

**CLAIMS**

1. A computer-implemented method for generating an optimized nucleotide sequence,  
comprising:
- 5
- (i) receiving an amino acid sequence, wherein the amino acid sequence encodes a  
peptide, polypeptide, or protein;
- (ii) receiving a first codon usage table, wherein the first codon usage table  
10 comprises a list of amino acids, wherein each amino acid in the table is  
associated with at least one codon and each codon is associated with a usage  
frequency;
- (iii) removing from the codon usage table any codons associated with a usage  
15 frequency which is less than a threshold frequency;
- (iv) generating a normalized codon usage table by normalising the usage frequencies  
of the codons not removed in step (iii); and
- 20
- (v) generating an optimized nucleotide sequence encoding the amino acid sequence  
by selecting a codon for each amino acid in the amino acid sequence based on  
the usage frequency of the one or more codons associated with the amino acid in  
the normalized codon usage table.
- 25 2. The method according to claim 1, wherein normalising comprises:
- (a) distributing the usage frequency of each codon associated with a first  
amino acid and removed in step (iii) to the remaining codons associated  
with the first amino acid; and
- 30

(b) repeating step (a) for each amino acid to produce the normalized codon usage table.

3. The method according to claim 2, wherein the usage frequency of the removed  
5 codons is distributed equally amongst the remaining codons.

4. The method according to claim 2, wherein the usage frequency of the removed  
codons is distributed amongst the remaining codons proportionally based on the usage  
frequency of each remaining codon.

10

5. The method according to any preceding claim, wherein selecting a codon for each  
amino acid comprises:

(a) identifying, in the normalized codon usage table, the one or more codons  
15 associated with a first amino acid of the amino acid sequence;

(b) selecting a codon associated with the first amino acid, wherein the probability of  
selecting a certain codon is equal to the usage frequency associated with the  
codon associated with the first amino acid in the normalized codon usage table;  
20 and

(c) repeating steps (a) and (b) until a codon has been selected for each amino acid in  
the amino acid sequence.

6. The method according to any preceding claim, wherein step (v) is performed a  
25 plurality of times to generate a list of optimized nucleotide sequences.

7. The method according to any preceding claim, wherein the threshold frequency is  
selectable by a user.

8. The method according to any preceding claim, wherein the threshold frequency is in the range of 5% - 30%, in particular 5%, 10%, or 15%, or 20%, or 25%, or 30%, or, in particular, 10%.
- 5 9. The method according to any one of claims 6 to 8, further comprising:
- screening the list of optimized nucleotide sequences to identify and remove optimized nucleotide sequences failing to meet one or more criteria.
10. The method according to claim 9, wherein screening the list of optimized nucleotide sequences comprises, for each of the one or more criteria:
- 10 determining whether each optimized nucleotide sequence in the list, or most recently updated list, of optimized nucleotide sequences meets the criterion; and
- updating the list of optimized nucleotide sequences by removing any nucleotide sequence from the list, or most recently updated list, if the nucleotide sequence does not meet the criterion.
- 15
11. The method according to claim 10, wherein determining whether each optimized nucleotide sequence in the list, or most recently updated list, of optimized nucleotide sequences meets the criterion comprises, for each nucleotide sequence:
- determining whether a first portion of the nucleotide sequence meets the criterion,
- 20 and wherein updating the list of optimized nucleotide sequences comprises:
- removing the nucleotide sequence if the first portion does not meet the criterion.
12. The method according to claim 11, wherein determining whether each optimized nucleotide sequence in the list, or most recently updated list, of optimized nucleotide sequences meets the criterion further comprises, for each nucleotide sequence:
- 25 determining whether one or more additional portions of the nucleotide sequence meets the criterion, wherein the additional portions are non-overlapping with each other and with the first portion, and wherein updating the list of optimized sequences comprises:

removing the nucleotide sequence if any portion does not meet the criterion, optionally wherein determining whether an optimized nucleotide sequence meets the criterion is halted when any portion is determined not to meet the criterion.

- 5      13.      The method according to claim 11 or 12, wherein the first portion and/or the one or more additional portions of the nucleotide sequence comprise a predetermined number of nucleotides, optionally wherein the predetermined number of nucleotides is in the range of: 5 to 300 nucleotides, or 10 to 200 nucleotides, or 15 to 100 nucleotides, or 20 to 50 nucleotides, e.g., 30 nucleotides, e.g., 100 nucleotides.

10

14.      The method according to any one of claims 9 to 13, wherein a first criterion comprises the nucleotide sequence not containing a termination signal, such that determining and updating comprise:

15                      determining whether each optimized nucleotide sequence in the list, or most recently updated list, of optimized nucleotide sequences contains a termination signal; and

                         updating the list of optimized nucleotide sequences by removing any nucleotide sequence from the list, or most recently updated list, if the nucleotide sequence contains one or more termination signals.

20

15.      The method according to claim 14, wherein the one or more termination signals has/have the following nucleotide sequence:

5'-X<sub>1</sub>ATCTX<sub>2</sub>TX<sub>3</sub>-3',

25

wherein X<sub>1</sub>, X<sub>2</sub> and X<sub>3</sub> are independently selected from A, C, T or G.

16.      The method according to claim 15, wherein the one or more termination signal has/have one or more of the following nucleotide sequences:

30

TATCTGTT; and/or

TTTTTT; and/or  
AAGCTT; and/or  
GAAGAGC; and/or  
TCTAGA.

5

17. The method according to claim 16, wherein the one or more termination signals has/have the following nucleotide sequence:

5'-X<sub>1</sub>AUCUX<sub>2</sub>UX<sub>3</sub>-3',

10

wherein X<sub>1</sub>, X<sub>2</sub> and X<sub>3</sub> are independently selected from A, C, U or G.

18. The method according to claim 17, wherein the one or more termination signals has/have one of the following nucleotide sequences:

15

UAUCUGUU; and/or  
UUUUUU; and/or  
AAGCUU; and/or  
GAAGAGC; and/or  
UCUAGA.

20

19. The method according to any one of claims 9 to 18, wherein a second criterion comprises the nucleotide sequence having a guanine-cytosine content within a predetermined guanine-cytosine content range, such that determining and updating comprise:

25

determining the guanine-cytosine content of each of the optimized nucleotide sequences in the list, or most recently updated list, of optimized nucleotide sequences, wherein the guanine-cytosine content of a sequence is the percentage of bases in the nucleotide sequence that are guanine or cytosine;

30

updating the list of optimized nucleotide sequences by removing any nucleotide sequence from the list, or most recently updated list, if its guanine-cytosine content falls outside the predetermined guanine-cytosine content range.

5

20. The method according to claim 19, wherein the predetermined guanine-cytosine content range is selectable by a user.

21. The method according to claim 19 or 20, wherein the predetermined guanine-cytosine content range is 15% - 75%, or 40% - 60%, or, in particular 30% - 70%.

22. The method according to any one of claims 9 to 21, wherein a third criterion comprises the nucleotide sequence having a codon adaptation index greater than a predetermined codon adaptation index threshold, such that determining and updating  
comprise:

determining the codon adaptation index of each of the optimized nucleotide sequences in the list, or most recently updated list, of optimized nucleotide sequences, wherein the codon adaptation index of a sequence is a measure of codon usage bias and can be a value between 0 and 1;

20 updating the list, or most recently updated list, of optimized nucleotide sequences by removing any nucleotide sequence if its codon adaptation index is less than or equal to the predetermined codon adaptation index threshold.

23. The method according to claim 22, wherein the codon adaptation index threshold is  
25 selectable by a user.

24. The method according to claim 22 or 23, wherein the codon adaptation index threshold is 0.7, or 0.75, or 0.85, or 0.9, or, in particular, 0.8.

30 25. The method according to any one of claims 9 to 24, wherein a fourth criterion comprises the nucleotide sequence not containing at least 2, for example 3, adjacent

identical codons, such that determining and updating comprise:

determining whether any optimized nucleotide sequence in the list, or most recently updated list, of optimized nucleotide sequences, containing at least 2, for example 3 or more, adjacent identical codons; and

5 updating the list, or most recently updated list, of optimized nucleotide sequences by removing any nucleotide sequence if it contains at least 2, for example 3 or more, adjacent identical codons.

26. The method according to claim 25, wherein the fourth criterion is applied only in  
10 respect of codons whose frequency in the normalized codon usage table is less than an adjacency rarity threshold, wherein the adjacency rarity threshold is between 10 and 50%, for example between 15 and 40 %, for example between 20 and 30%.

27. The method according to any preceding claim, wherein the amino acid sequence is  
15 received from a database of amino acid sequences.

28. The method according to claim 26, further comprising requesting the amino acid  
sequence from the database of amino acid sequences, wherein the amino acid sequence is received in response to the request.

20 29. The method according to any preceding claim, wherein the first codon usage table is received from a database of codon usage tables.

30. The method according to claim 29, further comprising requesting the first codon  
25 usage table from the database of codon usage tables, wherein the first codon usage table is received in response to the request.

31. The method according to any preceding claim, further comprising displaying at least  
one optimized nucleotide sequence on a screen.

30

32. A computer program comprising instructions which, when the program is executed by a computer, cause the computer to carry out the method of any preceding claim.
33. A data processing system comprising means for carrying out the method of any preceding claim.
34. A computer-readable data carrier having stored thereon the computer program of claim 32.
35. A data carrier signal carrying the computer program of claim 32.
36. A method for synthesizing a nucleotide sequence, comprising:  
performing the computer-implemented method of any one of claims 1 to 31 to generate at least one optimized nucleotide sequence; and  
synthesizing at least one of the generated optimized nucleotide sequences.
37. The method according to claim 36, wherein the method further comprises inserting the synthesized optimized sequence in a nucleic acid vector for use *in vitro* transcription.
38. The method according to claim 36 or 37, wherein the method further comprises inserting one or more termination signals at the 3' end of the synthesized optimized nucleotide sequence.
39. The method according to claim 38, wherein the one or more termination signals are encoded by the following nucleotide sequence:



wherein  $X_1$ ,  $X_2$  and  $X_3$  are independently selected from A, C, T or G.

40. The method according to claim 38 or 39, wherein the one or more termination signals are encoded by one or more of the following nucleotide sequences:

5                   TATCTGTT;  
                   TTTTTT;  
                   AAGCTT;  
                   GAAGAGC; and/or  
                   TCTAGA.

10   41. The method according to any one of claims 38 to 40, wherein more than one termination signal is inserted, and said termination signals are separated by 10 base pairs or fewer, e.g. separated by 5-10 base pairs.

42. The method according to claim 40, wherein the more than one termination signals are encoded by the following nucleotide sequence: (a) 5'-X<sub>1</sub>ATCTX<sub>2</sub>TX<sub>3</sub>-(Z<sub>N</sub>)-X<sub>4</sub>ATCTX<sub>5</sub>TX<sub>6</sub>-3' or (b) 5'-X<sub>1</sub>ATCTX<sub>2</sub>TX<sub>3</sub>-(Z<sub>N</sub>)-X<sub>4</sub>ATCTX<sub>5</sub>TX<sub>6</sub>-(Z<sub>M</sub>)-X<sub>7</sub>ATCTX<sub>8</sub>TX<sub>9</sub>-3', wherein X<sub>1</sub>, X<sub>2</sub>, X<sub>3</sub>, X<sub>4</sub>, X<sub>5</sub>, X<sub>6</sub>, X<sub>7</sub>, X<sub>8</sub> and X<sub>9</sub> are independently selected from A, C, T or G, Z<sub>N</sub> represents a spacer sequence of N nucleotides, and Z<sub>M</sub> represents a spacer sequence of M nucleotides, each of which are independently selected from A, C, T or G, and wherein N and/or M are independently 10 or fewer.

43. The method according to any one of claims 37 to 42, wherein the nucleic acid vector comprises an RNA polymerase promoter operably linked to the optimized nucleotide sequence, optionally wherein the RNA polymerase promoter is a SP6 RNA polymerase promoter or a T7 RNA polymerase promoter.

44. The method according to any one of claims 37 to 43, wherein the nucleic acid vector comprises a nucleotide sequence encoding a 5' UTR operably linked to the optimized nucleotide sequence.

30   45. The method according to claim 44, wherein the 5' UTR is different to the 5' UTR of a naturally occurring mRNA encoding the amino acid sequence.

46. The method according to claim 42, wherein the 5' UTR has the nucleotide sequence of SEQ ID NO: 16.
- 5 47. The method according to any one of claims 37 to 46, wherein the nucleic acid vector comprises a nucleotide sequence encoding a 3' UTR operably linked to the optimized nucleotide sequence.
- 10 48. The method according to claim 46, wherein the 3' UTR is different to the 3' UTR of a naturally occurring mRNA encoding the amino acid sequence.
49. The method according to claim 48, wherein the 3' UTR has the nucleotide sequence of SEQ ID NO: 17 or SEQ ID NO: 18.
- 15 50. The method according to any one of claims 37 to 49 wherein the nucleic acid vector is a plasmid.
51. The method according to claim 50, wherein the plasmid is linearized before *in vitro* transcription.
- 20 52. The method according to claim 50, wherein the plasmid is not linearized before *in vitro* transcription.
53. The method according to claim 52, wherein the plasmid is supercoiled.
- 25 54. The method according to any one of claims 36 to 53, wherein the method further comprises using at least one of the synthesized optimized nucleotide sequences in *in vitro* transcription to synthesize mRNA.
- 30 55. The method according to claim 54, wherein the mRNA is synthesized by a SP6 RNA polymerase.

56. The method according to claim 55, wherein the SP6 RNA polymerase is a naturally occurring SP6 RNA polymerase.
57. The method according to claim 55, wherein the SP6 RNA polymerase is a  
5 recombinant SP6 RNA polymerase.
58. The method according to claim 57, wherein the SP6 RNA polymerase comprises a tag.
- 10 59. The method according to claim 58, wherein the tag is a his-tag.
60. The method according to claim 54, wherein the mRNA is synthesized by a T7 RNA polymerase.
- 15 61. The method according to any one of claims 54 to 60, wherein the method further comprises a separate step of capping and/or tailing the synthesized mRNA.
62. The method according to any one of claims 54 to 60, wherein capping and tailing occurs during *in vitro* transcription.
- 20 63. The method according to any one of claims 54 to 62, wherein the mRNA is synthesized in a reaction mixture comprising NTPs at a concentration ranging from 1-10 mM each NTP, the DNA template at a concentration ranging from 0.01-0.5 mg/ml, and the SP6 RNA polymerase at a concentration ranging from 0.01-0.1 mg/ml.
- 25 64. The method according to claim 63, wherein the reaction mixture comprises NTPs at a concentration of 5 mM each NTP, the DNA template at a concentration of 0.1 mg/ml, and the SP6 RNA polymerase at a concentration of 0.05 mg/ml.
- 30 65. The method according to any one of claims 54 to 64, wherein the mRNA is synthesized at a temperature ranging from 37-56 °C.

66. The method according to any one of claims 63 to 65, wherein the NTPs are naturally-occurring NTPs.

5 67. The method according to any one of claims 63 to 65, wherein the NTPs comprise modified NTPs.

68. The method according to any one of claims 36 to 67, wherein the method further comprises transfecting the synthesized optimized nucleotide sequence into a cell either *in vitro* or *in vivo*.

10

69. The method according to claim 68, wherein the expression level of the protein encoded by the synthesized optimized nucleotide sequence in transfected cell is determined.

15 70. The method according to claim 68 or 69, wherein the functional activity of the protein encoded by the synthesized optimized nucleotide sequence is determined.

71. The method according to any one of claims 1 to 31, further comprising synthesizing a reference nucleotide sequence encoding the amino acid sequence and the at least one optimized nucleotide sequence according to the method of any one of claims 36 to 70, and  
20 contacting the reference nucleotide sequence and the at least one optimized nucleotide sequence with a separate cell or organism, wherein the cell or organism contacted with the at least one synthesized optimized nucleotide sequence produces an increased yield of the protein encoded by the optimized nucleotide sequence compared to the yield of the protein encoded by the reference nucleotide sequence produced by the cell or organism contacted  
25 with the synthesized reference nucleotide sequence.

72. The method of any one of claims 36 to 70, wherein the method further comprises producing a therapeutic composition comprising an mRNA encoding a therapeutic peptide, polypeptide, or protein for use in the delivery to or treatment of a subject.

30

73. The method of claim 72, wherein the mRNA encodes cystic fibrosis transmembrane conductance regulator (CFTR) protein.

74. The method according to any one of claims 1 to 31, wherein the at least one optimized nucleotide sequence, when synthesized, is configured to increase the expression of the protein encoded by the at least one optimized nucleotide sequence compared to the expression of the protein encoded by the reference nucleotide sequence, when synthesized.

75. The method of any one of claims 71 to 74, wherein the reference nucleotide sequence is (a) a naturally occurring nucleotide sequence encoding the amino acid sequence or (b) a nucleotide sequence encoding the amino acid sequence generated by a method other than the method according to any one of claims 1 to 31.

76. A synthesized optimized nucleotide sequence generated according to the methods of any one of claims 36 to 67 and 72 to 75 for use in therapy.

77. A method of treatment comprising administering the synthesized optimized nucleotide sequence generated according to the method of any one of claims 36 to 67 and 72 to 75 to a human subject in need of such treatment.

78. An *in vitro* synthesized nucleic acid comprising an optimized nucleotide sequence consisting of codons associated with a usage frequency which is greater than or equal to 10%; wherein the optimized nucleotide sequence:

(iv) does not contain a termination signal having one of the following nucleotide sequences:

5'-X<sub>1</sub>AUCUX<sub>2</sub>UX<sub>3</sub>-3', wherein X<sub>1</sub>, X<sub>2</sub> and X<sub>3</sub> are independently selected from A, C, U or G; and 5'-X<sub>1</sub>AUCUX<sub>2</sub>UX<sub>3</sub>-3', wherein X<sub>1</sub>, X<sub>2</sub> and X<sub>3</sub> are independently selected from A, C, U or G;

(v) does not contain any negative cis-regulatory elements and negative repeat elements; and

(vi) has a codon adaptation index greater than 0.8;

wherein, when divided into non-overlapping 30 nucleotide-long portions, each portion of the optimized nucleotide sequence has a guanine cytosine content range of 30% - 70%.

79. The *in vitro* synthesized nucleic acid of claim 77, wherein the optimized nucleotide sequence does not contain a termination signal having one of the following sequences: TATCTGTT; TTTTTT; AAGCTT; GAAGAGC; TCTAGA; UAUCUGUU; UUUUUU; AAGCUU; GAAGAGC; UCUAGA.

5

80. The *in vitro* synthesized nucleic acid of claim 78 or 79, wherein the nucleic acid is mRNA.

81. The *in vitro* synthesized nucleic acid of any one of claims 78 to 80 for use in  
10 therapy.