in the same species can have a translational kinetics value reflecting higher predicted translational pause properties of the codon pair. In contrast, an over-represented codon pair which is present with below-random frequency at boundary locations between autonomous folding units of proteins in the same species can have a translational kinetics value reflecting lower predicted translational pause properties of the codon pair.

[0143] As another example, the translational kinetics values for a codon pair can be refined by consideration of, for example, experimentally measured translation step times in one species and the degree to which codon pairs that correspond to measured pause sites in the first species are conserved across homologous proteins in other species, for example, in a multiple sequence alignment. When an over-represented codon pair in another species is aligned with above-random frequency to a codon pair that corresponds to a measured translation pause site in the first species, it can have a translational kinetics value reflecting higher predicted translational pause properties of that codon pair in the other species. In contrast, when an over-represented codon pair in another species is aligned with below-random frequency to a codon pair that corresponds to a measured
translation pause site in the first species, it can have a translational kinetics value reflecting lower predicted translational pause properties of that codon pair in the other species.

[0144] In various embodiments described herein, translational kinetics values for codon pairs, including refined translational kinetics values, can be determined. The translational kinetic values can be organized according to the likelihood of causing a translational pause or slowing based on any method known in the art. In one example, the translational kinetic values for two or more codon pairs, up to all codon pairs, in an organism are determined, and the mean translational kinetics value and associated standard deviation are calculated. Based on this, the translational kinetics value for a particular codon pair can be described in terms of the multiple of standard deviations the translational kinetics value for the particular codon pair differs from the mean translational kinetics value. Accordingly, reference herein to mean translational kinetics values and standard deviations, whether or not applied to a particular expression of translational kinetics value, can be applied to any of a variety of expressions of translational kinetics values provided herein.

**Graphical analysis of translational kinetics**

[0145] Also provided herein are methods of analyzing translational kinetics of an mRNA into polypeptide encoded by a gene in a host organism by determining translational kinetics values for codon pairs in the host organism and generating a graphical display of the translational kinetics values of actual codon pairs of an original polypeptide-encoding nucleotide sequence of a heterologous gene as a function of codon position. Such a graphical display provides a visual display of the predicted translational influence, including translational pause or slowing for numerous or all codon pairs of a polypeptide-encoding nucleotide sequence. This visual display can be used in methods of modifying polypeptide-encoding nucleotide sequences in order to thereby modify the predicted translational kinetics of the mRNA into polypeptide in methods such as those provided herein. For example, the graphical displays can be used to identify one or more codon pairs to be modified in a polypeptide-encoding nucleotide sequence. The graphical displays can be used in analyzing a polypeptide-encoding nucleotide sequence prior to modifying the polypeptide-encoding nucleotide sequence, or can be used in analyzing a modified polypeptide-encoding nucleotide sequence to determine, for example, whether or not further modifications are desired.

-56-
[0146] Methods for creating and using graphical displays can be performed according to any method known in the art, as exemplified in U.S. Patent Publication No. 2007/0298503, published on December 27, 2007, and U.S. Patent Publication No. 2007/0275399, published on November 29, 2007, which are incorporated by reference herein in their entireties. In particular, graphical displays as described therein can be created to illustrate the translational kinetics of an original or redesigned polypeptide-encoding nucleotide sequence in the native or a heterologous organism, or to illustrate differences and/or similarities of translation kinetic of a polypeptide-encoding nucleotide sequence in which one or more codon pairs have been modified. Additionally, numerous normalized graphical displays can be created to illustrate differences and/or similarities of translation kinetics of a polypeptide-encoding nucleotide sequence when expressed in two or more different organisms.

[0147] The graphical displays can be created using translational kinetics values based on any of the methods for determining translational kinetics values provided herein or otherwise known in the art. For example, chi-squared as a function of codon pair position, chi-squared 2 as a function of codon position, or chi-squared 3 as a function of codon pair position, translational kinetics values thereof, empirical measurement of translational pause of codon pairs in a host organism, estimated translational pause capability based on observed presence and/or recurrence of a codon pair at predicted pause site, and variations and combinations thereof as provided herein.

[0148] The exact format of the graphical displays can take any of a variety of forms, and the specific form is typically selected for ease of analysis and comparison between plots. For example, the abscissa typically lists the position along the nucleotide sequence or polypeptide sequence, and can be represented by nucleotide position, codon position, codon pair position, amino acid position, or amino acid pair position. In such instances, the ordinate typically lists the translational kinetics value of the codon pair, such as, but not limited to, a translational kinetics value of codon pair frequency, including, but not limited to the z score of chisq1, the z score of chisq2, the z score of chisq3, the empirically measured value, and the refined translational kinetics value. In alternative embodiments, the sequence position can be plotted along the ordinate and the translational kinetics value can be plotted along the abscissa.

[0149] As an example, a graphical display of translational kinetics is depicted in Figure 1, where each positive deflection or peak describes a predicted translational pause or slowing at the nucleotide location as defined by the abscissa.
Comparing plots

[0150] Also contemplated herein are methods in which a set of graphical displays, including at least a first graphical display and a second graphical display, are prepared. These sets of displays can be compared in order to determine the difference in predicted translational efficiency or translational kinetics of the two plots. The plots can differ according to any of a variety of criteria. For example, each plot can represent a different polypeptide-encoding nucleotide sequence, each plot can represent a different host organism, each plot can represent differently determined translational kinetics values, or any combination thereof. As will be apparent to one skilled in the art, any number of different graphical displays can be compared in accordance with the methods provided herein, for example, 2, 3, 4, 5, 6, 7, 8 or more different graphical displays can be compared. Typically, two plots will represent different polypeptide-encoding nucleotide sequences, the same sequence in different host organisms, or different sequences in different host organisms.

[0151] Comparison of different graphical displays can be used to analyze the predicted change in translational kinetics as a result of the difference represented by the graphical displays. For example, comparison of the same polypeptide-encoding nucleotide sequence in different host organisms can be used to analyze any predicted transcriptional pauses that can be removed. Accordingly, provided herein are methods of analyzing translational kinetics of an mRNA into polypeptide in a host organism by comparing two graphical displays to understand or predict the differences in translational kinetics of the mRNA into polypeptide, where the differences in the graphical displays can be as a result of, for example, a difference in the polypeptide-encoding nucleotide sequence or a difference in the host organism. Upon determination of the differences in translational kinetics, it can be evaluated whether or not the change in translational kinetics as a result of the underlying difference between the two graphical displays is desirable. Such comparison methods also can lead to an identification of further modifications, e.g., further modifications to the polypeptide-encoding nucleotide sequence to further improve translational kinetics. Accordingly, it is contemplated herein that such comparison methods can be carried out iteratively.

[0152] In embodiments where it is desired to improve expression of a polypeptide-encoding nucleotide sequence in a particular heterologous host, a graphical display of the translational kinetics values of codon pairs for the original polypeptide-
encoding nucleotide sequence in the heterologous host can be compared to a graphical
display of the translational kinetics values of codon pairs for a modified polypeptide-
encoding nucleotide sequence in the heterologous host, and it can be determined whether
or not the modification to the polypeptide-encoding nucleotide sequence resulted in
improved translational kinetics.

Methods of expressing polypeptide

[0153] Also provided herein is an expression system, comprising an
expression vector in a host organism, wherein the expression vector includes a DNA
sequence of the embodiments provided herein operably linked to an expression control
sequence. As used herein, an expression vector is a DNA or RNA vector that is capable
of transforming a host cell and of effecting expression of a specified nucleic acid
molecule. Typically, the expression vector is also capable of replicating within the host
cell. Expression vectors can be either prokaryotic or eukaryotic, and are typically viruses
or plasmids.

[0154] The term operably linked refers to functional linkage between a nucleic
acid expression control sequence (such as a promoter, or array of transcription factor
binding sites) and a second nucleic acid sequence, wherein the expression control
sequence directs transcription of the nucleic acid corresponding to the second sequence.
An operably linked expression vector can also include secretion signals and other
modifying sequences, and may encode chaperones and proteins for a variety of organisms
and systems.

[0155] Also provided herein are methods of expressing a polypeptide-
encoding nucleotide sequence generated by the methods provided herein. Methods of
expressing polypeptides from polypeptide-encoding nucleotide sequences are known in
the art, as exemplified, for example, by the techniques described in Maniatis et al., 1989,
Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, N.Y. and
Ausubel et al., 2006, Current Protocols in Molecular Biology, Greene Publishing
Associates and Wiley Interscience, N.Y. The methods include inserting a polypeptide-
encoding nucleotide sequence designed by the methods provided herein into a cell, and
expressing the polypeptide-encoding nucleotide sequence under conditions suitable for
gene expression. Additionally provided expression methods include cell-free expression
systems as known in the art, where such methods include providing a polypeptide-
encoding nucleotide sequence designed by the methods provided herein and contacting
the polypeptide-encoding nucleotide sequence with a cell-free expression system under conditions suitable for protein translation.

Systems and methods for degrading cellulose

[0156] Also provided herein are systems and methods for degrading cellulose, comprising one or more host organisms that collectively include DNA sequences operably encoding at least two of the following enzymes: endo-1,4-β-glucanase, exo-1,4-β-D-glucanase, and β-D-glucosidase. In certain aspects, one or more, or all of the enzymes are heterologous to the one or more host organisms. In certain aspects, the translational kinetics of each of the DNA sequences encoding the enzymes has been increased by silent permutation or conservative amino acid substitution of at least 1, 2, or 3 codon pairs present in the original sequence for each enzyme. A silent permutation is a change to one or more nucleotides of a codon such that the encoded amino acid does not change. In certain aspects, the at least 1, 2 or 3 substituted codon pairs are predicted to cause a translational pause or slowing in the host organism, and the substituting codon pair is typically a codon pair not predicted to cause a translational pause or slowing in the host organism.

[0157] In some aspects, the one or more host organisms are selected from the group consisting of: *Saccharomyces cerevisiae*, *Pichia pastoris*, *Escherichia coli*, *Bombyx mori*, *Spodoptera frugiperra*, *Drosophila melanogaster*, *Kluyveromyces lactis*, *Zymomonas mobilis* and *Schizosaccharomyces pombe*.

[0158] In some aspects, each encoded enzyme in the system has at least a 50%, 60%, 70%, 80%, and more typically at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% amino acid sequence identity to the with the original sequence of the enzyme.

[0159] In some aspects, one or more of the endo-1,4-β-glucanase, exo-1,4-β-D-glucanase, and β-D-glucosidase enzymes in the system retains at least 75% of the enzymatic activity of the enzyme encoded by the original sequence under conditions suitable for degradation of cellulose. Methods for measuring the activity of the enzymes in the system are known in the art. For example, the incorporated materials of U.S. Patent No. 6,566,113 provide methods for measuring the activity of celllobiohydrolases that have been recombinantly expressed.

[0160] Also provided are methods of hydrolyzing a carbohydrate comprising providing a carbohydrate comprising at least one glycosidic bond, providing a
polypeptide encoded by any of the polynucleotides provided herein, and contacting said carbohydrate with said polypeptide under conditions that permit said polypeptide to hydrolyze at least one glycosidic bond of said carbohydrate, whereby at least one glycosidic bond of said carbohydrate is hydrolyzed. In some such embodiments, the carbohydrate is cellulose. In some such embodiments, the carbohydrate comprises two or more β-1,4-linked glucose units. Typically such methods can be performed using the cells and systems provided herein. Such methods can be performed in order to provide smaller polysaccharides and/or monosaccharides which can be used by a cell or processed extracellularly according to any one of a variety of known methods in the art.

[0161] The following examples are included for illustrative purposes only and are not intended to limit the scope of the invention.

EXAMPLE 1

[0162] This example describes optimization of a DNA sequence encoding TrCBH-II for expression in yeast.

[0163] Chi-squared values for S. cerevisiae were determined using previously reported methods (Hatfield and Gutman, "Codon Pair Utilization Bias in Bacteria, Yeast, and Mammals" in Transfer RNA in Protein Synthesis, Hatfield, Lee and Pirtle Eds. CRC Press (Boca Raton, LA) 1993). Briefly, non-redundant protein coding regions for S. cerevisiae was obtained from GenBank sequence database (75,403 codon pairs in 177 sequences for S. cerevisiae) to determine an observed number of occurrences for each codon pair. The expected number of occurrences of each codon pair was calculated under the assumption that the codon pairs are used randomly. The chi-squared value "chisql" was generated by the expected and observed values determined. The chisql was recalculated to remove any influence of non-randomness in amino acid pair frequencies, yielding "chisq2." The chsq2 was re-calculated to remove any influence of non-randomness in dinucleotide frequencies, yielding "chisq3." z scores of chisq3 were calculated by determining the mean chisq3 value and corresponding standard deviation for all codon pairs, and normalizing each chisq3 value to be reported in terms of number of standard deviations from the mean chisq3 values.

[0164] The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to optimize codon usage for S. cerevisiae. The DNA sequence encoding TrCBH-II (SEQ ID NO: 1) was derived from GenBank accession number M16190 by removing untranslated sequence (5' untranslated region and introns).
A graphical display for the native gene (SEQ ID NO: 1) encoding the TrCBH-II protein (SEQ ID NO: 2) in T. reesei was prepared by plotting z scores of translational kinetics values for codon pair utilization in T. reesei as a function of codon pair position. The graphical display is provided in Figure 1.

A graphical display for the native gene (SEQ ID NO: 1) encoding the TrCBH-II protein (SEQ ID NO: 2) in S. cerevisiae was prepared by plotting z scores of translational kinetics values for codon pair utilization in S. cerevisiae as a function of codon pair position. The graphical display is provided in Figure 2A.

The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to no longer contain codon pairs having z scores in S. cerevisiae greater than 3. The resulting nucleotide sequence (SEQ ID NO: 3) was found to encode a protein (SEQ ID NO: 4) with 100% amino acid sequence identity to wild-type TrCBH-II (SEQ ID NO: 2). A graphical display for the codon pair utilization-modified gene (SEQ ID NO: 3) encoding the TrCBH-II protein (SEQ ID NO: 4) expressed in S. cerevisiae was prepared by plotting z scores of translational kinetics values for codon pair utilization in S. cerevisiae as a function of codon pair position. The graphical display is provided in Figure 2B.

**EXAMPLE 2**

This example describes optimization of a DNA sequence encoding TrCBH-II for expression in bacteria.

Chi-squared values for E. coli were determined as described in Example 1, with the following differences. Briefly, non-redundant protein coding regions for E. coli were obtained from GenBank sequence database (75,096 codon pairs in 237 sequences for E. coli) to determine an observed number of occurrences for each codon pair. The expected number of occurrences of each codon pair was calculated under the assumption that the codon pairs are used randomly. Chi-squared values chisq1, chisq2, chisq3 and z scores of chisq3 were calculated as described in Example 1.

The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to optimize codon usage for E. coli. A graphical display for the native gene (SEQ ID NO: 1) encoding the TrCBH-II protein (SEQ ID NO: 2) in E. coli was prepared by plotting z scores of translational kinetics values for codon pair utilization in E. coli as a function of codon pair position. The graphical display is provided in Figure 3A.
The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to no longer contain codon pairs having \( z \) scores in \( E. coli \) greater than 3. The resulting nucleotide sequence (SEQ ID NO: 9) was found to encode a protein (SEQ ID NO: 10) with 100% amino acid sequence identity to wild-type TrCBH-II (SEQ ID NO: 2). A graphical display for the codon pair utilization-modified gene (SEQ ID NO: 9) encoding the TrCBH-II protein (SEQ ID NO: 10) expressed in \( E. coli \) was prepared by plotting \( z \) scores of translational kinetics values for codon pair utilization in \( E. coli \) as a function of codon pair position. The graphical display is provided in Figure 3B.

**EXAMPLE 3**

This example describes optimization of a DNA sequence encoding TrCBH-II for expression in \( P. pastoris \).

Chi-squared values for \( P. pastoris \) were determined as described in Example 1, with the following differences. Briefly, non-redundant protein coding regions for \( P. pastoris \) were obtained from GenBank sequences to determine an observed number of occurrences for each codon pair. The expected number of occurrences of each codon pair was calculated under the assumption that the codon pairs are used randomly. Chi-squared values chisq1, chisq2, chisq3 and \( z \) scores of chisq3 were calculated as described in Example 1.

The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to optimize codon usage for \( P. pastoris \). A graphical display for the native gene (SEQ ID NO: 1) encoding the TrCBH-II protein (SEQ ID NO: 2) in \( P. pastoris \) was prepared by plotting \( z \) scores of translational kinetics values for codon pair utilization in \( P. pastoris \) as a function of codon pair position. The graphical display is provided in Figure 4A.

The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to no longer contain codon pairs having \( z \) scores in \( P. pastoris \) greater than 3. The resulting nucleotide sequence (SEQ ID NO: 15) was found to encode a protein (SEQ ID NO: 16) with 100% amino acid sequence identity to wild-type TrCBH-II (SEQ ID NO: 2). A graphical display for the codon pair utilization-modified gene (SEQ ID NO: 15) encoding the TrCBH-II protein (SEQ ID NO: 16) expressed in \( P. pastoris \) was prepared by plotting \( z \) scores of translational kinetics values for codon pair utilization in \( P. pastoris \) as a function of codon pair position. The graphical display is provided in Figure 4B.
EXAMPLE 4

[0176] This example describes optimization of a DNA sequence encoding TrCBH-II for expression in \textit{K. lactis}.

[0177] Chi-squared values for \textit{K. lactis} were determined as described in Example 1, with the following differences. Briefly, non-redundant protein coding regions for \textit{K. lactis} were obtained from GenBank sequences to determine an observed number of occurrences for each codon pair. The expected number of occurrences of each codon pair was calculated under the assumption that the codon pairs are used randomly. Chi-squared values chisql, chisq2, chisq3 and \( z \) scores of chisq3 were calculated as described in Example 1.

[0178] The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to optimize codon usage for \textit{K. lactis}. A graphical display for the native gene (SEQ ID NO: 1) encoding the TrCBH-II protein (SEQ ID NO: 2) in \textit{K. lactis} was prepared by plotting \( z \) scores of translational kinetics values for codon pair utilization in \textit{K. lactis} as a function of codon pair position. The graphical display is provided in Figure 5A.

[0179] The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to no longer contain codon pairs having \( z \) scores in \textit{K. lactis} greater than 3. The resulting nucleotide sequence (SEQ ID NO: 15) was found to encode a protein (SEQ ID NO: 16) with 100\% amino acid sequence identity to wild-type TrCBH-II (SEQ ID NO: 2). A graphical display for the codon pair utilization-modified gene (SEQ ID NO: 15) encoding the TrCBH-II protein (SEQ ID NO: 16) expressed in \textit{K. lactis} was prepared by plotting \( z \) scores of translational kinetics values for codon pair utilization in \textit{K. lactis} as a function of codon pair position. The graphical display is provided in Figure 5B.

EXAMPLE 5

[0180] This example describes optimization of a DNA sequence encoding TrCBH-II for expression in \textit{Z. mobilis}.

[0181] Chi-squared values for \textit{Z. mobilis} were determined as described in Example 1, with the following differences. Briefly, non-redundant protein coding regions for \textit{Z. mobilis} were obtained from GenBank sequences to determine an observed number of occurrences for each codon pair. The expected number of occurrences of each codon pair was calculated under the assumption that the codon pairs are used randomly. Chi-
squared values chisq1, chisq2, chisq3 and z scores of chisq3 were calculated as described in Example 1.

[0182] The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to optimize codon usage for Z. mobilis. A graphical display for the native gene (SEQ ID NO: 1) encoding the TrCBH-II protein (SEQ ID NO: 2) in Z. mobilis was prepared by plotting z scores of translational kinetics values for codon pair utilization in Z. mobilis as a function of codon pair position. The graphical display is provided in Figure 6A.

[0183] The nucleotide sequence for the gene encoding the TrCBH-II protein was modified to no longer contain codon pairs having z scores in Z. mobilis greater than 3. The resulting nucleotide sequence (SEQ ID NO: 15) was found to encode a protein (SEQ ID NO: 16) with 100% amino acid sequence identity to wild-type TrCBH-II (SEQ ID NO: 2). A graphical display for the codon pair utilization-modified gene (SEQ ID NO: 15) encoding the TrCBH-II protein (SEQ ID NO: 16) expressed in Z. mobilis was prepared by plotting z scores of translational kinetics values for codon pair utilization in Z. mobilis as a function of codon pair position. The graphical display is provided in Figure 6B.

EXAMPLE 6

[0184] Expression in E. coli of the codon optimized, codon pair utilization-based modification (Hot-Rrod) from Example 2 and native TrCBH-II protein is examined by Western blot analysis. Each vector is transformed into E. coli strain Top 10 (F-mcrA h[6mrr-hsdRMS-mcrBQ $\phi$ 80lacZ $\delta$M15 $\delta$lacX74 deoR recA1 araD139 $\delta$(ara-leu) 7697 galU galK rpsL (StrR) endA1 nupG). An overnight culture is inoculated at 1:100 into 5 ml of LB medium plus 100µg/ml ampicillin and grown at 37°C to OD$_{600}$ of 0.5. Protein expression is induced by addition of 0.002 or 0.02% L-arabinose and grown for 3hrs at 37°C. Cells are harvested by centrifugation and the cell pellets are resuspended in phosphate buffered saline. Cells are disrupted by sonication and supernatant and pellet fractions are resolved in a 4-20% SDS-polyacrylamide gel (Pierce). Proteins are transferred to Immobilon-P (Millipore, Bedford, MA) and are incubated with rabbit polyclonal anti-CBH-II antibody diluted 1:20,000. Rabbit IgG is visualized using a HRP-conjugated secondary antibody and ECL + Plus (Amersham, Buckinghamshire, UK) according to manufacturer's instructions.
Western blot analysis demonstrates that changes to a polypeptide-encoding nucleic acid sequence can increase expression of the polypeptide, particularly when the polypeptide is heterologously expressed.

Since modifications will be apparent to those of skill in this art, it is intended that this invention be limited only by the scope of the appended claims.
WHAT IS CLAIMED IS:

1. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 of the following codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof:
   - CCCTCT (nucleotides 463-468)
   - GGCCAA (nucleotides 94-99)
   - CAGTTT (nucleotides 565-570)
   - GATATC (nucleotides 703-708)
   - GTGGAA (nucleotides 691-696)
   - GGATTT (nucleotides 1192-1197)
   - GGTATT (nucleotides 1198-1203).

2. The DNA sequence of Claim 1, in which at least 5 of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

3. The DNA sequence of Claim 1, in which at least 3 of the following codon pair replacements have been made:
   - CCCTCT (nucleotides 463-468) replaced with CCTTCT
   - GGCCAA (nucleotides 94-99) replaced with GGTCGA
   - CAGTTT (nucleotides 565-570) replaced with CAATTT
   - GATATC (nucleotides 703-708) replaced with GACATT
   - GTGGAA (nucleotides 691-696) replaced with GTTGAA
   - GGATTT (nucleotides 1192-1197) replaced with GGTTC
   - GGTATT (nucleotides 1198-1203) replaced with GGAATT.

4. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 of the following codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof:
   - CTCGGT (nucleotides 760-765)
ATTGCC (nucleotides 631-636)
GACAGC (nucleotides 1285-1290)
GTCTGG (nucleotides 88-93)
GTCTGG (nucleotides 1246-1251)
TTGCTG (nucleotides 1231-1236)
GTTGGTG (nucleotides 571-576)
ACGCTG (nucleotides 22-27)
ACGCTG (nucleotides 31-36)
GACTGG (nucleotides 1168-1173)
GCCGGA (nucleotides 559-564)
CTGGTG (nucleotides 748-753).

5. The DNA sequence of Claim 4, in which at least 5 of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

6. The DNA sequence of Claim 4, in which at least 3 of the following codon pair replacements have been made:
   
   CTCGGT (nucleotides 760-765) replaced with CTGGGT
   ATTGCC (nucleotides 631-636) replaced with ATTGCG
   GACAGC (nucleotides 1285-1290) replaced with GACTCT
   GTCTGG (nucleotides 88-93) replaced with GTGTTG
   GTCTGG (nucleotides 1246-1251) replaced with GTTGGG
   TTGCTG (nucleotides 1231-1236) replaced with CTGCTG
   GTGGTG (nucleotides 571-576) replaced with GTGGTT
   ACGCTG (nucleotides 22-27) replaced with ACCCTC
   ACGCTG (nucleotides 31-36) replaced with ACCCTG
   GACTGG (nucleotides 1168-1173) replaced with GATTGG
   GCCGGA (nucleotides 559-564) replaced with GCGGGC
   CTGGTG (nucleotides 748-753) replaced with CTGGTT.

7. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 of the
following codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof:

- CAGTTT (nucleotides 565-570)
- TTTGAC (nucleotides 1303-1308)
- TCGTTT (nucleotides 1240-1245)
- GGCCAA (nucleotides 94-99)
- AAGAAT (nucleotides 541-546)
- AAGAAT (nucleotides 934-939)
- GCCAAA (nucleotides 649-654)
- GTCAAG (nucleotides 1252-1257)
- GGTATT (nucleotides 1198-1203)
- ATCAAC (nucleotides 808-813)
- GGCCAT (nucleotides 865-870)
- CTTCCA (nucleotides 835-840)
- GATATC (nucleotides 703-708)
- TCGTTG (nucleotides 1228-1233).

8. The DNA sequence of Claim 7, in which at least 5 of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

9. The DNA sequence of Claim 7, in which at least 3 of the following codon pair replacements have been made:

- CAGTTT (nucleotides 565-570) replaced with CAATTT
- TTTGAC (nucleotides 1303-1308) replaced with TTTGAT
- TCGTTT (nucleotides 1240-1245) replaced with TCTTTT
- GGCCAA (nucleotides 94-99) replaced with GGACAA
- AAGAAT (nucleotides 541-546) replaced with AAAAAT
- AAGAAT (nucleotides 934-939) replaced with AAAAAC
- GCCAAA (nucleotides 649-654) replaced with GCTAAA
- GTCAAG (nucleotides 1252-1257) replaced with GTTAAA
- GGTATT (nucleotides 1198-1203) replaced with GGAATC
- ATCAAC (nucleotides 808-813) replaced with ATTAAT
- GGCCAT (nucleotides 865-870) replaced with GGACAC
- CTTCCA (nucleotides 835-840) replaced with TTGCCT
GATATC (nucleotides 703-708) replaced with GATATA
TCGTTG (nucleotides 1228-1233) replaced with TCATTG.

10. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 of the following codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof:
   GGCCAA (nucleotides 94-99)
   CAGTTT (nucleotides 565-570)
   GATATC (nucleotides 703-708)
   TATTTG (nucleotides 853-858)
   GGCCAT (nucleotides 865-870)
   TCGTTG (nucleotides 1228-1233)
   TTTGTC (nucleotides 1243-1248)
   TTCCAA (nucleotides 1363-1368).

11. The DNA sequence of Claim 10, in which at least 5 of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

12. The DNA sequence of Claim 10, in which at least 3 of the following codon pair replacements have been made:
   GGCCAA (nucleotides 94-99) replaced with GGTCAA
   CAGTTT (nucleotides 565-570) replaced with CAATTC
   GATATC (nucleotides 703-708) replaced with GACATT
   TATTTG (nucleotides 853-858) replaced with TATTTA
   GGCCAT (nucleotides 865-870) replaced with GGACAT
   TCGTTG (nucleotides 1228-1233) replaced with TCTTTA
   TTTGTC (nucleotides 1243-1248) replaced with TTCGTT
   TTCCAA (nucleotides 1363-1368) replaced with TTCCAG.

13. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2, wherein at least 3 of the
following codon pairs of SEQ ID NO: 1 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof:

<table>
<thead>
<tr>
<th>Codon Pair</th>
<th>Nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTGCCT</td>
<td>55-60</td>
</tr>
<tr>
<td>GCCAAT</td>
<td>370-375</td>
</tr>
<tr>
<td>GCTATT</td>
<td>406-411</td>
</tr>
<tr>
<td>GCCGGA</td>
<td>559-564</td>
</tr>
<tr>
<td>GCCAAT</td>
<td>778-783</td>
</tr>
<tr>
<td>TTGGCA</td>
<td>967-972</td>
</tr>
<tr>
<td>AAGCTG</td>
<td>1051-1056</td>
</tr>
<tr>
<td>GCTATT</td>
<td>1066-1071</td>
</tr>
<tr>
<td>GCCAAT</td>
<td>1084-1089</td>
</tr>
<tr>
<td>ACCGGA</td>
<td>1147-1152</td>
</tr>
<tr>
<td>ACCGGA</td>
<td>1189-1194</td>
</tr>
<tr>
<td>GGTATT</td>
<td>1198-1203</td>
</tr>
<tr>
<td>GACAGC</td>
<td>1285-1290</td>
</tr>
<tr>
<td>GATGCC</td>
<td>1327-1332</td>
</tr>
<tr>
<td>GCCTTG</td>
<td>1330-1335</td>
</tr>
<tr>
<td>CAGCTT</td>
<td>1381-1386</td>
</tr>
</tbody>
</table>

14. The DNA sequence of Claim 13, in which at least 5 of the specified codon pairs have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof.

15. The DNA sequence of Claim 13, in which at least 3 of the following codon pair replacements have been made:

<table>
<thead>
<tr>
<th>Codon Pair</th>
<th>Nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTGCCT</td>
<td>55-60</td>
</tr>
<tr>
<td>GCCAAT</td>
<td>370-375</td>
</tr>
<tr>
<td>GCTATT</td>
<td>406-411</td>
</tr>
<tr>
<td>GCCGGA</td>
<td>559-564</td>
</tr>
<tr>
<td>GCCAAT</td>
<td>778-783</td>
</tr>
<tr>
<td>TTGGCA</td>
<td>967-972</td>
</tr>
<tr>
<td>AAGCTG</td>
<td>1051-1056</td>
</tr>
<tr>
<td>GCTATT</td>
<td>1066-1071</td>
</tr>
<tr>
<td>GCCAAT</td>
<td>1084-1089</td>
</tr>
<tr>
<td>ACCGGA</td>
<td>1147-1152</td>
</tr>
<tr>
<td>ACCGGA</td>
<td>1189-1194</td>
</tr>
<tr>
<td>GGTATT</td>
<td>1198-1203</td>
</tr>
<tr>
<td>GACAGC</td>
<td>1285-1290</td>
</tr>
<tr>
<td>GATGCC</td>
<td>1327-1332</td>
</tr>
<tr>
<td>GCCTTG</td>
<td>1330-1335</td>
</tr>
<tr>
<td>CAGCTT</td>
<td>1381-1386</td>
</tr>
</tbody>
</table>

AAGCTG replaced with AAATTG
GCTATT replaced with GCCATT
GCCAAT replaced with GCCAAT
ACCGGA replaced with ACCGGA
ACCGGA (nucleotides 1189-1194) replaced with ACAGGT
GGTATT (nucleotides 1198 - 1203) replaced with GGAATC
GACAGC (nucleotides 1285-1290) replaced with GATTCT
GATGCC (nucleotides 1327-1332) replaced with GACGCC
GCCTTG (nucleotides 1330-1335) replaced with GCCCTT
CAGCTT (nucleotides 1381-1386) replaced with CAGTTG.

16. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least three codon pairs of the wild-type sequence that are predicted to cause a translational pause in the host organism have been replaced with codon pairs that are predicted to be less likely to cause a translational pause therein, and wherein the host organism is not human, E. coli or S. cerevisiae.

17. The DNA sequence of Claim 16, wherein said at least three codon pairs of the wild-type sequence that are predicted to cause a translational pause are highly-overrepresented codon pairs.

18. The DNA sequence of Claim 16, wherein a codon pair predicted to be less likely to cause a translational pause is a codon pair that has a translational kinetics value greater than 1.5 times the standard deviation of translational kinetics values for the host organism.

19. A cellobiohydrolase-encoding DNA sequence, having at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least three codon pairs of the wild-type sequence that are predicted to cause a translational pause in the host organism have been replaced with codon pairs that are predicted to be less likely to cause a translational pause therein, wherein the host organisms are as follows:

*Pichia pastoris*
*Oryctolagus cuniculus* (rabbit)
*Macaca fascicularis* (Long-tailed monkey)
*Macaca mulatta* (Monkey)
*Escherichia coli* K12 W31 10
*Escherichia coli* UTI89
*Escherichia coli* O157:U7 EDL933
*Escherichia coli* O157:H7 str. Sakai
20. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least 3 codon pairs present in SEQ ID NO:1 and which encode amino acids 27-62 of SEQ ID NO: 2 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof, and wherein at least three replacement codon pair is predicted to be less likely to cause a translational pause in the heterologous host organism relative to the respective wild type codon pair when expressed in the heterologous host organism.

21. The cellobiohydrolase-encoding DNA sequence of Claim 20, wherein the translational kinetics value of at least one replacement codon pair when expressed in the heterologous host organism is no more than 150% of the translational kinetics value for the wild type codon pair when expressed in the native organism.

22. The cellobiohydrolase-encoding DNA sequence of any of Claims 20-21, wherein no replacement codon encoding amino acids 27-62 of SEQ ID NO: 2 has a z score for expression in the heterologous host that is more than 200% of the z score of the wild type codon pair TCCAAC when expressed in the native organism.

23. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least 3 codon pairs present in SEQ ID NO:1 and which encode amino acids 107-471 of SEQ ID NO: 2 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof, and wherein at least three replacement codon pair is predicted to be less likely to cause a translational pause in the heterologous host organism relative to the respective wild type codon pair when expressed in the heterologous host organism.

24. The cellobiohydrolase-encoding DNA sequence of Claim 23, wherein the translational kinetics value of at least one replacement codon pair when expressed in the
heterologous host organism is no more than 150% of the translational kinetics value for the wild type codon pair when expressed in the native organism.

25. The cellobiohydrolase-encoding DNA sequence of any of Claims 23-24, wherein no replacement codon encoding amino acids 107-471 of SEQ ID NO: 2 has a z score for expression in the heterologous host that is more than 200% of the z score of the wild type codon pair GCAAAG when expressed in the native organism.

26. A cellobiohydrolase-encoding DNA sequence, wherein the encoded sequence has at least a 75% amino acid sequence identity with amino acids 27-471 of wild-type cellobiohydrolase as set forth in SEQ ID NO: 2 and is adapted for expression in a heterologous host organism, wherein at least 1, 2 or 3 codon pairs present in SEQ ID NO: 1 and which encode amino acids 62-107 of SEQ ID NO: 2 have been replaced with different codon pairs encoding identical amino acids or conservative amino acid substitutions thereof, and wherein at least one replacement codon pair is predicted to be equally or more likely to cause a translational pause in the heterologous host organism relative to the respective wild type codon pair when expressed in the heterologous host organism.

27. The cellobiohydrolase-encoding DNA sequence of Claim 26, wherein the translational kinetics value of at least one replacement codon pair when expressed in the heterologous host organism is at least 75% of the translational kinetics value for the wild type codon pair when expressed in the native organism.

28. The cellobiohydrolase-encoding DNA sequence of any of Claims 26-27, wherein at least one replacement codon encoding amino acids 62-107 of SEQ ID NO: 2 has a z score for expression in the heterologous host that is more than 75% of the z score of the wild type codon pair TCTACT when expressed in the native organism.


30. An isolated polynucleotide comprising the DNA sequence of SEQ ID NOs: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 or 23.

31. An isolated polypeptide encoded by the DNA sequence of any of Claims 1-28, provided that the amino acid sequence of said polypeptide is not SEQ ID NO: 2.

32. An expression system, comprising:

an expression vector in a host organism, wherein the expression vector includes the polynucleotide of Claim 29 or Claim 30 operably linked to an expression control sequence.
33. A system for degrading cellulose, comprising:
    one or more host organisms that collectively include polynucleotides operably encoding the following enzymes:
    endo-1,4-β-glucanase,
    exo-1,4-β-D-glucanase, and
    β-D-glucosidase;
    wherein the enzymes are heterologous to the one or more host organisms, and wherein translational kinetics of each of the polynucleotides encoding the enzymes has been modified to replace at least three codon pairs present in the original sequence for each enzyme, wherein the at least three codon pairs are predicted to cause a translational pause in the host organism, and wherein said modification results in silent permutation or conservative amino acid substitution of said at least three codon pairs.

34. The system of Claim 32 or Claim 33, wherein said one or more host organisms are selected from the group consisting of: *Saccharomyces cerevisiae*, *Pichia pastoris*, *Escherichia coli*, *Bombyx mori*, *Spodoptera frugiperda*, *Drosophila melanogaster*, *Kluyveromyces lactis*, *Zymomonas mobilis* and *Schizosaccharomyces pombe*.

35. The system of any of Claims 32-34, wherein each encoded enzyme has at least a 75% amino acid sequence identity with the original sequence of said enzyme.

36. The system of any of Claims 32-35, wherein said exo-1,4-β-D-glucanase retains at least 75% of the enzymatic activity of wild-type TrCBH-II (SEQ ID NO: 2) under normal physiological conditions.

37. A cell comprising the polynucleotide of Claim 29 or Claim 30.

38. The cell of Claim 37, wherein said cell expresses the polypeptide encoded by said polynucleotide.

39. A method of introducing a polynucleotide into a host cell comprising:
    providing a host cell; and
    contacting said host cell with the polynucleotide of Claim 29 or Claim 30 under conditions that permit the polynucleotide to be introduced into the host cell.

40. A method of expressing a polypeptide comprising:
    providing a cell comprising the polynucleotide of Claim 29 or Claim 30; and
    placing the cell under conditions that permit the cell to express the polypeptide encoded by the DNA sequence,
whereby said encoded polypeptide is expressed by said cell.

41. A method of hydrolyzing a carbohydrate comprising:

providing a carbohydrate comprising at least one glycosidic bond;

providing a polypeptide encoded by the polynucleotide of Claim 29 or Claim 30;

and

contacting said carbohydrate with said polypeptide under conditions that permit said polypeptide to hydrolyze at least one glycosidic bond of said carbohydrate,

whereby at least one glycosidic bond of said carbohydrate is hydrolyzed.
Figure 1
Figure 3

A

B
Figure 4

A

B
Figure 5

A

B
Figure 6

A

B
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

INV. C12N9/24 C12P7/10 C12N5/10

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)

C12N C12P

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, WPI Data, EMBASE, MEDLINE, FSTA, COMPENDEX, BIOSIS, Sequence Search

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with Indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>

Further documents are listed in the continuation of Box C.

See patent family annex.

Name and mailing address of the ISA/
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epi nl,
Fax. (+31-70) 340-3016

Date of the actual completion of the International search

22 August 2008

Date of mailing of the International search report

05/09/2008

Authorized officer

Bassi as, Ioannis
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with Indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>DATABASE Geneseq [Online] 25 August 2005 (2005-08-25), &quot;T. reesei cell ohi ohydrolase II (CeI 6A) cDNA.&quot; XP002492982 retrieved from EBI accession no. GSN:AEA49934 Database accession no. AEA49934</td>
<td>-----</td>
</tr>
</tbody>
</table>
Continuation of Box II.2

Claims Nos.: 1-29, 31-41 (all partially)

The present claims 1-29 relate to an extremely large number of possible compounds. Support and disclosure in the sense of Article 6 and 5 PCT is to be found however for only a very small proportion of the compounds, namely those comprising the specific DNA sequences of SEQ ID Nos: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 or 23. The non-compliance with the substantive provisions is to such an extent, that a meaningful search over the whole scope of the claims 1-29 and 31-41 was not possible.

The DNA sequence of interest is defined by a certain percentage of identity to the encoded amino acid sequence whereby at least 3 codon pairs which encode the amino acids 27-471 of the wild-type cell′ohydrolase as set forth in SEQ ID NO: 2 should be replaced by codon pairs which encode identical amino acids or conservative amino acid substitutions. By the definition of up to 16 different specific codons which can be replaced and their possible permutations amongst them, the number of possible codons defining every single amino acid and additionally the different amounts of amino acids belonging to the same functional group ("conservative amino acids") such claims embrace a huge number of possible DNA sequences. The true scope of such claims is thus rendered fully unclear (Article 6 PCT).

The definition of the products furthermore via translational kinetics values and/or z scores for an expression of said DNA in a heterologous host without a restriction to specific hosts is also not clear in the sense of Article 6 PCT. Due to these uncertainties and the very large number of possible compounds being embraced, it is not clear what the true scope of said claims is and how a meaningful search can be performed over the whole possible scope of the claims 1-29. These objections apply also to claims 31-41 which directly or indirectly refer to the subject-matter of claims 1-29.

Only the polynucleotides which comprise the specific DNA sequences of SEQ ID Nos: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 or 23 are clear and sufficiently supported by the description and thus fulfil the requirements of Articles 5 and 6 PCT. Consequently, the search of the application was restricted to said polynucleotides namely cell′ohydrolase-encoding DNA sequences comprising the DNA sequences of SEQ ID Nos: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 or 23 (as defined in claim 30) and to related products (isolated polypeptides encoded by said DNA sequences, expression systems, host cells and systems for degrading cellulose comprising said polynucleotides) as defined in claims 31-38, and to related methods (for introducing a polynucleotide into a host cell, for expressing a polypeptide or for hydrolyzing a carbohydrate) according to claims 39-41.

The applicant’s attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary
examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guideline C-VI, 8.2), should the problems which led to the Article 17(2)PCT declaration be overcome.
INTERNATIONAL SEARCH REPORT

Box No. II Observations where certain claims were found unsearchable (Continuation of Item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
   because they relate to subject matter not required to be searched by this Authority, namely:

2. ☑ Claims Nos. 1-29, 31-41 (al l p a r t i a l l y)
   because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
   see FURTHER INFORMATION sheet PCT/ISA/210

3. ☐ Claims Nos.:
   because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of Invention is lacking (Continuation of Item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers allsearchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☑ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest
- ☐ The additional search fees were accompanied by the applicant’s protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant’s protest but the applicable protest fee was not paid within the time limit specified in the International.
- ☐ No protest accompanied the payment of additional search fees.

Form PCT/ISA/210 *continuation of first sheet (2) (April 2005)
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>US 2007298503 A1</td>
<td>27-12-2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 2007130650 A2</td>
<td>15-11-2007</td>
</tr>
</tbody>
</table>