

things, three objectives: (i) how to allocate the count of synonymous codons of certain amino acid at first, (ii) how to place a synonymous codon into its most suitable location, and (iii) how to reduce the adverse but accidentally generated subsequences and/or motifs. In some examples, these three objectives are quantified as the harmony index, the codon context index, and the outlier index. During optimization, the objectives are considered using a multi-objective algorithm such as the nondominated sorting genetic algorithm III (NSGA-III) or a variant thereof. Specifically, the objectives can be calculated, for a given candidate nucleic acid sequence, with reference to known characteristics of highly-expressed genes. In some examples, various known adverse motifs and/or features (e.g., as identified from literature) are removed from one or more optimized sequences before gene synthesis and protein expression.

[0006] Accordingly, the disclosure provides a systematic method whereby preferably all or most of the parameters and factors affecting protein expression including, but not limited to, codon harmony, codon usage (e.g., synonymous codon distribution), codon context index, cis-acting mRNA destabilizing motifs, RNase splicing sites, GC-content, ribosome binding site (RBS), mRNA secondary structure of the genes (e.g., mRNA free energy), and repetitive element are taken into consideration to improve and optimize the nucleic acid sequences to boost the protein expression of genes in expression systems, such as in expression host cells including both eukaryotic and prokaryotic cells such as mammalian, insect, yeast, bacterial, algal, and in cell-free expression system.

[0007] In one example aspect, there is provided a computer-implemented method for optimizing a nucleic acid sequence for expression of a protein in a host, comprising: a) receiving an initial population set, wherein the initial population set comprises a plurality of initial candidate nucleic acid sequences capable of expressing the protein; and b) performing, based on the initial population set, optimization of a harmony index, a codon context index, and an outlier index using a computer-assisted NSGA-III algorithm or a variant thereof, thereby obtaining a plurality of optimized nucleic acid sequences capable of expressing the protein, wherein the harmony index of a candidate nucleic acid sequence is indicative of consistency of usage frequency distribution of synonymous codons between a plurality of highly expressed genes and the candidate nucleic acid sequence, wherein the codon context index of the candidate nucleic acid sequence is a measure for placing a synonymous codon into a suitable location, and wherein

the outlier index of the candidate nucleic acid sequence is a measure of negative effect of a plurality of predetermined sequence features on the candidate nucleic acid sequence.

[0007a] In one aspect, there is provided a computer-implemented method for optimizing a nucleic acid sequence for expression of a protein in a host, comprising: a) receiving an initial population set, wherein the initial population set comprises a plurality of initial candidate nucleic acid sequences capable of expressing the protein; and b) performing, based on the initial population set, optimization of a harmony index, a codon context index, and an outlier index using a computer-assisted multi-objective genetic algorithm selected from the group consisting of a NSGA-III algorithm, an EliteNSGA-III algorithm, a MAM-MOIA algorithm, or a MOLA algorithm, thereby obtaining a plurality of optimized nucleic acid sequences capable of expressing the protein, wherein the harmony index of a candidate nucleic acid sequence is indicative of consistency of usage frequency distribution of synonymous codons between a plurality of highly expressed genes and the candidate nucleic acid sequence, and is calculated based on a formula: $H = 1 - D(F_{hs}, F_s)$, wherein $D()$ indicates a distance function measuring a distance between two vectors; wherein F_{hs} is a vector comprising frequencies of synonymous codons of a plurality of amino acids within a plurality of highly expressed genes; and wherein F_s is a vector comprising of frequencies of synonymous codons of the plurality of amino acids within a coding gene of the candidate nucleic acid sequence, wherein the codon context index of the candidate nucleic acid sequence is a measure for placing a synonymous codon into a suitable location, and is calculated based on a formula: $CC = 1 - D(F_{hcc}, F_{tcc})$, wherein F_{hcc} is a vector comprising frequencies of synonymous codon pairs of two continual amino acids within a plurality of highly expressed genes; and wherein F_{tcc} is a vector comprising frequencies of synonymous codon pairs of two continual amino acids within a coding gene of the candidate nucleic acid sequence, and wherein the outlier index of the candidate nucleic acid sequence is a measure of a negative effect of a plurality of predetermined sequence features on the candidate nucleic acid sequence, and is calculated based on a formula: $O = \sum_{i=1}^N w_i \times f_i(x)$, wherein N is the number of the plurality of predetermined sequence features; wherein $f_i(x)$ denotes a penalty scoring function of the i th sequence feature of the plurality of predetermined sequence features; and wherein w_i denotes a relative weight associated with $f_i(x)$.

[0008] In some examples, the method further comprises providing an output indicative of at least one optimized nucleic acid sequence of the plurality of optimized nucleic acid sequences.

[0009] In some examples, receiving an initial population set comprises: receiving a protein sequence; generating the initial population set based on the received protein sequence.

[0010] In some examples, receiving an initial population set comprises: receiving a nucleic acid sequence; translating the received nucleic acid sequence into a protein sequence; generating the initial population set based on the protein sequence.

[0011] In some examples, the initial population set is of a predetermined size.

[0012] In some examples, the initial population set includes binary representations of the plurality of initial candidate nucleic acid sequences.

[0013] In some examples, performing optimization of a harmony index, a codon context index, and an outlier index comprises: maximizing the harmony index; maximizing the codon context index; and minimizing the outlier index.

[0014] In some examples, performing optimization of a harmony index, a codon context index, and an outlier index comprises: calculating, for each initial candidate nucleic acid sequence of the initial population set, a respective harmony index value, a respective codon context index value, and a respective outlier index value for a respective initial candidate nucleic acid sequence; based on the calculating, assigning a plurality of fitness values corresponding to the plurality of initial candidate nucleic acid sequences; based on the plurality of fitness values, sorting the plurality of initial candidate nucleic acid sequences; and including a subset of the sorted plurality of initial candidate nucleic acid sequences in a subsequent population set. In some examples, the plurality of fitness values includes the harmony index, the codon context index, and the outlier index for the candidate nucleic acid sequence.

[0042] In one aspect, there is provided a program product stored on a recordable medium for optimizing a nucleic acid sequence for expression of a protein in a host, the program product comprising a computer software for carrying out any of the methods described herein.

[0043] In one example aspect, there is provided an isolated nucleic acid molecule comprising the optimized nucleic acid sequence obtained from any of the methods described herein.

[0044] In one example aspect, there is provided a vector comprising the above-mentioned isolated nucleic acid molecule.

[0045] In one example aspect, there is provided a recombinant host cell comprising the above-mentioned isolated nucleic acid molecule or the above-mentioned vector.

[0046] In one aspect, there is provided a method for expressing a protein in a host cell, the method comprising: (a) obtaining an optimized nucleic acid sequence for expression of the protein in the host cell using any of the methods described herein, (b) synthesizing a nucleic acid molecule comprising the optimized nucleic acid sequence; (c) introducing the nucleic acid molecule into the host cell to obtain a recombinant host cell; and (d) cultivating the recombinant host cell under conditions to allow expression of the protein from the optimized nucleic acid sequence.

DESCRIPTION OF THE FIGURES

[0047] **FIG. 1** depicts a block diagram of an exemplary process for codon optimization, in accordance with some embodiments.

[0048] **FIG. 2A** depicts an exemplary pipeline for constructing and executing an algorithm for optimizing a sequence (e.g., a nucleic acid sequence) for expression of a protein in a host, in accordance with some embodiments.

[0049] **FIG. 2B** depicts an exemplary general workflow of genetic algorithm, in accordance with some embodiments.

[0050] **FIG. 3** depicts Western blot result of optimized GFP and JNK3A1 relative to their wild type, in accordance with some embodiments.

CLAIMS

What is claimed is:

1. A computer-implemented method for optimizing a nucleic acid sequence for expression of a protein in a host, comprising:

- a) receiving an initial population set, wherein the initial population set comprises a plurality of initial candidate nucleic acid sequences capable of expressing the protein; and

- b) performing, based on the initial population set, optimization of a harmony index, a codon context index, and an outlier index using a computer-assisted multi-objective genetic algorithm selected from the group consisting of a NSGA-III algorithm, an EliteNSGA-III algorithm, a MAM-MOIA algorithm, or a MOLA algorithm~~variant thereof~~, thereby obtaining a plurality of optimized nucleic acid sequences capable of expressing the protein,

wherein the harmony index of a candidate nucleic acid sequence is indicative of consistency of usage frequency distribution of synonymous codons between a plurality of highly expressed genes and the candidate nucleic acid sequence, and is calculated based on a formula: $H = 1 - D(F_{hs}, F_{ts})$,

wherein $D()$ indicates a distance function measuring a distance between two vectors;

wherein F_{hs} is a vector comprising frequencies of synonymous codons of a plurality of amino acids within a plurality of highly expressed genes; and

wherein F_{ts} is a vector comprising of frequencies of synonymous codons of the plurality of amino acids within a coding gene of the candidate nucleic acid sequence,

wherein the codon context index of the candidate nucleic acid sequence is a measure for placing a synonymous codon into a suitable location, and is calculated based on a formula: $CC = 1 - D(F_{hcc}, F_{tcc})$,

wherein F_{hcc} is a vector comprising frequencies of synonymous codon pairs of two continual amino acids within a plurality of highly expressed genes; and

wherein F_{tcc} is a vector comprising frequencies of synonymous codon pairs of two continual amino acids within a coding gene of the candidate nucleic acid sequence, and

wherein the outlier index of the candidate nucleic acid sequence is a measure of a negative effect of a plurality of predetermined sequence features on the candidate nucleic acid sequence, and is calculated based on a formula: $O = \sum_{i=1}^N w_i \times f_i(x)$,

wherein N is the number of the plurality of predetermined sequence features;

wherein $f_i(x)$ denotes a penalty scoring function of the i th sequence feature of the plurality of predetermined sequence features; and

wherein w_i denotes a relative weight associated with $f_i(x)$.

2. The method according claim 1, further comprising providing an output indicative of at least one optimized nucleic acid sequence of the plurality of optimized nucleic acid sequences.
3. The method of any of claims 1 and 2, wherein receiving an initial population set comprises:
 - receiving a protein sequence; and
 - generating the initial population set based on the received protein sequence.
4. The method of any of claims 1 and 2, wherein receiving an initial population set comprises:
 - receiving a nucleic acid sequence;
 - translating the received nucleic acid sequence into a protein sequence; and
 - generating the initial population set based on the protein sequence.
5. The method of any of claims 1-4, wherein the initial population set is of a predetermined size.

6. The method according to any of claims 1-5, wherein the initial population set includes binary representations of the plurality of initial candidate nucleic acid sequences.
7. The method of any of claims 1-6, wherein performing optimization of a harmony index, a codon context index, and an outlier index comprises:
 - maximizing the harmony index;
 - maximizing the codon context index; and
 - minimizing the outlier index.
8. The method of any of claims 1-7, wherein performing optimization of a harmony index, a codon context index, and an outlier index comprises:
 - calculating, for each initial candidate nucleic acid sequence of the initial population set, a respective harmony index value, a respective codon context index value, and a respective outlier index value for a respective initial candidate nucleic acid sequence;
 - based on the calculating, assigning a plurality of fitness values corresponding to the plurality of initial candidate nucleic acid sequences;
 - based on the plurality of fitness values, sorting the plurality of initial candidate nucleic acid sequences; and
 - including a subset of the sorted plurality of initial candidate nucleic acid sequences in a subsequent population set.
9. The method of claim 8, further comprising:
 - generating an offspring population based on the initial population; and
 - including the offspring population in the subsequent population set.
10. The method of claim 9, wherein the offspring population is generated via binary tournament selection, crossover/recombination, mutation, or any combination thereof.
11. The method of any of claims 8-10, wherein the initial population set and the subsequent population set are of the same size.

12. The method of any of claims 1-11,

wherein performing optimization of a harmony index, a codon context index, and an outlier index comprises a plurality of iterations,

wherein the i-th iteration of the plurality of iterations comprises:

receiving a population set of nucleic acid sequences corresponding to the (i-1)th iteration;

associating each nucleic acid sequence of the population set corresponding to the (i-1)th iteration with a non-domination level;

sorting the nucleic acid sequences in the population set corresponding to the (i-1)th iteration based on the associated non-domination levels;

generating a population set corresponding to the i-th iteration, wherein the population set corresponding to the i-th iteration includes a subset of the sorted nucleic acid sequences corresponding to the (i-1)th iteration and an offspring population generated based on the sorted nucleic acid sequences corresponding to the (i-1)th iteration; and

determining, based on one or more terminating conditions, whether to proceed to a (i+1)th iteration using the population set corresponding to the i-th iteration.

13. The method of claim 12, wherein associating each nucleic acid sequence with a non-domination level comprises: calculating, for each nucleic acid sequence of the population set corresponding to the (i-1)th iteration, a respective harmony index value, a respective codon context index value, and a respective outlier index value.

14. The method of any of claims 10-11, wherein generating a population set corresponding to the i-th iteration comprises:

associating at least one nucleic acid sequence of the sorted nucleic acid sequence corresponding to the (i-1)th iteration with one of a plurality of predetermined reference points.

15. The method of any of claims 10-12, wherein the one or more terminating conditions includes: a fixed number of iterations reached, best fitness reached a plateau and no better results produced, a minimum criteria of near-optimal solution satisfied by some solutions, or any combination thereof.

~~16. The method according to any of claims 1-15, wherein the harmony index of a candidate nucleic acid sequence is calculated based on a formula: $H = 1 - D(F_{hs}, F_{ts})$,
 — wherein D() indicates a distance function;
 — wherein F_{hs} includes a vector comprising frequencies of synonymous codons of a plurality of amino acids within a plurality of highly expressed genes; and
 — wherein F_{ts} includes a vector comprising of frequencies of synonymous codons of the plurality of amino acids within a coding gene of the candidate nucleic acid sequence.~~

~~17. The method according to claim 16, wherein D() indicates a function measuring a distance between two vectors.~~

~~18.16.~~ The method of any of claims 1-15-17, wherein D() is a distance function that comprises aincludes, but is not limited to: Euclidean distance, a Cosine distance, a Manhattan distance, or a Minkowski distance betweenof two vectors.

~~19.17.~~ The method according to any of claims 1-1618, wherein a frequency of a synonymous codon of the plurality of highly expressed genes or a candidate nucleic acid sequence is defined as: $F_{sij} = \frac{\text{total occurancy of synonymous codon } j}{\text{total occurancy of amino acid } i}, \forall i \in \{A, C, D, E, F, G, H, I, K, L, N, P, Q, R, S, T, V, Y\} \text{ and } \exists j \in 59 \text{ synonymous codons.}$

~~20. The method according to any of claims 1-19, wherein the codon context index of a candidate nucleic acid sequence is calculated based on a formula: $CC = 1 - D(F_{hcc}, F_{tcc})$,
 — wherein D() indicates a distance function;
 — wherein F_{hcc} comprises a vector comprising frequencies of synonymous codon pairs of two continual amino acids within a plurality of highly expressed genes; and~~

—wherein F_{tee} comprises a vector comprising frequencies of synonymous codon pairs of two continual amino acids within a coding gene of the candidate nucleic acid sequence;

21. The method according to claim 20, wherein $D()$ indicates a function measuring a distance between two vectors.

22. The method of claim 21, wherein $D()$ is a distance function that includes, but is not limited to: Euclidean distance, a Cosine distance, a Manhattan distance, or a Minkowski distance of two vectors.

23.18. The method according to any of claims ~~1-17~~20-22, wherein a frequency of a synonymous codon pair of the plurality of highly expressed genes or a candidate nucleic acid sequence is defined as: $F_{ccij} = \frac{\text{total occurance of synonymous codon pair } j}{\text{total occurance of amino acid pair } i}, \forall i \in \text{the permutation of two amino acids besides MM, MW, WW and WM}; \exists j \in 3717 \text{ codon pairs.}$

24. The method according to any of claims 1-23, wherein the outlier index is calculated based on a formula: $O = \sum_{i=1}^N w_i \times f_i(x)$,

—wherein N is the number of the plurality of predetermined sequence features;

—wherein $f_i(x)$ denotes a penalty scoring function of the i th sequence feature of the plurality of predetermined sequence features; and

—wherein w_i denotes a relative weight associated with $f_i(x)$.

25.19. The method according to any of claims ~~1-18~~24, wherein the plurality of predetermined sequence features includes:

GC-content value,

CIS elements,

repetitive elements,

RNA splicing sites,

ribosome binding sequences,

minimal free energy of mRNA, or
any combination thereof.

26.20. The method according to claim 1924, wherein the plurality of predetermined sequence features is identified based on a selected expression system.

~~27. The method according to any of claims 1-26, wherein a variant of the NSGA-III algorithm includes the EliteNSGA-III algorithm or a NSGA-II based immune algorithm.~~

28.21. The method according to any of claims 1-~~27~~20, wherein performing optimization of a harmony index, a codon context index, and an outlier index comprises:
ranking the plurality of optimized nucleic acid sequences by descending order of harmony index, then by descending order of codon context index, and then by ascending order of outlier index; and
selecting one or more top-ranked optimized nucleic acid sequences for synthesis.

29.22. The method according to any of claims 1-21~~28~~, further comprising:
c) removing a predetermined adverse subsequence or motif from an optimized nucleic acid sequence of the plurality of optimized nucleic acid sequences.

30.23. The method according to claim 22~~29~~, wherein the predetermined adverse subsequence or motif is identified based on analysis of a plurality of text portions.

31.24. The method according to claim 22~~29~~, wherein removing the predetermined adverse subsequence or motif comprises:
identifying the predetermined adverse subsequence or motif in the optimized nucleic acid sequence;
identifying a plurality of synonymous codons based on identified predetermined adverse subsequence or motif; and

selecting a synonymous codon from the plurality of synonymous codons for substitution with the identified predetermined adverse subsequence in the optimized nucleic acid sequence.

~~32.25.~~ The method according to any of claims 1-~~2431~~, wherein at least one of the harmony index, the codon context index, and the outlier index is calculated based on one or more characteristics of a plurality of highly-expressed genes from one or more databases.

~~33.26.~~ The method according to claim ~~2532~~, wherein the one or more characteristics include codon frequency, synonymous codon frequency, codon pair frequency, or a combination thereof.

~~34.27.~~ The method according to any of claims 1-~~2633~~, further comprising: setting one or more parameters, wherein the one or more parameters include a size of a population set, a number of divisions, a distribution index for simulated binary crossover, a crossover rate for simulated binary crossover, a mutation rate for bit flip mutation, a distribution index for bit flip mutation, or any combination thereof.

~~35.28.~~ A non-transitory computer-readable storage medium storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of an electronic device, cause the electronic device to carry out the methods of any one of claims 1-~~2734~~.

~~36.29.~~ A system for optimizing a nucleic acid sequence for expression of a protein in a host, the system comprising:
one or more processors;
a memory; and
one or more programs, wherein the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for carrying out the method of any one of claims 1-~~2734~~.

~~37.~~30. An electronic device for optimizing a nucleic acid sequence for expression of a protein in a host, the device comprising means for carrying out the method of any one of claims 1-~~27~~34.

~~38.~~31. A program product stored on a recordable medium for optimizing a nucleic acid sequence for expression of a protein in a host, the program product comprising a computer software for carrying out the methods of any one of claims 1-~~27~~34.

~~39. An isolated nucleic acid molecule comprising the optimized nucleic acid sequence obtained from the method of any one of claims 1-34.~~

~~40. A vector comprising the isolated nucleic acid molecule of claim 39.~~

~~41. A recombinant host cell comprising the isolated nucleic acid molecule of claim 39 or the vector of claim 40.~~

~~42.~~32. A method for expressing a protein in a host cell, the method comprising:

- (a) obtaining an optimized nucleic acid sequence for expression of the protein in the host cell using a method of any one of claim 1-~~27~~34;
- (b) synthesizing a nucleic acid molecule comprising the optimized nucleic acid sequence;
- (c) introducing the nucleic acid molecule into the host cell to obtain a recombinant host cell; and
- (d) cultivating the recombinant host cell under conditions to allow expression of the protein from the optimized nucleic acid sequence.