Abstract: The invention relates to methods useful for optimizing expression of a foreign gene in an organism or tissue of interest. In particular, the invention relates to methods for ranking genes expressed in an organism or tissue of interest and methods for modifying nucleotide sequences for particular uses by making changes in codon usage to modulate expression of a gene. In one aspect, the invention provides a method for optimizing the nucleotide sequence for expression in a specific organism or tissue by recursively overlapping the nucleotide sequence of a target gene with the nucleotide sequences of ranked genes in a tissue of interest.

In another aspect, the method provides for the construction of an optimized version of the target gene which retains the original amino acid coding information of the target gene.
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published: without international search report and to be republished upon receipt of that report.
SEQUENCE OPTIMIZATION FOR EXPRESSION OF A FOREIGN GENE

[001] This application claims priority to U.S. Provisional Application No. 60/949,349 filed on July 12, 2007 which is hereby incorporated by reference in its entirety.

[002] All patents, patent applications and publications cited herein are hereby incorporated by reference in their entirety. The disclosures of these publications in their entireties are hereby incorporated by reference into this application in order to more fully describe the state of the art as known to those skilled therein as of the date of the invention described herein.

BACKGROUND OF THE INVENTION

[003] Nucleotide sequences contain a wealth of information in addition to the information needed to encode proteins. For example, genomic nucleotide sequences contain transcription factor binding sites, restriction enzyme binding sites, splicing signals, mRNA stability signals, and the like. Many of these regulatory mechanisms have been proposed to function in tissue specific manner. It is also known that levels of tRNAs vary between tissues of a given organism and that the relative tRNA abundance correlates to the codon usage of subsets of highly expressed genes in those tissues (Dittmar et al, 2006). Similarly, cis acting elements encoded in genes may function in a host dependent manner. It is possible that cellular regulators of these nucleotide sequences may differ in their abundance or availability in different tissues thereby altering the functional consequence of the presence of such cis acting elements encoded in a given gene. In order to modulate the level of expression of a given gene in a given tissue without changing the protein coding content of the gene, these cis acting nucleotide sequences can be optimized for gene expression in an organism or tissue specific manner. The present invention addresses this need.

SUMMARY OF THE INVENTION

[004] The present invention provides methods useful for optimizing expression of a foreign gene in an organism or tissue of interest. The present invention also provides methods for ranking genes expressed in an organism or tissue of interest and methods for modifying nucleotide sequences for particular uses by making changes in codon usage to modulate expression of a gene.
[005] The present invention also provides methods for optimizing a the nucleotide sequence for expression in a specific organism or tissue by recursively overlapping the nucleotide sequence of a target gene with the nucleotide sequences of ranked genes in a tissue of interest. In one embodiment, the method provides for the construction of an optimized version of the target gene which retains the original amino acid coding information of the target gene.

[006] The present invention provides a method for optimizing a nucleotide sequence for expression in an organism or tissue of interest, the method comprising, (i) providing a nucleotide sequence “b”, which encodes an amino acid sequence “s”, for optimization, (ii) providing a group of one or more genes “tn” from the organism or tissue of interest use as templates to optimize the nucleotide sequence, (iii) ranking the group of one or more genes in “tn” according to their expression level in the organism or tissue of interest to create a group of ranked genes “t”, (iv) selecting an amino acid word “w0” of length “L” from the amino acid sequence “s”, (v) sequentially examining the ranked genes from group “t” for a nucleotide sequence that can potentially encode the amino acid word “w0” without regard to the translation frame, (vi) eliminating the C-terminal amino acid from the polypeptide sequence “w0” and adding the next N-terminal amino acid in the corresponding amino acid sequence in “s” to generate amino acid sequence “w1”, (vii) eliminating the three nucleotides at the 5’ end of the nucleotide sequence “n0” to generate a new nucleotide sequence “n01”, (viii) replacing the nucleotide sequence of “n0” with the nucleotide sequence of “n1”, and (ix) recursively repeating the steps (v) to (viii) until the end of the amino acid sequence “s”, thereby generating the optimized nucleotide sequence “o”. In one embodiment, the nucleotide sequence “b” can encode a therapeutic protein. In another embodiment, the nucleotide sequence “b” can encode an immunogenic protein. In other embodiments, the nucleotide sequence “b” can be selected from the group consisting of: the genome a eukaryotic organism, the genome of a prokaryotic organism, the genome of a virus, an expression vector sequence, a plasmid sequence, a cloned cDNA sequence, and the an expressed sequence tag (EST). Hypothetical sequences are used in a schematic illustration of one embodiment of the above method in Figure 1.

[007] Various modifications of the above method are contemplated. For example in some embodiments, 1 to 20 genes can be ranked in step (iii). In other embodiments, 21 to 100 genes, 101 to 10000 genes or more than 10000 genes can be ranked in step (iii). In yet other
embodiments, the word length “L” can be two amino acids or the word length “L” can be three amino acids in length or the word length “L” can be four amino acids in length or the word length “L” can be five amino acids in length or the word length “L” can be six amino acids in length or the word length “L” can be seven amino acids in length or the word length “L” can be eight amino acids in length or the word length “L” can be nine amino acids in length. In other embodiments the word length “L” is ten or more amino acids in length.

[008] In yet further embodiments the word length “L” is not a fixed value and is established by recursively scanning the group of ranked nucleotide sequences in group “t” for the longest possible “word size” by beginning with the longest possible “word size” in the amino acid sequence “s” and shortening the “word size” of “w0” by removing the C-terminal amino acid of “w0” recursively until a nucleotide sequence in the group of ranked nucleotide sequences “t” is identified.

[009] The optimized nucleotide sequence “o” can contain one or more mRNA stability signals, signals that increase the rate of transcription, signals that increase protein translation, protein binding sites, transcription factor binding sites, promoter sequences, enhancer sequences, or splice sites not present in the starting nucleotide sequence “b”. In some embodiments, the optimized nucleotide sequence “o” differs from the nucleotide sequence “b” by one or more of the sequence motifs identified is an mRNA stability signal, an mRNA instability signal, a signal that increases the rate of transcription, a signal that decreases the rate of transcription, a signal involved in protein translation, a protein binding site, a transcription factor binding site, a promoter sequence, an enhancer sequence, a repressor sequence, a silencer sequence, a splice site, a restriction enzyme site, or a viral latency signal.

[0010] In some embodiments, step (ix) is repeated one time. In other embodiments, step (ix) is repeated multiple times. In yet other embodiments, step (ix) is repeated until the end of the amino acid sequence “s” is reached. In yet further embodiments, steps (i) through (ix) are performed multiple times with different ranked genes or gene rankings or criteria for generating the group “t”.

[0011] In other embodiments of the invention, the genes in group “t” can comprise genes derived from the same host whereas in other embodiments the genes in group “t” comprise
genes derived different hosts. In yet further embodiments, the genes in group “t” can comprise artificial genes, recombinant genes or theoretical genes.

[0012] In some embodiments of the above method, the ranking can be based on experimental data, on information available in a repository or on a theoretical model.

[0013] In other variations of the above method, the method is performed iteratively with different genes in group “t” to generate multiple nucleotide sequence “o”. In other embodiments, multiple nucleotide sequences “o” can be used to generate a consensus nucleotide sequence. Methods for generating a single consensus nucleotide sequence from multiple nucleotide sequences on the basis of sequence homology are readily known to those skilled in the art.

[0014] In other embodiments of the invention, the genes in group “t” can be ranked on a set of one or more criteria comprising gene expression, mRNA abundance, mRNA steady state levels, mRNA stability, mRNA export, mRNA translation, mRNA localization, mRNA transcription, mRNA splicing, mRNA dicing, mRNA secondary structure, mRNA tertiary structure, mRNA binding to agents, mRNA binding to proteins, mRNA binding to other RNAs, mRNA binding to DNA, sequence homology, phylogenetic analysis, the presence or absence of sequence motifs, RNA or DNA topology, RNA or DNA architecture, protein expression, the organization of sequence motifs, binding to transcription factors, binding to enhancers, binding to repressors, specificity of expression, expression patterns, functional significance, biological significance, developmental significance, pathological significance, disease significance, infection significance, virulence, infectivity, replication, gene function or any combination thereof.

[0015] In some embodiments, the genes in group “t” can comprise genes derived from the same host and in other embodiments of the method described above, the genes in group “t” comprise genes derived from different hosts.

[0016] In another aspect, the invention provides a method for optimizing the production of a protein in an organism or tissue of interest, the method comprising, (a) selected a nucleotide sequence to be optimized, and (b) optimizing the nucleotide sequence using the method of method comprising, (i) providing a nucleotide sequence “b”, which encodes an amino acid sequence “s”, for optimization, (ii) providing a group of one or more genes “tn” from the organism or tissue of interest use as templates to optimize the nucleotide sequence, (iii)
ranking the group of one or more genes in “tn” according to their expression level in the organism or tissue of interest to create a group of ranked genes “t”, (iv) selecting an amino acid word “w0” of length “L” from the amino acid sequence “s”, (v) sequentially examining the ranked genes from group “t” for a nucleotide sequence that can potentially encode the amino acid word “w0” without regard to the translation frame, (vi) eliminating the C-terminal amino acid from the polypeptide sequence “w0” and adding the next N-terminal amino acid in the corresponding amino acid sequence in “s” to generate amino acid sequence “w1”, (vii) eliminating the three nucleotides at the 5’ end of the nucleotide sequence “n0” to generate a new nucleotide sequence “n01”, (viii) replacing the nucleotide sequence of “n0” with the nucleotide sequence of “n1”, and (ix) recursively repeating the steps (v) to (viii) until the end of the amino acid sequence “s”, thereby generating the optimized nucleotide sequence “o”. In one embodiment, the protein can be a therapeutic protein or an immunogenic protein.

[0017] In a further aspect, the invention provides an optimized HIV or lentivirus sequence for use in an HIV or lentivirus vaccine composition wherein the HIV or lentivirus sequence comprises a sequence selected from the group consisting of optimized Gag sequence of SEQ ID NO: 1.

[0018] In yet another aspect, the invention provides a method for optimizing the production of a protein in a host, the method comprising, (a) obtaining a nucleotide sequence encoding a protein to be expressed in the host, (b) optimizing the sequence of the nucleotide sequence to result in improved production of the protein in the host wherein the method comprises, (i) selecting a nucleotide sequence “b”, which encodes an amino acid sequence “s”, for optimization, (ii) selecting a group of one or more genes “tn” from the organism or tissue of interest use as templates to optimize the nucleotide sequence, (iii) ranking the group of one or more genes in “tn” according to their expression level in the organism or tissue of interest, (iv) selecting an amino acid word “w0” of length “L” from the amino acid sequence “s”, (v) sequentially examining the ranked genes from group “t” for a nucleotide sequence that can potentially encode the amino acid word “w0” without regard to the translation frame, (vi) eliminating the C-terminal amino acid from the polypeptide sequence “w0” and adding the next N-terminal amino acid in the corresponding amino acid sequence in “s” to generate amino acid sequence “w1”, (vii) eliminating the three nucleotides at the 5’ end of the nucleotide sequence “n0” to generate a new nucleotide sequence “n01”, (viii) replacing the nucleotide sequence of “n0” with the nucleotide sequence of “n1”, and (ix) recursively
repeating the steps (v) to (viii) until the end of the amino acid sequence “s”, thereby generating the optimized nucleotide sequence “o”.

[0019] In one embodiment, the word size used is two or more amino acids. In another embodiment, the nucleotide sequence “b” selected from the group consisting of: the genome of a eukaryotic organism, the genome of a prokaryotic organism, the genome of a virus, an expression vector, a plasmid, a cloned cDNA, and an expressed sequence tag (EST). In yet another embodiment, the genes in group “t” are selected from the group consisting of: the genome of a eukaryotic organism, the genome of a prokaryotic organism, the genome of a virus, an expression vector, a plasmid, a cloned cDNA, and an expressed sequence tag (EST). In yet another embodiment, the amino acid sequence encoded by the nucleotide sequences “b” and “o” are identical. In still another embodiment, the nucleotide sequences in step (v) exist in the same translation frame and actually encode the same amino acid sequence word “w0”. In another embodiment, the steps (i) through (ix) are independently repeated for a nucleotide sequence “b” using two or more ranking criteria or with different genes in group “t” to generate multiple optimized nucleotide sequences “o” and the multiple optimized nucleotide sequences “o” are used to generate a consensus sequence “o”.

[0020] In some embodiments, the protein subject to optimization may be a therapeutic protein, an immunogenic protein or a protein that may be suitable for use in a vaccine composition. In another embodiment, the nucleotide sequence encoding the protein is located in, or may be inserted into, a vector. In yet further embodiments, the vector may be an expression vector or an expression vector may be adapted for administration to the host as a vaccine. In other embodiments, the vector may be a viral vector or viral vector may be adapted for administration to the host as a vaccine.

[0021] In one embodiment, the nucleotide sequence encoding the protein is located in, or may be inserted into, a recombinant virus. In some embodiments, the recombinant virus is adapted for administration to the host as a vaccine or the recombinant virus is an attenuated virus. In further embodiments, the host is a eukaryote or a eukaryotic cell, the host is a prokaryote or a prokaryotic cell, the host is a bacterium, the host is a yeast cell, the host is a mammal or a mammalian cell, the host is a primate or a primate cell, the host is a human or a human cell, the host is a mouse or a mouse cell, the host is a goat or a goat cell, the host is a sheep or a sheep cell, the host is a bird or a bird cell, the host is a chicken or a chicken cell, the host is an insect or an insect cell, the host is a transgenic animal or a cell from a
transgenic animal or the host is a cell from a cultured cell line. In further embodiments, the
cell line is selected from the group consisting of: a Chinese hamster ovary (CHO) cell line,
the mouse myeloma NS0 cell line, a baby hamster kidney (BHK) cell line, the human embryo
kidney 293 (HEK-293) cell line, a chicken embryo fibroblast cell line, the human C6 cell
line, a Madin-Darby canine kidney (MDCK) cell line, and the Sf9 insect cell line.

[0022] In another aspect, the invention provides a method for identifying an agent that affects
viral RNA production, viral protein, viral particle production or inhibition or stimulation of
viral latency, the method comprising, (a) providing a control cell containing at least one non-
optimized viral nucleic acid sequence and a test cell containing at least one optimized viral
nucleic acid sequence, (b) contacting the control cell and the test cell with one or more
agents, (c) measuring viral RNA production, viral protein, viral particles production or
inhibition or stimulation of viral latency, and (d) comparing the measured viral RNA
production, viral protein, viral particles production or inhibition or stimulation of viral
latency in the test cell and the control cell and identifying at least one agent that affects viral
RNA production, viral protein, viral particles production or inhibition or stimulation of viral
latency, wherein an increase or decrease in viral RNA production, viral protein, viral
particles production or inhibition or stimulation of viral latency indicates that the agent
affects viral RNA production, viral protein, viral particles production or inhibition or
stimulation of viral latency.

[0023] In another aspect, the invention provides a method for identifying an agent that binds
differentially to an non-optimized nucleotide sequence and a corresponding optimized
nucleotide, the method comprising, (a) providing at least one non-optimized viral nucleic acid
sequence and a least one optimized viral nucleic acid sequence, (b) contacting the non-
optimized nucleotide sequence and the optimized nucleotide sequence with one or more
agents, (c) measuring binding of the agent to the optimized and non-optimized nucleic acid
sequences, and (d) comparing the measured binding of the agent to the optimized and non-
optimized nucleic acid sequences in and identifying at least one agent that binds differentially
to the optimized and non-optimized nucleic acid sequences, wherein an increase of decrease
in binding of the agent with the optimized nucleic acid sequence compared to the non-
optimized nucleic acid sequence indicates that the agent binds differentially to the optimized
and non-optimized nucleic acid sequences. In yet another aspect, the invention provides for
an agent that inhibits or stimulates binding of an agent to an optimized nucleic acid sequence identified using any of the methods of the invention.

[0024] In yet another aspect, the invention provides a method for identifying changes in protein characteristics that arise as a result of the introduction of silent mutations upon optimization of a nucleic acid encoding the protein, the method comprising, (a) providing a protein translated by a from a non-optimized nucleic acid sequence encoding the protein and a protein translated from an optimized nucleic acid sequence, (b) measuring the characteristics of the protein translated from the non-optimized protein sequence and the characteristic of the protein translated from the optimized protein sequence, and (c) comparing the measured characteristics of the protein translated from the non-optimized protein sequence and the characteristic of the protein translated from the optimized protein sequence, wherein a difference in the characteristic of the protein translated from the non-optimized protein sequence and the characteristic of the protein translated from the optimized protein sequence indicates that the introduction of silent mutations during optimization of a nucleic acid change protein characteristics. In one embodiment, of the above aspect, the protein characteristics include: tertiary structure, immunogenicity, incorporation into a viral structure, protein expression, protein stability, enzymatic activity, polymerase activity, transcriptase activity, protein binding, small molecule binding, binding to an agent or protein localization.

BRIEF DESCRIPTION OF THE FIGURES

[0025] Figure 1 shows the nucleotide sequence of the HIV-1 Gag gene prior to optimization (SEQ ID NO:1).

[0026] Figure 2 shows the nucleotide sequence of the HIV-1 Gag gene after optimization (SEQ ID NO:2).

[0027] Figure 3 shows a schematic illustration of the steps of one embodiment of the invention is shown. Hypothetical sequences are illustrated for clarity. The nucleotide and amino acid sequences shown in this figure are not intended to be limiting and are not intended to actually correspond to any known biological sequences. The schematic figure is for illustrative purposes only. The steps are divided among two figures. Figure 3A shows steps (i) to (v) and Figure 3B shows steps (vi) to (ix). The steps outlined are (i) providing a nucleotide sequence “b”, which encodes an amino acid sequence “s”, for optimization, (ii)
providing a group of one or more genes “tn” from the organism or tissue of interest use as templates to optimize the nucleotide sequence, (iii) ranking the group of one or more genes in “tn” according to their expression level in the organism or tissue of interest, (iv) selecting an amino acid word “w0” of length “L” from the amino acid sequence “s”, (v) sequentially examining the ranked genes from group “t” for a nucleotide sequence that can potentially encode the amino acid word “w0” without regard to the translation frame, (vi) eliminating the C-terminal amino acid from the polypeptide sequence “w0” and adding the next N-terminal amino acid in the corresponding amino acid sequence in “s” to generate amino acid sequence “w1”, (vii) eliminating the three nucleotides at the 5’ end of the nucleotide sequence “n0” to generate a new nucleotide sequence “n01”, (viii) replacing the nucleotide sequence of “n0” with the nucleotide sequence of “n1”, and (ix) recursively repeating the steps (v) to (viii) until the end of the amino acid sequence “s”, thereby generating the optimized nucleotide sequence “o”.

DETAILED DESCRIPTION OF THE INVENTION

Definitions

[0028] The singular forms "a," "an," and "the" include plural references unless the content clearly dictates otherwise. Thus, for example, reference to a "virus" includes a plurality of such viruses.

[0029] The term “word”, is used herein to define any string of two or more amino acids in an amino acid sequence. For example, certain embodiments of the invention involve selecting a word size before applying further calculations described herein to optimize the nucleotide sequence of a given gene.

[0030] The terms “rank”, “ranking” and “ranked” refer to the process of categorizing, grading or conferring a hierarchal order upon genes selected as templates for the method of optimization described herein. The characteristics used for ranking include, but are not limited to, gene expression levels, protein expression levels, biological significance, functional significance and any other characteristics known to those skilled in the art.

[0031] The terms “optimization”, “optimize” and “optimized”, as used herein refer to any mutation, modification or alteration of a nucleic acid sequence by the methods described herein. These terms are not limiting and do not refer explicitly to modifications that increase
the level of expression of a gene. In on embodiment, optimization increases the level of expression of a gene in a given organism, tissue, cell type or cell.

[0032] As used herein, the term "homolog" refers to a nucleotide sequence sharing at least about 70%, 80%, 90% or more identity with the nucleotide sequences referred to herein, such as the wild-type viral nucleotide sequences referred to herein.

[0033] The term "homolog" is also used to refer to proteins with amino acid sequences, or nucleic acids having nucleotide sequences that exhibit a certain percent sequence identity with the nucleotide sequence of a reference nucleic acid, or the amino acid sequence of a reference protein, such as the viral proteins referred to herein or the wild-type nucleic acid sequences referred to herein. In certain embodiments, a homolog exhibits at least about 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or more identity with a reference nucleic acid or protein sequence. In some embodiments, homologs of the proteins described herein have a substantially similar structure and/or function and/or immunogenicity to the wild type viral proteins described herein.

[0034] The term "organism", as used herein, includes all multicellular and unicellular life forms such as for example, animals or animal cells, plants or plant cells, bacteria, fungi, yeasts, protozoans, protists and the like. The term “organism” also includes any living structure that contains nucleic acid and is capable of reproduction. Unless stated otherwise, the term “organism” as used herein should also be construed to encompass viruses. Unless stated otherwise, the term “organism” may be used interchangeable with the term “host”

[0035] As used herein, a "virus" includes any infectious particle having a protein coat surrounding an RNA or DNA core of genetic material. The term “virus”, as used herein, also refers to all strains, isolates, and clades of all DNA and RNA viruses. Viruses include, but are not limited to all Adenoviruses, Alfamoviruses, Allyxviruses, Alloviruses, Alphacryptoviruses, Alphalipotrichaviruses, Alphanodaviruses, Alphapapillomaviruses, Alpharetroviruses, Alphaherpesviruses, Amdoviruses, Ampeloviruses, Aphthoviruses, Aquabirnaviruses, Aquareoviruses, Arenaviruses, Arteriviruses, Ascoviruses, Asfivirus, Avagenoviruses, Avianadenoviruses, Avibirnaviruses, Avihepadnaviruses, Avipoxviruses, Avulaviruses, Babuviruses, Badnaviruses, Barnaviruses, Bdellomicroviruses, Begomoviruses, Betacryptoviruses, Betapapillomaviruses, Betaretroviruses,
Spiromicroviruses, Spumaviruses, Suipoxviruses, Tectiviruses, Teschoviruses, Thetapapillomaviruses, Thogotoviruses, Tombusviruses, Topocuviruses, Toroviruses, Tospoviruses, Totiviruses, Trichoviruses, Tritimoviruses, Tungroviruses, Tymoviruses, Varicelloviruses, Vesiculoviruses, Vesiviruses, Vitiviruses, Waikaviruses, Whispoviruses, Xipapillomaviruses, Yatapoxviruses, Zetapapillomaviruses or any combination thereof.

[0036] The term “retrovirus”, as used herein, refers to all strains, isolates, and clades of all retroviruses including, but not limited to all alpharetroviruses, betaretroviruses, deltaretroviruses, epsilonretroviruses, gammaretroviruses, spumaviruses, and lentiviruses.

[0037] The term “lentivirus”, as used herein, refers to all strains, isolates, and clades of all lentiviruses, including but not limited to, bovine immunodeficiency viruses, equine infectious anemia viruses (EIAV), feline immunodeficiency viruses (FIV), caprine arthritis encephalitis viruses, visna/maedi viruses, type 1 human immunodeficiency viruses (HIV-1), type 2 human immunodeficiency viruses (HIV-2) and simian immunodeficiency viruses (SIV).

[0038] The term “HIV”, as used herein refers to all strains, isolates, and clades of both HIV-1 and HIV-2. Thus, unless stated otherwise, when the term HIV is used without specifying a type (i.e. without specifying type 1 or type 2) it is to be assumed that both HIV-1 and HIV-2 are referred to, including all strains, isolates, and clades of HIV-1 and HIV-2.

[0039] The terms “protein” and “peptide”, as used herein, refer to polymeric chain(s) of amino acids. Although the term “peptide” is generally used to refer to relatively short polymeric chains of amino acids, and the term “protein” is used to refer to longer polymeric chain of amino acids, there is some overlap in terms of molecules that can be considered proteins and those that can considered peptides. Thus, the terms “protein” and “peptide” may be used interchangeably herein, and when such terms are used they are not intended to limit in anyway the length of the polymeric chain of amino acids referred to. Unless otherwise stated, the terms “protein” and “peptide” should be construed as encompassing all fragments, derivatives, variants, homologs, and mimetics of the specific proteins mentioned, and may comprise naturally occurring amino acids or synthetic amino acids.

[0040] The terms “vaccine” and “immunogenic composition” are used interchangeably herein to refer to agents or compositions capable of inducing an immune response against a virus.
[0041] In a one embodiment, the present invention provides vaccines capable of inducing an immune response against a lentivirus such as HIV-1, HIV-2, SIV, FIV or EIAV. The terms “vaccine” and “immunogenic composition” encompass prophylactic or preventive vaccines and therapeutic vaccines. The vaccine compositions of the invention may also be cross-reactive with, and effective against, multiple different viruses. For example, the immunogenic compositions of the invention may be cross-reactive with, and effective against, multiple different types of lentivirus and/or multiple different types of immunodeficiency virus. Similarly, the immunogenic compositions of the invention may be cross-reactive between different strains and clades of the same virus. For example, an immunogenic composition according to the present invention that is effective against one strain of HIV or lentivirus may also be effective against multiple strains of HIV or lentivirus. A prophylactic vaccine is one administered to subjects who are not infected with the pathogenic agent against which the vaccine is designed to protect. For example, a prophylactic vaccine will prevent a pathogenic agent from establishing an infection in a vaccinated subject, i.e. it will provide complete protective immunity. However, even if it does not provide complete protective immunity, a prophylactic vaccine may still confer some protection to a subject. For example, a prophylactic vaccine may decrease the symptoms, severity, and/or duration of a disease caused by a pathogenic agent. A therapeutic vaccine, is administered to reduce the impact of an infection in a subject already infected with a pathogenic agent. A therapeutic vaccine may decrease the symptoms, severity, and/or duration of a disease caused by a pathogenic agent.

[0042] As used herein the terms “protein vaccine”, “proteinaceous vaccine” and “subunit vaccine” are used interchangeably to refer to vaccines that contain a viral protein component.

[0043] The term “therapeutic protein” is used herein to refer to a protein that, when administered to a subject, is useful for the treatment, amelioration, or prevention of a disease or disorder. The term “immunogenic protein” is used herein to refer to a protein that, when administered to a subject, is capable of stimulating an immune response.

[0044] The term “agent”, as used herein, is used generically to refer to any molecule, such as a protein, peptide, or pharmaceutical, including but not limited to, agents that bind to optimized nucleic acid sequences, agents that inhibit or stimulate binding of another agent to an optimized nucleic acid sequence, vaccines that contain or are made from optimized nucleic acids, molecules that are co-administered with the vaccines of the invention, and the like.
[0045] The term “subject” as used herein, refers to any animal to whom a vaccine or agent according to the present invention may be administered, including humans and other mammalian species.

[0046] The terms “pathogen”, “pathogenic agent” and “infectious agent” are used interchangeably herein to encompass, inter alia, bacteria, viruses (including bacteriophages), fungi, yeast, protozoans (such as the malaria parasite), protists, and prions (such as the prions that cause transmissible spongiform encephalopathies such as Creutzfeldt-Jakob disease).

[0047] The term “host” refers to any organism or any cell (including, but not limited to animals, animal cells, plants, plant cells, bacteria and fungi) which may be (a) infected by an “infectious agent” or (b) used to grow and/or amplify a nucleic acid or a nucleic acid containing organism or agent, (c) which may be used to express any nucleic acid sequence or (d) which may require treatment or vaccination. Organisms in need of treatment or vaccination may also be referred to as “subjects”. The term “host” includes, inter alia, cells used to amplify viruses, vectors, or plasmids, and cells used to express recombinant proteins.

[0048] The term “tissue”, as used herein, may refer to a particular tissue, cells of a particular type, a cells of a particular lineage, cells of a particular developmental state, cells of a particular stage of differentiation, differentiate cells and cells of a given state.

[0049] The term “mutant”, as used herein, refers to a modified nucleic acid or protein that has been altered (or “mutated”) by insertion, deletion and/or substitution of one or more nucleotides or amino acids. For example, the term mutant is used to refer to nucleic acid altered to disrupt a “sequence motif”, for example by substituting one or more nucleotides in the sequence motif with another nucleotide, or inserting one or more nucleotides to disrupt the sequence motif, or deleting one or more nucleotides in the sequence motif without substituting them for other nucleotides. The term “mutating” refers to the process of making such mutants.

[0050] The term “wild type” or “WT”, as used herein, refers to nucleic acids, and to organisms, cells, viruses, vectors, and the like, that have not been manipulated artificially to either disrupt a sequence motif or optimize their expression. The term “wild type” also refers to proteins encoded by such nucleic acids. Thus, the term “wild type” includes naturally occurring nucleic acids, viruses, vectors, cells and proteins. However, in addition, the term “wild type” includes non-naturally occurring nucleic acids, viruses, cells and proteins. For
example, unless otherwise stated, nucleic acids, viruses, vectors and cells that have been altered genetically are encompassed by the term “wild type” provided that those nucleic acids, viruses and cells have not been genetically altered with the intention of optimizing expression or disrupting a sequence motif therein.

[0051] The term “tissue specific”, as used herein, refers to the particular properties of a tissue or cell type that differentiate it from other types of tissues or cell types either within the same organism or between organisms. These particular properties comprise the proteomic profile of a given cell type (expressed or otherwise), the genetic profile of a given cell type (expressed or otherwise), the behavior of a given cell type in vitro and in vivo as well as other specific properties that enable those skilled in the art to differentiate a given cell type and some other cell type. The term “tissue specific”, as used herein, also refers to the expression profile of a gene, RNA or protein in a subset of tissues, cell types or cell states.

[0052] Other definitions are provided throughout the specification.

[0053] In one embodiment, the present invention is directed to a method for designing a nucleotide sequence optimized for expression in a target tissue by overlapping sequences derived from highly expressed native genes in the target tissue.

[0054] Because the genetic code is degenerate, nucleotide sequences can differ from each other at the nucleotide level but encode the same protein or peptide. There is selective pressure for the frequency and order of amino acids in the proteins encoded by the nucleotide sequences. However, in nature there is also often selective pressure for particular codon usage and AT/GC content that differs among organisms. Differing frequencies of codon usage among organisms can reflect the availability or abundance of tRNAs in protein translation. Thus, in order to provide for robust expression of a given gene in a foreign host, those skilled in the art appreciate that the introduction of mutations in a given gene to mimic codon usage in the target host can appreciably increase gene expression levels. The degeneracy of the genetic code allows for these mutations to be introduced in a silent manner such that the modifications do not alter the amino acid coding content of the gene.

[0055] In addition to differences in the abundance of tRNAs among tissues and cell types occurs within the same organism. The relative tRNA abundance among tissues can also be correlated to codon usage within cohorts of highly expressed genes in those tissues (Dittmars
et al, 2006). Such differences in codon usage may have functional consequences on the expression of a gene.

[0056] Cis acting elements that govern the production of RNAs and proteins encoded in the sequence of a gene may be located in coding sequences or non-coding sequences. Typically, if the location of the regulatory element is within a non-coding region, functional variability in the sequence of the element is not bound by the requirements of codon usage or a requirement to retain information for encoding the appropriate amino acid. When such regulatory elements are located within coding regions of genes, potential variability within the nucleotide sequence of the regulatory element may be restricted by a necessity to retain amino acid coding information. Nevertheless, given the redundancy in the genetic code, the nucleotide sequence of coding regions retain the potential to store functional cues for the regulation of gene expression without any impact on amino acid coding content. Such changes that change the nucleotide sequence without changing the amino acid coding sequence of a gene or RNA are “silent” changes.

[0057] Several examples of silent polymorphisms in coding nucleotide sequences have been described within splicing boundaries. Such examples include cis acting elements that regulate RNA splicing and RNA. Silent mutations can also impact the function of a gene encoded protein. A silent mutation (C3435T) in the Multi Drug Resistance 1 gene (MDR1) causes a change in the substrate specificity of the MDR1 protein in the absence of any differences between the encoded amino acid sequence of the wild type and “mutant” genes. The functional differences between the two alleles may be a result of differences in co-translational folding of the newly synthesized protein. Such silent changes in the nucleotide sequence of a gene have the potential to be to function as bone fide motifs and exercise an effect on a variety of biological processes. As is readily evident to one skilled in the art, such functional cues can comprise motifs recognized by cellular proteins involved in the regulation of gene expression, gene transcription, protein translation, RNA stability, RNA export, RNA localization, chromatin remodeling, DNA methylation as well as other modes of regulation. Such motifs can also comprise elements that dictate the formation of tertiary structures in DNA and secondary structure in RNA. Both have been implicated in regulating gene expression.

[0058] Since many genes display a tissue specific expression profile, it is possible that sequence motifs in the nucleotide sequence of genes highly expressed in a given tissue harbor
regulatory information. Further, this information may be encoded within the latitude provided by codon degeneracy and this information may potentiate gene expression at the level of transcription and/or translation.

**Method for identifying optimizing nucleotide sequences**

[0059] In the present invention, a recursive method is used to optimize the nucleotide sequence of a chosen gene for a selected organism, tissue or cell type.

[0060] In one embodiment, the present invention provides methods that enable optimization of the nucleotide sequence of a gene for expression in a specific organism or tissue by recursively overlapping the nucleotide sequence of a target gene with the nucleotide sequences of ranked genes in a tissue of interest. The method provides for the construction of an optimized version of the target gene which retains the original amino acid coding information of the target gene.

[0061] In one embodiment of the invention, the methods described herein are used to optimize expression of a lentiviral gene in a host. Lentiviruses belong to the Retrovirus family of viruses. The term “lenti” is Latin for "slow". Lentiviruses are characterized by having a long incubation period and the ability to infect neighboring cells directly without having to form extracellular particles. Their slow turnover, coupled with their ability to remain intracellular for long periods of time, make lentiviruses particularly adept at evading the immune response in infected subjects. Lentiviruses include immunodeficiency viruses, such as human immunodeficiency virus (HIV), simian immunodeficiency virus (SIV), feline immunodeficiency virus (FIV), and equine infectious anemia viruses (EIAV). Lentivirus infection can cause serious illness, and, if left untreated, can be fatal. In recent years several anti-retroviral drugs and drug cocktails that reduce viral load and ameliorate the symptoms of HIV infection have been developed. However, despite their successes, these drugs generally fail to eradicate the viral infection altogether. Instead the virus persists, often in a latent state, in infected subjects. There have also been multiple attempts to generate vaccines against lentiviral diseases such as HIV. However, to date, no vaccine is commercially available. Thus, there exists a need in the art to develop new drugs and vaccines against lentiviruses such as HIV.

[0062] The method comprises the following steps. Step 1: providing a target nucleotide sequence for optimization, wherein the nucleotide sequence encodes a protein having a
particular amino acid sequence. The optimization may involve optimization of expression in a particular organism or tissue of interest. For illustrative purposes, this nucleotide sequence may be termed “b” and the amino acid sequence that it encodes may be referred to as “s”.

Step 2: providing a group of one or more genes from the organism or tissue of interest for use as templates for the method of optimization. For illustrative purposes, this group may be termed “tn”. Step 3: ranking the group of one or more template genes from “tn” on the basis of one or more criteria, such as expression level in the organism or tissue of interest. For illustrative purposes, the group of one or more ranked genes may be termed “t”. In one embodiment, the group of one or more ranked genes are expressed in a tissue. Step 4: selecting a contiguous peptide sequence or “amino acid word” of defined length “L” from the amino acid sequence encoded by “s”. For illustrative purposes, this peptide sequence may be termed “w0”. The length “L” of “w0” may be referred to as its “word size” or “word length”. In one embodiment, the peptide sequence “w0” selected in step 4 is from a region between the C and N-termini of the amino acid sequence “s”. In one embodiment, the peptide sequence “w0” selected in step 4 includes the N-terminal amino acid of the amino acid sequence “s”. In one embodiment, the “word size” of the peptide “w0” is 2 amino acids in length. In another embodiment, the “word size” of the peptide “w0” is 3 amino acids in length. In another embodiment, the “word size” of the peptide “w0” is 4 amino acids in length. In yet another embodiment, the “word size” of the peptide sequence “w0” in step 4 is not a fixed value and is established by recursively scanning the group of ranked nucleotide sequences in group “t” for the longest possible “word size” by beginning with the longest possible “word size” in the amino acid sequence “s” and shortening the “word size” of “w0” by removing the C-terminal amino acid of “w0” recursively until a nucleotide sequence in the group of ranked nucleotide sequences “t” is identified. For example, for an amino acid sequence MPTGFYLPMTFDW (SEQ ID NO: 3) “s” of 15 amino acids in length, the initial word size is 15 amino acids. If a nucleotide sequence that can encode the word of 15 amino acids cannot be found in the group of nucleic acids in group “t”, then the word size is shortened to 14 and the group of nucleic acids “t” is scanned for a sequence that can encode the amino acid sequence MPTGFYLPMTFDW (SEQ ID NO: 4). The process is repeated recursively until a suitable nucleic acid sequence is identified in group “t”.

[0063] Step 5: sequentially examining the ranked genes from group “t” for the presence of a nucleotide sequence that encodes the polypeptide sequence “w0”. It is necessary to consider the translation frame of the contiguous nucleotide sequence in the ranked gene for this step as
the nucleotide sequences should actually encode the amino acid word “w0”. This nucleotide sequence in the ranked gene is termed “n0”. Step 6: removing the N-terminal amino acid from “w0” and adding to the amino acid immediately C-terminally adjacent to the corresponding amino acid sequence in the target protein to the C-terminal end of “w0”. This new amino acid sequence may be termed “w1”. Step 7: removing the first three 5’ nucleotides from the nucleotide sequence termed “n0”. This new nucleotide sequence may be termed “n01”. Step 8: replacing the nucleotide sequence of “n0” with the nucleotide sequence of “n1”, and Step 9: recursively repeating the steps 5 to 8 until the end of the amino acid sequence “s”, thereby generating the optimized nucleotide sequence “o”.

[0064] The steps above can be performed in the order described above. However, some of the steps may be performed in different orders or may be performed concurrently. For example, steps 1 and 2 may be performed in the inverse order, or may be performed simultaneously.

[0065] Step 1 of the above embodiment involves providing a nucleotide sequence “b” for optimization, wherein the nucleotide sequence encodes a protein having an amino acid sequence “s”. The optimization may involve optimization of expression in a particular organism or tissue of interest. There is no limitation on the origin, length or sequence of the nucleotide sequence selected in this step. The nucleic acid sequence may be DNA or RNA. The nucleotide sequence to be optimized may be any nucleotide sequence for which it is desired to optimize expression. For example, the nucleotide sequence to be optimized may encode a therapeutically useful protein for which it is desired to optimize expression in a particular organism or tissue of interest. Alternatively, the nucleotide sequence to be optimized may form part of a vaccine and may be optimized to improve production of the encoded protein in a vaccinated organism or tissue. In another embodiment, the nucleotide sequence encoding an amino acid sequence “s” need not be known. A hypothetical “seed” nucleotide sequence encoding the amino acid sequence “s” and of “word length” 3*L encoding the first “L” amino acids of sequence “s” can readily be generated by one skilled in the art. In some embodiments, protein “s” encoded by nucleotide sequence “b” may correspond to a mutated or modified variant of a given protein. In another embodiment, the protein “s” may correspond to a recombinant protein or a fusion protein. In other embodiments, the selected amino acid sequence designated as “s” may be an unknown protein. In other embodiments, the protein “s” may be a hypothetical protein that has not yet been produced, and/or may be a non-naturally occurring protein. The physical nucleotide
and/or amino acid molecules selected in this step need not be obtained or generated. Instead only virtual molecules or the amino acid or nucleotide sequences of these molecules need to be obtained generated, for example using a computer.

[0066] In one embodiment, the nucleic acid sequence to be optimized chosen in this step encodes an HIV or lentivirus protein, and may be useful in an HIV or lentivirus vaccine.

[0067] Step 2 of the above embodiment involves providing a group of one or more genes from the organism or tissue of interest for use a templates during optimization. There is no limitation of the origin, size, number or identities of the genes chosen for inclusion in the group “tn”. In one embodiment, the group of genes chosen for inclusion in this group are all of the genes of an organism of interest, or all of the genes expressed in a tissue of interest.

[0068] Step 3 of the above embodiment involves ranking genes in the group “tn”. This ranking is broadly defined and may make rank the sequences according to a variety of parameters such as level of expression in the organism of interest, or level of expression in the tissue of interest. In some embodiments, the process of ranking may comprise ordering the sequences according to a measure of DNA transcription, DNA replication, RNA reverse transcription, RNA replication, RNA stability, RNA localization, RNA export, protein stability, protein translation, steady state RNA levels, steady state protein levels, chromatin remodeling, DNA methylation, DNA replication, RNA replication, or any other parameter that influences the level of expression of a given gene. In another embodiment, the ranking may be subjective. In yet other embodiments, the ranking may reflect biological function or activity (such as for example, enzymatic activity, ability to stimulate an immune response, etc.), of may involve a relative measure of expression in one tissue or cell type relative to one or more other tissues or cell types of an organism, and the like. In one embodiment, the ranking is based on expressed mRNA levels. In another embodiment, the ranking is based on expressed protein levels.

[0069] The expression levels of these genes may by experimentally determined by standard investigative techniques such as microarray expression profiling, PCR, in situ hybridization, northern blotting, promoter assays, luciferase assays, immunofluorescence, western blotting, immunoprecipitation or any other suitable method known in the art. Similarly, gene expression information may be obtained from articles or publications containing the information or from publicly available information repositories such as the Gene Expression
Database (available at http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml) or the Gene Expression Omnibus (available at http://www.ncbi.nlm.nih.gov/geo/) or any of the public expression databases. Gene expression information from one organism may be applied to make inferences on the expression of homologous genes from other organisms. The sequences of the genes may also be determined by making reference to publicly available databases such as the GenBank database (available at the National Center for Biotechnology Information (NCBI) at http://www.ncbi.nlm.nih.gov/), the UCSC Genome browser (available at http://genome.ucsc.edu/cgi/hgGateway) or any of the public genome databases.

Alternatively, the sequence may be determined using any technique known in the art, including standard cloning and sequencing techniques. Suitable techniques for isolating, cloning, and determining the sequence of nucleic acids are well known in the art. See for example, Sambrook et al. (2001) Molecular Cloning: A Laboratory Manual, 3rd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y. The actual nucleotide molecules of the ranked genes need not be generated. Instead only virtual molecules need be generated, i.e. the ranking should be determined, for example using a computer, but the actual nucleic acid molecule having the sequence of the ranked molecules need not be produced.

[0070] Step 4 of the above embodiment involves selecting a peptide sequence or amino acid “word” from the amino acid sequence “s” encoded by the nucleotide sequence “b”. The peptide sequence (or amino acid “word”) selected in this step may termed “w0” and the length (or “word size”) of the peptide, as measured in amino acids, may be referred to as “L”. This peptide sequence or amino acid “word” may be chosen from any region of the amino acid sequence “s” and may include amino acids at either the amino or carboxy terminal regions of the protein of interest. In one embodiment, the peptide sequence selected in this step includes the N-terminal amino acid of protein “s”. In another embodiment, the peptide sequence selected in step 4 includes an N-terminal amino acid residue of a functional domain of protein “s”.

[0071] The length “L” of an amino acid “word” must contain at least two amino acids, but the upper limit on word length “L” is variable. One of skill in the art can select a suitable word length. In one embodiment, the large word size is selected to scan the group of nucleic acids in the group “t”. If a nucleic acid sequence in the group “t” that encodes the amino acid sequence in “w0” cannot be found, then the length of “w0” is decreased by one by removing the C-terminal amino acid and the group of nucleic acids is scanned again. The process is
repeated recursively by shortening the length of the word “w0” until a nucleic acid sequence in the group “t” is identified. In the Examples provided herein, word lengths of 3 or 4 amino acids was chosen for the optimization of the HIV or lentivirus proteins Gag, Pol and Nef in human muscle tissues. Shorter or longer word lengths could have been chosen if desired, taking into account the above considerations.

[0072] Once a suitable word length has been chosen, routine methods can be used to identify and count each word. For example the nucleotide sequence MPRLTY (SEQ ID NO: 5) contains the 2 “letter” words MP (SEQ ID NO: 6), PR (SEQ ID NO: 7), RL (SEQ ID NO: 8), LT (SEQ ID NO: 9), and TY (SEQ ID NO: 10), the 3 letter words MPR (SEQ ID NO: 11), PRL (SEQ ID NO: 12), RL (SEQ ID NO: 13), and LTY (SEQ ID NO: 14), and the four letter words MPRL (SEQ ID NO: 15), PRLT (SEQ ID NO: 16), RLTY (SEQ ID NO: 17). This type of word identification and word counting can be performed using standard methods known in the art in order to identify and count words of a given length in a given real genome. In one embodiment, the “word size” of this peptide is 3. In another embodiment, the “word size” of this peptide is 4. In yet another embodiment, the “word size” of this peptide is 5. In a further embodiment, the “word size” of this peptide is not a fixed value and is instead defined by the largest possible length of the amino acid sequence in “s”. In this embodiment, “w0” is used to scan the group of genes in “t” to identify the occurrence of a nucleotide sequence capable of coding for the amino acid sequence in “w0”. If a corresponding nucleotide sequence is not found within the nucleic acid sequences of group “t”, then the length of “w0” is shortened by one by removing the C-terminal amino acid for “w0” and the process is repeated recursively until a “hit” is identified. If at any stage a matching hit cannot be identified in any of the embodiments described herein, a nucleotide sequence can be assigned for any “w0” of a “word length” equal to one. This assignment may be arbitrary or random. Directed assignment methods are also readily apparent to those skilled in the art. In one embodiment, this assignment for the amino acid “w0” of word length 1 may be chosen to match the nucleotide sequence encoding that particular amino acid in the sequence “s”. In another embodiment, this assignment for the amino acid “w0” of word length 1 may be chosen on the basis of preferential codon usage of genes that are highly expressed in a cell type, tissue type or organism.

[0073] Step 5 of the above embodiment involves identifying a nucleotide sequence matching hit coding for the amino acid “word” selected in step 4 from the cohort of ranked genes “t”
from step 3. Several approaches to identify a nucleotide sequence coding for the amino acid “word” selected in step 4 from the cohort of ranked genes “i” from step 3 are readily apparent to one skilled in the art. In one embodiment, the first matching hit in the highest ranked gene of group “i” is selected. In another embodiment, the “hit” is selected by counting the all of the occurrences of a suitable nucleotide sequence in the database of ranked genes and then weighing the occurrences by the rank of the corresponding genes, wherein the higher ranked genes are given a higher weight. In one embodiment, it is required that the selected nucleotide sequence adhere to the translation frame of the amino acid sequence selected in step 2. In another embodiment, it is required that the selected nucleotide sequence reside wholly within an protein coding region of a gene. The nucleic acid sequence selected in this step for its potential to code for the amino acid “w0” is termed “n0”.

[0074] In step 6 of the above embodiment, the amino acid word selected in step 5 is shortened by eliminating the amino acid residue at the amino-terminal end of the amino acid sequence. This truncated version of “w0” is then lengthened to the original word size (i.e. word length “L”) by the addition of the corresponding amino acid in the starting protein “s” to generate the peptide “w1”.

[0075] In step 7 of the above embodiment, three nucleotides from the 5’ end of the nucleotide sequence of “n0” are removed to generate the nucleotide sequence “n01”.

[0076] Step 8 of the above embodiment involves replacing the nucleotide sequence of “n0” with the nucleotide sequence of “n1”, and step 9 of the above embodiment involves recursively repeating the steps 5 to 8 until the end of the amino acid sequence “s”, thereby generating the optimized nucleotide sequence “o”. At the end of the process, the starting nucleotide sequence “b” and the optimized nucleotide sequence “o” will encode the same amino acid sequence, and they will retain the same translation frame as one another, even though their nucleotides sequences will differ.

[0077] In one embodiment, the present invention is directed to methods for optimizing the production of proteins in hosts. Such methods can be used, inter alia, to optimize the production of therapeutically useful proteins, or to optimize vaccines that contain protein-coding nucleic acid sequences so as to improve the production of the proteins in a vaccinated host.
[0078] For example, in one embodiment the present invention provides a method for optimizing the production of a protein in a host by optimizing a nucleotide sequence that encodes the protein, wherein the mutations result in improved production of the protein in the host.

[0079] The methods of the invention can be used to optimize the expression of any protein. In some embodiments, the protein whose expression is optimized is a therapeutic protein. In other embodiments, the protein whose expression is optimized is an immunogenic protein, such as an immunogenic protein that can be administered to a subject as a component of a proteinaceous vaccine. In yet other embodiments, the immunogenic protein is one that is expressed in a subject from a nucleic acid present in a vaccine composition. Examples of vaccine compositions that contain nucleic acids include, but are not limited to, attenuated viral vaccines and various vector-based vaccines.

[0080] The methods of the invention can be used to optimize the production of proteins in various hosts, including but not limited to, eukaryotes, prokaryotes, bacteria and yeasts. For example, the host may be any wild-type, mutant, or transgenic animal or plant, or any cell or cell-line derived therefrom. In certain embodiments, the host is a mammal, such as a human, or a cell or cell line derived from a mammal. In other embodiments, the host may be an insect cell or an insect cell line. In yet other embodiments the host is a cellular system or culture that can be used to produce large quantities or proteins for therapeutic uses. In other embodiments, the host may be a subject in need of vaccination.

[0081] In one embodiment, the present invention is directed to a viral nucleic acid that has been mutated to change one or more nucleic acids to optimize expression in a organism, tissue type or cell type manner. In some embodiments, the viral nucleic is from an HIV or lentivirus virus. In other embodiments, the viral nucleic acid that has been mutated to change one or more nucleic acids in the either the gag, pol, or env genes.

[0082] In another embodiment, the present invention is directed to a method for producing optimized expression of a viral nucleic acid having one or more nucleotides modified, in, or derived from, any location in the viral genome, including coding and non-coding regions. In one embodiment, if the optimized nucleotide sequence is in a region of the viral nucleic acid that encodes a protein, and the nucleotide sequence is changed such that it does not adversely
affect the structure, function or immunogenicity of the protein encoded by the viral nucleic acid. In further embodiments, the viral nucleic acid is an HIV or lentivirus nucleic acid.

[0083] In another embodiment, the present invention is directed to a mutant virus having a genome that been mutated to optimize expression of one or more viral nucleotide sequences. In some embodiments, the mutant virus is a mutant HIV or lentivirus virus.

[0084] In yet another embodiment, the present invention is directed to a recombinant virus that is not a virus but that contains a viral nucleic acid sequence that has been mutated to optimize expression of a viral nucleic acid. In one embodiment, the mutant viral nucleic acid is a mutant HIV or lentivirus nucleic acid.

[0085] In a further embodiment, the present invention is directed to a viral protein expressed from a viral nucleic acid sequence that has been modified to optimize expression. In one embodiment, the invention is directed to an HIV or lentivirus protein expressed from a HIV or lentivirus nucleic acid sequence that has been modified to change one or more nucleotides in the viral nucleotide sequence.

[0086] In another embodiment, the present invention is directed to a virus vaccine comprising a viral nucleic acid sequence that has been mutated to optimize expression. In one embodiment, the invention is directed to an HIV or lentivirus vaccine comprising an HIV or lentivirus nucleic acid sequence that has been mutated to optimize expression.

[0087] In another embodiment, the present invention is directed to a virus vaccine capable of higher protein expression than the corresponding wild-type virus, wherein the virus vaccine comprises a nucleic acid that has been optimized for expression. In one embodiment, the present invention is directed to an HIV or lentivirus vaccine capable of higher protein expression than the corresponding wild-type HIV or lentivirus virus, wherein the HIV or lentivirus vaccine comprises a nucleic acid sequence that has been optimized to provide greater expression than the wild-type HIV or lentivirus virus nucleic acid sequence.

[0088] In another embodiment, the present invention is directed to a viral vaccine comprising a protein produced from a viral nucleic acid that has been mutated to optimize expression. In one embodiment, the present invention is directed to an HIV or lentivirus vaccine comprising a protein produced from an HIV or lentivirus nucleic acid that has been mutated optimize expression.
[0089] In another embodiment, the invention is directed to a composition comprising a vaccine as provided by the present invention, and an additional component selected from the group consisting of pharmaceutically acceptable diluents, carriers, excipients and adjuvants.

[0090] In yet another embodiment, the invention is directed to a method for immunizing a subject against a virus comprising administering to the subject an effective amount of a vaccine of present invention. In one embodiment, the invention is directed to a method for immunizing a subject against a virus, comprising administering to the subject an effective amount of a virus that has been mutated to optimize expression. In another embodiment, the invention is directed to a method for immunizing a subject against HIV or lentivirus, comprising administering to the subject an effective amount of a nucleic acid encoding a virus protein that has been mutated to optimize expression. In yet another embodiment, the invention is directed to a method for immunizing a subject against a virus, comprising administering to the subject an effective amount of a viral protein produced from a viral nucleic acid that has been mutated to optimize expression. In some embodiments, the invention is directed to methods for immunizing a subject against HIV or lentivirus.

[0091] In another embodiment, the invention is directed to methods for identifying agents that inhibit or stimulate production of viral RNA, production of virus protein or production of virus particles, or that inhibit or stimulate viral latency. In one embodiment, the method comprises providing a control cell containing at least one viral nucleic acid sequence containing at least one optimized nucleic acid mutation and a test cell containing at least one viral nucleic acid of the corresponding wild type sequence, contacting the test cell and the control cell with one or more agents, and identifying at least one agent that inhibits or stimulates production of viral RNA, production of virus protein or production of virus particles, or that inhibits or stimulates viral latency, in the test cell as compared to the control cell. In one embodiment, the agents inhibit or stimulate production of HIV or lentivirus RNA, production of HIV or lentivirus protein or production of HIV or lentivirus particles, or inhibit or stimulate HIV or lentivirus latency.

[0092] These and other embodiments of the invention are described further in the accompanying specification, drawings, and claims

**Tissue Specific Expression of Gene**
[0093] Tissue specific gene expression functions to define the expression range of a given gene to a subset of tissues or cell types within an organism. The expression range need not be defined in an absolute manner and rather may be measured in relative terms over the expression profile of a gene. To be expressed in a tissue specific manner, the DNA segment encoding a gene is typically coupled to one or more cis acting regulatory elements that regulate the expression profile of the gene. Such regulatory elements comprise, but are not limited to, elements that promote transcription, enhance transcription, silence transcription, modulate transcription such that it is responsive to extracellular and intracellular cues, regulate stability of the encoded RNA, regulate splicing of the encoded RNA, regulate export of the encoded RNA, regulate localization of the encoded RNA, regulate translation from the encoded RNA. Within the scope of this invention, genes expressed in a tissue specific manner suitable for ranking in the optimization method described above will be apparent to those skilled in the art. Also apparent to those skilled in the art is that the expression profile of a given gene in one organism is frequently a reliable indicator of the expression pattern of homologs in phylogenetically related organisms. One such example is that the gene encoding creatine kinase is highly expressed in skeletal muscle tissues in both mice and humans. Therefore, in the absence of data of creatine kinase expression in human tissues, it would be reasonable to hypothesize that the general expression pattern of this gene in humans will resemble the expression pattern of creatine kinase in mice. In the absence of data for a given species, the method described herein provides a method for inferring expression profiles from homologs in different species. Examples of genes expressed in a tissue or cell type specific manner are provided below. These examples are non-limiting and other types of regulated gene expression are apparent to those skilled in the art.

Adipose Tissue

[0094] Non-limiting examples of genes expressed in human adipose tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, geRMA dataset comprises, but is not limited to, adiponectin, DKK1 (dickkopf homolog 1), FADS1 (fatty acid desaturase 1), LPL (lipoprotein lipase), PLIN (perilipin), PTX3 (pentraxin-related gene, rapidly induced by IL-1 beta) and SRPX (sushi-repeat-containing protein, X-linked).

Bone Marrow
[0095] Non-limiting examples of genes expressed in human bone marrow tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, geRMA dataset comprises, but is not limited to, CA1 (carbonic anhydrase I), ELA2 (elastase 2), HBG1 (hemoglobin, gamma A), MPO (myeloperoxidase), PRG3 (proteoglycan 3) and S100P (S100 calcium binding protein P).

Heart

[0096] Non-limiting examples of genes expressed in human heart tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, geRMA dataset comprises, but is not limited to, FHL2 (four and a half LIM domains 2), HRC (histidine rich calcium binding protein), MB (myoglobin), MYOZ2 (myozin 2) and TPM1 (tropomyosin 1 alpha).

Kidney

[0097] Non-limiting examples of genes expressed in human kidney tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, geRMA dataset comprises, but is not limited to, ABP1 (amiloride binding protein 1), BHMT (betaine-homocysteine methyltransferase), FXYD2 (FXYD domain containing ion transport regulator 2) SLC3A1 (solute carrier family 3), SPP1 (secreted phosphoprotein 1) and UMOD (uromodulin).

Liver

[0098] Non-limiting examples of genes expressed in human liver tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, geRMA dataset comprises, but is not limited to, 39763_at (Human hemopexin gene), CBP2 (Carboxypeptidase B2), FABP1 (fatty acid binding protein 1), HB (haptoglobin), HPX (hemopexin) and SDS (serine dehydratase).

Lung
[0099] Non-limiting examples of genes expressed in human liver tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, gcRMA dataset comprises, but is not limited to, AGER (advanced glycosylation end product-specific receptor), CLIC3 (chloride intracellular channel 3), EMP2 (epithelial membrane protein 2), SCNN1A (sodium channel, nonvoltage-gated 1 alpha), SFTPA2 (surfactant, pulmonary-associated protein A2) and TPSB2 (tryptase beta 2).

**Lymph Node**

[00100] Non-limiting examples of genes expressed in human lymph node tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, gcRMA dataset comprises, but is not limited to, CCL21 (chemokine (C-C motif) ligand 21), CD52 (CD52 antigen (CAMPATH-1 antigen)), GPNMB (glycoprotein (transmembrane) nmb), MS4A1 (membrane-spanning 4-domains, subfamily A, member 1), PTPRC (protein tyrosine phosphatase, receptor type, C) and TRBC1 (T cell receptor beta constant 1).

**Pancreas**

[00101] Non-limiting examples of genes expressed in human pancreatic tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, gcRMA dataset comprises, but is not limited to, CEL (carboxyl ester lipase), CLPS (colipase, pancreatic), CTRB1 (chymotrypsinogen B1), CUZD1 (CUB and zona pellucida-like domains 1), INS (Insulin) and TRY6 (trypsinogen C).

**PB-CD14+Monocytes**

[00102] Non-limiting examples of genes expressed in human PB-CD14 positive monocytes at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, gcRMA dataset comprises, but is not limited to, AIF1 (allograft inflammatory factor 1), BZRP (benzodiazapine receptor), CD14 (CD14 antigen), CIAS1
(cold autoinflammatory syndrome 1), HRB2 (HIV-1 rev binding protein 2) and KYNU (kynureninase).

**PB-CD19+Bcells**

[00103] Non-limiting examples of genes expressed in human PB-CD19 positive B Cells at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, gcRMA dataset comprises, but is not limited to, BANK1 (B-cell scaffold protein with ankyrin repeats 1), BLNK (B-cell linker), CD19 (CD19 antigen), CXCR4 (chemokine (C-X-C motif) receptor 4), FREB (Fc receptor homolog expressed in B cells) and PAX5 (paired box gene 5).

**PB-CD56+NKCells**

[00104] Non-limiting examples of genes expressed in human PB-CD56 positive Natural Killer Cells at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, gcRMA dataset comprises, but is not limited to, CD160 (CD160 antigen), GNLY (granulysin), GZMH (granzyme H), KIR3DL2 (killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 2), KLRB1 (killer cell lectin-like receptor subfamily B, member 1) and NKG7 (natural killer cell group 7 sequence).

**Skeletal Muscle**

[00105] Non-limiting examples of genes expressed in human skeletal muscle tissues at levels 100 fold over their median expression in other tissues are available at the Gene Expression Atlas (available at: http://symatlas.gnf.org/SymAtlas/). This group of genes from the GNF1H, gcRMA dataset comprises, but is not limited to, ACTA1 (actin, alpha 1, skeletal muscle), CKM (creatinine kinase), CMYA5 (cardiomyopathy associated 5), LDB3 (LIM domain binding 3), MYOZ1 (myozinin 1) and TNNI1 (troponin I).

**Other Types of Tissue Specific Expression**

[00106] The invention described herein also provides for other modes of tissue specific gene expression. The examples provided above are not intended to be limiting in any way and other cell or tissue specific modes of expression are envisioned. Furthermore, there is no
limitation that the grouping or ranking of the genes selected for the optimization process be limited to one single criterion. In one embodiment, the genes are grouped and ranked on the basis of their expression pattern in two cell or tissue types. In another embodiment, the genes are grouped and ranked on the basis of their expression pattern in 3 to 10 cell or tissue types. In yet another embodiment, the genes are grouped and ranked on the basis of their responsiveness to biological cue, such as inflammatory responses. In yet another embodiment, the genes are ranked on the basis of gene expression in a cell type specific manner, such as immune responses in macrophages. Other relevant examples are evident to those skilled in the art.

**Computer Systems**

**[00107]** The methods of the present invention can be performed using a computer. In one embodiment, the invention involves the use of a computer system which is adapted to allow input of the sequences of proteins and ranked genes and which includes computer code for performing one or more of the steps of the various methods described herein. For example, the present invention encompasses a computer program that includes code for performing one or more of generating protein sequences, generating gene sequences, ranking gene sequences, computing each of the steps in the above method sequentially or simultaneously, and the like.

**[00108]** The computer systems of the invention can comprise a means for inputting data such as the sequence of proteins and ranked genes, a processor for performing the various calculations described herein, and a means for outputting or displaying the result of the calculations. Typically, that result will be an optimized sequence of a given protein against a given set of ranked genes.

**[00109]** One of skill in the art can readily create computer code for executing the methods of the invention, using any suitable computer code language or system known in the art, such as “C” for example.

**Optimization of Sequences**

**[00110]** The methods of the present invention have many different uses and applications, some of which are described below. In some embodiments, nucleic acid sequences can be optimized by the methods of the invention for expression in a particular
foreign host organism. Optimization of such sequences can be achieved by ranking template
genes from a target host according to a given criteria. In one embodiment, the ranking is based
on gene expression and in other embodiments, the ranking may be based on the levels of
RNA or protein expression of a gene. Specific examples provided below relate to viral
nucleic acids, but such examples are not intended to be limiting and the methods of the
invention can be used to optimize any nucleic acid including nucleic acids sequences form
eukaryotic, prokaryotic of fungal sources. The nucleic acid by be optimized by the methods
of the present invention need not exist and may be a hypothetical or artificial sequence with
no known homologous sequences. In other embodiments, nucleic acid sequences to be
optimized and the nucleic acid sequences ranked as temples in the methods of the present
invention may be from the same host organism or be expressed in the same tissues or cell
type. Some applications of the methods of the present invention may be for increasing the
expression of a particular gene in a host or cell type. In one non-limiting example, the
methods of the invention may be useful for the production of viral proteins, such as HIV or
lentivirus proteins, that are otherwise expressed at low levels in human cells. The methods of
the present invention also enable optimization of nucleic acid sequences to increase
expression in a foreign host or a specific cell type. In other embodiments, the methods of the
present invention can be used to optimize specific regions or putative functional elements of a
nucleic acid sequence. For example, there are many elements that are known to regulate the
gene expression. Some examples non-limiting include RNA slice sites, translation initiation
sites or elements that regulate the stability of an RNA molecule. The efficacy and functional
consequences of such elements can depend on the environment in which they function. For
example, a splice site that functions at high efficiency in one cell type or organism may
function at a lower level in a different cell type or organism. Modification of a nucleotide
sequence by the methods of the present invention can by used to optimize a nucleic acid
sequence according to topological biases in nucleotide sequences. For example, the presence
of a particular nucleic acid sequence proximal to a splice site in a cohort of genes that are
highly expressed in a given cell type may be used to optimized a foreign nucleic acid
sequence by the methods of the invention by biasing the ranking of template genes according
the proximity of nucleic acids to a splice site. Splice sites are a non limiting example and any
other such functional elements may be contemplated in the ranking and optimization of
nucleic acid sequences. Other applications for the methods of the present invention will be
readily apparent to those of skill in the art.
HIV Biology

[00111] The HIV genome encodes several proteins, some of which are produced as poly-proteins that produce different functional entities upon proteolytic cleavage. All of the proteins encoded by the HIV genome, including but not limited to poly-proteins and their proteolytic cleavage products, are within the scope of the invention, and may be referred to herein as “HIV proteins”, “HIV peptides”, HIV poly-proteins”, “proteins of the invention”, “polyproteins of the invention” or “peptides of the invention”.

[00112] For example, the HIV genome encodes the pr55 GAG protein, which can be cleaved by a viral protease into p17MA, p24CA, p7 and p6 proteins that make up the core of the virus. The Pr160 GAG-POL precursor protein produces a polymerase poly-protein that is made by translational frame shifting and is subsequently cleaved into a reverse transcriptase (RT), RNAase H, a protease (PR) and an integrase (IN). The Gp160 envelope protein is cleaved by a cellular protease into a Gp120 protein (ENV) that attaches the virus to the CD-4 receptor and the co-receptor and a Gp41 trans-membrane protein that fuses the viral envelope into the host cell membrane. These are the major structural proteins and enzymes of the HIV particle. Two additional proteins, Tat and Rev, regulate the transcription of the integrated provirus (Tat) and the transport of unspliced mRNAs (GAG-POL and viral genomic RNA) from the nucleus to the cytoplasm (Rev). These activities regulate the expression of viral genes and the temporal events during the replication cycle. Four accessory proteins: Vpu, Vif, Vpr and Nef are adaptors that form complexes with cellular proteins and enhance infectious virus production in vivo, but have more minimal phenotypes in some cell cultures. All of these HIV proteins are within the scope of the invention.

[00113] The HIV virus attaches to T-cells and monocytes using its gp120 ENV protein, which binds to the CD-4 protein and co-receptors at the cell surface. T-cells express the cytokine receptor CXCR4, monocytes express the CCR5 co-receptor, and peripheral blood lymphocytes express both. Genetic alterations in the ENV protein produce monocyte tropic viruses (R5 viruses), T-cell tropic viruses (X4 viruses) and dual tropic viruses. The co-receptor is essential for ENV attachment and Gp41 fusion. People who lack a normal HIV co-receptor, for example because they carry the CCR5/delta 32 polymorphism, are almost completely resistant to HIV infection. (See, for example, Deng et al., Nature (1996) Vol. 381, p 661-6 and Samson et al., Nature, (1996) Vol. 382 p722-5). Drugs that block the CCR5 co-receptor are now in clinical trials.
After fusion of the HIV virus with the cell, the virus core particle copies the viral RNA into a DNA copy using reverse transcriptase (RT) and the particle moves to the nucleus where it integrates the viral DNA into the cellular genome using integrase (IN). The long terminal repeat (LTR) DNA sequences contain a number of transcription factor binding sites that are essential to produce viral mRNAs from the incorporated DNA. In addition to a TATA element, there are two NF kappa b sites and three Sp-1 sites in the LTR which are recognized by cellular transcription factors. In addition sites for LEF, ETS and USF transcription factors are also present. These cellular transcription factors help to initiate transcription, but this occurs at a very low level. After TAT is produced, it binds to a cellular protein (cyclin T1) and the TAT-cyclT1 complex then binds to an RNA loop structure (called TAR) in the viral mRNA. The TAT-cyclT1 complex next binds the CDK9 protein kinase, which phosphorylates the carboxy-terminal end of one of the subunits of RNA polymerase-II. This is required for the efficient initiation of viral RNA transcription (which increases by 100 fold). Over 30 different viral RNA molecules are produced by these events. They fall into three categories: (1) unspliced RNAs that are used to make GAG, POL and the intact viral genomes, (2) partially-spliced RNAs of about 5.0Kb in size, that are employed to make ENV, Vif, Vpu, and Vpr proteins, and (3) small, spliced RNAs (1.7-2.0Kb) that are translated into REV, TAT and Nef. The transport of these RNAs out of the nucleus is most efficient for the fully spliced mRNAs. Thus, early after infection, only TAT, REV and NEF are made efficiently. TAT binding to TAR then increases the rate of transcription by 100 fold. The larger, unspliced or poorly spliced mRNAs are transported into the cytoplasm more efficiently only after the REV protein is made and binds to the Rev-responsive element (RRE) in the ENV gene. In this way, the synthesis of TAT and REV regulate timing of the viral life cycle.

In addition to TAT, there is a second set of signals in the HIV genome that reduce the steady state levels of viral RNA in cells. These are referred to as inhibitory nucleotide signal sequences (INS sequences). Putative INS-containing regions have been identified previously in the gag/pol regions of the HIV genome (see Schneider et al., Journal of Virology, 1997, Vol. 71, p. 4892-4903). In the prior study by Schneider et al. the region containing putative INS sequences was mutated to eliminate AUUA (SEQ ID NO: 18) pentanucleotides and to decrease AU content without altering the coding capacity of the region. It was found that these mutations resulted in an increase in the level of HIV RNA by up to 70-130 fold. With the INS sequences mutated and in the presence of a functional REV,
the increase in HIV RNA in an infected cell was 160 fold higher than without these two distinct functions.

[00116] It is believed that the HIV virus gains an advantage by having a low steady state level of viral RNA. It has been proposed that a virus that replicates rapidly and kills the cell rapidly produces less virus per-cell than one that employs a slower cycle. Indeed, rapid efficient virus production and cell killing, as is observed with polio viruses for example, often leads to complete immune clearance of the virus and immunity to subsequent infections. In contrast, viruses which remain intracellular for prolonged periods are able to evade, or at least reduce the effectiveness of, the immune response against them.

[00117] Although Schneider et al. did determine that mutations in the putative INS-containing region of the gag/pol genes were additive, they did not identify the sequence(s) of the INS. Schneider et al. demonstrated that the INS-containing region(s) acted to lower the steady state levels of viral RNA in an infected cell, and discussed four possible mechanisms for this effect. These included (1) a lowered rate of transcription, (2) an increased rate of RNA degradation of HIV RNA, (3) a lowered rate of transport of HIV RNA out of the nucleus, and (4) an altered efficiency of translation of the RNAs. Schneider et al. eliminated possibility (4), but could not distinguish among the other three possibilities.

[00118] The assembly of HIV virus particles takes place at the cell membrane where the GAG-POL poly-protein packages the viral RNA. The viral particles bud off from the plasma membrane of the host cell, which now contains the HIV ENV and gp41 proteins. After release of the virus particle the viral protease (PR) cleaves the GAG-POL poly-protein giving rise to mature and infectious particles. Vif is believed to be involved in assembly of the virus. Also, NEF and Vpu are involved in the degradation of cellular CD-4 protein and release of viruses from the membrane of infected cells. The Vpr protein is incorporated into the virion at assembly and appears to play a role early in infection, transporting the particle into the nucleus, while Vif antagonizes an anti-viral cellular enzyme activity in the cell. The released infectious virus attaches to and infects additional cells and the progressive infection and killing of CD-4 lymphocytes results in an incapacitated immune system.

Optimization of sequences for protein production

[00119] Recombinant proteins have many applications, for example as therapeutic agents and as components of proteinaceous vaccines. These recombinant proteins are
generally produced in host cells that have been transformed or transfected with expression vectors containing a nucleotide sequence that encodes the protein, under the control have a suitable promoter. Often recombinant proteins are expressed and produced in cell types of a species different than that from which the nucleotide sequence is derived. For example Amgen’s recombinant human erythropoietin product is produced in cultured hamster ovary (CHO) cells, and recombinant human G-CSF, the active ingredient in the commercial product Neupogen®, is produced in E. coli bacterial cells. In such situations, the nucleotide sequence encoding the recombinant protein may not contain certain sequence motifs that are present in the genome of the host cells, or may contain additional sequence motifs that are absent in the host cell. These differences may adversely affect the expression of foreign recombinant proteins in host cells. For example, the host genome may contain certain sequence motifs required for mRNA stabilization in the host that are absent in the recombinant nucleotide sequence, or the recombinant nucleotide sequence may contain certain sequence motifs that inhibit or decrease the efficiency of protein expression in the host. Thus, it may be useful to mutate the nucleotide sequence encoding the recombinant protein to add one or more of the host-specific optimized nucleic acid sequences or to remove one or more of the source species sequence, so as to optimize production of the recombinant protein in the host cells. For example, if a recombinant human protein is to be expressed in hamster cells, it may be desirable to add one or more hamster-specific optimized nucleic acid sequences to the nucleotide sequence that encodes the recombinant human protein. Similarly, if a recombinant human protein is to be expressed in insect cells, such as using the baculovirus expression system, it may be desirable to add one or more insect-specific sequence motifs to the nucleotide sequence that encodes the recombinant human protein.

[00120] There are many variations on the above concepts, all of which are within the scope of the present invention. For example, any nucleotide sequence encoding a recombinant protein may be optimized using the methods described herein, including, but not limited to, sequences encoding any eukaryotic, prokaryotic, plant, animal, bacterial, yeast, insect, mammalian, primate, human, hamster, mouse, goat, sheep, bird or chicken recombinant protein.

[00121] Similarly, the host system in which the recombinant nucleotide protein is to be produced may be any suitable cellular expression system known in the art, including, but not limited to, eukaryotic expression systems, prokaryotic expression systems, plant expression
systems, animal expression systems, bacterial expression systems, yeast cell expression systems, insect cell expression systems, mammalian cell expression systems, primate cell expression systems, human cell expression systems, hamster cell expression systems, mouse cell expression systems, goat cell expression systems, sheep cell expression systems, bird cell expression systems, chicken cell expression systems, and the like. The host expression system may also be any cell line suitable for recombinant protein expression, including, but not limited to, Chinese hamster ovary (CHO) cells, mouse myeloma NS0 cells, baby hamster kidney cells (BHK), human embryo kidney 293 cells (HEK-293), human C6 cells, Madin-Darby canine kidney cells (MDCK) and Sf9 insect cells. The expression system may also be an entire organism, such as a transgenic plant or animal. For example, the expression system may be a transgenic sheep or cow that capable of expression of recombinant proteins that are secreted into the milk, or a recombinant plant capable of expressing recombinant proteins. Any suitable host system for recombinant protein expression known in the art can be used in accordance with the methods of the present invention.

[00122] As stated above, the nucleotide sequence encoding the recombinant protein can be altered in multiple ways to make it more compatible with the host’s cellular environment. In one embodiment, the methods of the present invention are used to identify sequence motifs present in the nucleotide sequence encoding the recombinant protein that are either over- or under-represented in the host genome. In a next step, the functional consequences of the sequence motifs can be determined. This can be done by mutating the sequence motifs in either the nucleotide sequence encoding the recombinant protein or in the host genome, and testing the effect of these mutations on certain biological properties such as rate of mRNA production, stability of mRNA, rate or protein production, protein stability, cleavage by restriction enzymes, and the like. In a further step, the nucleotide sequence encoding the recombinant protein can be optimized by making mutations to remove or disrupt one or more disadvantageous sequence motifs or to add or create one or more advantageous sequence motifs. Any suitable mutation methods known in the art, such as those described herein, may be used.

Optimization of Vector Sequences

[00123] In another embodiment, the methods of the invention can be used to optimize the sequence of various vectors, such as vectors used for expression of recombinant proteins, vectors used for gene therapy, vectors used as vaccines, and the like. Such vectors may be,
for example, plasmid vectors or viral vectors (i.e. vectors that comprise, or are derived from a viral genome). Methods for modifying nucleotide sequences that encode recombinant proteins and which may be inserted into vector backbones, are described below. However, the methods of the present invention can also be used to optimize the vector backbone itself. For example, many vectors themselves encode various proteins. For example, viral vectors may encode various viral proteins. In some situations it may be desirable to optimize a vector by eliminating or minimizing expression of proteins encoded by a vector backbone. In other situations it may be desirable to optimize a vector to increase expression of proteins encoded by a vector backbone. Vector sequences can be altered in the same ways as described above for protein-coding sequences in order to achieve these results. For example, the methods of the present invention may be used to identify sequence motifs present in the vector backbone that are either over- or under-represented as compared to the host genome. The functional consequences of these sequence motifs can be determined. This can be done by mutating the sequence motifs in either the vector or in the host genome, and testing the effect of these mutations on certain biological properties such as rate of production of vector-encoded mRNAs, stability of vector-encoded mRNAs, rate or production of vector-encoded proteins, stability of vector-encoded proteins, and the like. Then, the nucleotide sequence of the vector backbone may be optimized by performing mutations to remove one or more disadvantageous sequences in the vector backbone, or to add one or more advantageous sequences to the vector backbone. Any suitable mutation methods known in the art, such as those described herein, may be used.

Optimization of Vaccines

[00124] The methods described above for optimization of sequences for protein production and optimization of vector sequences can be used to optimize vaccines, including, but not limited to, attenuated viral vaccines, killed viral vaccines, viral vector vaccines, DNA vaccines, and protein vaccines.

[00125] Attenuated viruses are viruses that have been altered to weaken them, such that they no longer cause disease but may still stimulate an immune response. There are many ways in which a virus may be attenuated. For example, a virus can be attenuated by removal or disruption of viral sequences required for causing disease, while leaving intact those sequences encoding antigens recognized by the immune system. Attenuated viruses may or may not be capable of replication in host cells. Attenuated viruses that are capable of
replication are useful because the virus is amplified in vivo after administration to the subject, thus increasing the amount of immunogen available to stimulate an immune response. The methods of the invention can be used to identify sequence motifs that are either under- or over-represented in a viral strain as compared to its host, and mutate these sequence motifs to increase the level of attenuation of a virus and/or to increase its immunogenicity in a host. For example, mutations can be made to disrupt or remove sequence motifs that are involved in the virulence of the viral strain or to add sequence motifs that suppress the virulence of the viral strain in its hosts. If the attenuation methods used involve disrupting or deleting sequence motifs within the virus genome, these mutations can be made sufficiently large in size or number such that the chance reversion of the virus to a non-attenuated form is close to zero.

[00126] Killed or inactivated viral vaccines are generally non-functional and do not express viral genes or replicate in a vaccinated subject. However, the methods of the invention may be used to facilitate expansion and growth of a viral strain in vitro or ex vivo prior to inactivation of the virus. For example, by mutating one or more inhibitory sequence motifs in a virus, the rate of viral expansion in host cells may be increased, such that larger amounts of the virus can be produced in the host cells and then inactivated for use as a vaccine.

[00127] The methods of the present invention may also be used to optimize DNA vaccines and viral vector vaccines. For example, DNA vaccines or viral vector vaccines may comprise nucleotide sequences that encode certain immunogenic proteins in the context of a plasmid vector or viral vector backbone. The methods described above can be used to optimize expression of the nucleotide sequences that encode the immunogenic proteins, and also to optimize the sequence of the plasmid vector or viral vector backbone, for example by decreasing the expression of vector-encoded proteins.

[00128] The methods of the invention may also be used to optimize proteinaceous vaccines, such as proteinaceous vaccines produced by production of a recombinant proteins in a cellular host expression system. The methods described above can be used to optimized the nucleic acid encoding the protein for expression in the cellular host expression system.

**Mutation Methods**
In some embodiments, the present invention involves mutating nucleotide sequences to add/create or remove/disrupt sequence motifs. Such mutations can be made using any suitable mutagenesis method known in the art, including, but not limited to, site-directed mutagenesis, oligonucleotide-directed mutagenesis, positive antibiotic selection methods, unique restriction site elimination (USE), deoxyuridine incorporation, phosphorothioate incorporation, and PCR-based mutagenesis methods. Details of such methods can be found in, for example, Lewis et al. (1990) Nucl. Acids Res. 18, p3439; Bohnsack et al. (1996) Meth. Mol. Biol. 57, p1; Vavra et al. (1996) Promega Notes 58, 30; Altered Sites®II in vitro Mutagenesis Systems Technical Manual #TM001, Promega Corporation; Deng et al.. (1992) Anal. Biochem. 200, p81; Kunkel et al. (1985) Proc. Natl. Acad. Sci. USA 82, p488; Kunke et al. (1987) Meth. Enzymol. 154, p367; Taylor et al. (1985) Nucl. Acids Res. 13, p8764; Nakamaye et al. (1986) Nucl. Acids Res. 14, p9679; Higuchi et al. (1988) Nucl. Acids Res. 16, p7351; Shimada et al. (1996) Meth. Mol. Biol. 57, p157; Ho et al. (1989) Gene 77, p51; Horton et al. (1989) Gene 77, p61; and Sarkar et al. (1990) BioTechniques 8, p404. Numerous kits for performing site-directed mutagenesis are commercially available, such as the QuikChange® II Site-Directed Mutagenesis Kit from Stratagene Inc. and the Altered Sites® II in vitro mutagenesis system from Promega Inc. Such commercially available kits may also be used to optimize sequences.

Vaccines

The methods and composition of the present invention may be particularly useful for the production of vaccines. The low amounts of lentiviral particles produced during an infection cycle, coupled with their ability to remain intracellular for extended periods of time, limits the exposure of lentiviruses such as HIV to the immune system. This property is advantageous to the virus but adversely affects the ability to generate an effective vaccine. For example, viral vaccines that are designed to infect and replicate in host cells may produce low levels of progeny and remain hidden in host cells for extended periods of time. Consequently, such vaccines may not be able to effectively trigger an immune response and immunological memory. Similarly, DNA vaccines which encode one or more viral antigens are likely to express low levels of the antigen in the host, in turn limiting the effectiveness of the DNA vaccine in generating an immune response and immunological memory.

The methods of the present invention provide the ability to generate mutant viruses that have optimized nucleotide sequences and therefore have increased steady state
levels of viral RNA, increased expression of viral-encoded protein, increased infection cycles and increased exposure to the immune system. Such mutant viruses would be useful as viral vaccines. Vaccines that comprise, or are derived from, such mutant viruses are described in more detail below. The present invention also raises the possibility of generating mutant viral nucleic acid sequences that produce virally encoded proteins at a much higher rate, and/or in much larger quantities, than would otherwise be the case. Such mutant nucleic acids could be useful as DNA vaccines, as described in more detail below. Furthermore, such mutant nucleic acids could also be useful for production of viral proteins for use in protein vaccines. Vaccines that comprise, or are derived from, such proteins are also described in more detail below.

[00132] The present invention encompasses both prophylactic/preventive vaccines and therapeutic vaccines. A prophylactic vaccine is one administered to subjects who are not infected with the disease against which the vaccine is designed to protect. For example, a preventive vaccine may prevent a virus from establishing an infection in a vaccinated subject, i.e. it will provide complete protective immunity. However, even if it does not provide complete protective immunity, a prophylactic vaccine may still confer some protection to a subject. For example, a prophylactic vaccine may decrease the symptoms, severity, and/or duration of the disease. In the case of HIV, a prophylactic vaccine may prevent or delay the progression to full-blown AIDS even if it is not sufficient to provide complete protective immunity. A therapeutic vaccine, is administered to reduce the impact of a viral infection in subjects already infected with that virus. A therapeutic vaccine may decrease the symptoms, severity, and/or duration of the disease. In the case of HIV, administration of a therapeutic vaccine may prevent or delay the progression to full-blown AIDS.

[00133] The present invention encompasses any and all types of vaccine that comprise a nucleic acid having an optimized sequence, or that are produced from a nucleic acid having an optimized sequence. Several different types of vaccine are described herein. However, one of skill in the art will recognize that there are other types of vaccines that may be used, and other methods for producing vaccines. The present invention is not limited to the specific types of vaccines illustrated. Instead, it encompasses any and all vaccines that comprise a nucleic acid having an optimized nucleotide sequence, or that are produced from a nucleic acid having an optimized sequence.
[00134] The present invention encompasses viral vaccines. The term “viral vaccine” as used herein includes attenuated viral vaccines, inactivated viral vaccines and viral vector vaccines. The present invention also encompasses DNA vaccines and proteinaceous or “subunit” vaccines, each of which is described below. It should be noted that there is significant overlap among the various types of vaccines. For example, viral vaccines may comprise nucleic acids that are the same as, or similar to those used to make DNA vaccines. Similarly, DNA vaccines and viral vaccines may express proteins that are the same as, or similar to, those used to make proteinaceous vaccines. Thus, the description provided for any one type of vaccine below should not be construed as being useful for only that vaccine type. Instead all of the description regarding any one type of vaccine can be used and applied interchangeably to any and all of the types of vaccines encompassed by the present invention.

Attenuated viral vaccines

[00135] In one embodiment, the present invention provides attenuated viral vaccines having one or more optimized nucleotides. Attenuated viruses are viruses that have been altered to weaken them, such that they no longer cause disease, but may still stimulate an immune response. There are many ways in which a virus may be attenuated. For example, a virus can be attenuated by removal or disruption of viral sequences required for causing disease, while leaving intact those sequences encoding antigens recognized by the immune system. Attenuated viruses may or may not be capable of replication in host cells. Attenuated viruses that are capable of replication are useful because the virus is amplified in vivo after administration to the subject, thus increasing the amount of immunogen available to stimulate an immune response.

[00136] According to the present invention, a suitable attenuated viral strain may be obtained or generated and one or more of the nucleotide sequences in the attenuated viral strain mutated to an optimized nucleotide sequence. Several attenuated live viral vaccines have been shown to be useful in protecting against lentiviral infection. For example, live attenuated simian immunodeficiency viruses (SIV) have been used to protect primates against challenge with SIV. See, for example, Daniel et al., “Protective effects of a live attenuated SIV vaccine with a deletion in the nef gene” (1992) Science 258, p1938; Almond et al., “Protection by attenuated simian immunodeficiency virus in macaques against challenge with virus-infected cells” (1995) Lancet 345, p1342. The methods of attenuation and attenuated viral strains disclosed in these references may be used in conjunction with the present
invention. Other methods of attenuation have been described by Desrosiers et al. ("Identification of highly attenuated mutants of simian immunodeficiency virus" (1998) J. Virol. 72, p1431) and Guan et al. ("Construction and in vitro properties of a series of attenuated simian immunodeficiency viruses with all accessory genes deleted" (2001) J. Virol. 75, p4056). It should be noted that SIV is a commonly used model for HIV, and attenuation methods useful in SIV may also be useful for HIV. Published patent application WO/2001/007637 describes a live attenuated HIV vaccine modified to replicate only in the presence of a tetracycline analogue. Various other live attenuated HIV strains have been developed, for example delta 4, which is HIV-1 lacking the nef, vpr, vpu, and Nef-responsive element or NRE genes, and delta kURN, which is based on the delta 4 vaccine strain but has an additional deletion in the gene encoding the NFkB-binding element. There are also several articles describing how live attenuated HIV vaccines may be generated. See for example, Mills et al. "Live attenuated HIV vaccines: a proposal for further research and development." (2000) AIDS Res Hum Retroviruses 16, p1453. Any such methods for attenuation may be used in accordance with the present invention. If the attenuation methods used involve deletions within the lentiviral genome or within lentiviral nucleic acids, these mutations can be made large enough to reduce the chance reversion. For example, 20 bases or more can be deleted if such methods are used.

**Killed Viral Vaccines**

[00137] In another embodiment, the present invention provides killed or inactivated viral vaccines having one or more optimized nucleotide sequences. Such vaccines are generally non-functional and thus do not express viral genes or replicate in the vaccinated subject. However, the methods of the invention may be used to facilitate expansion and growth of virus in vitro or ex vivo prior to inactivation of the virus. For example, by optimizing one or more nucleotide sequences in a virus to an optimized nucleotide sequence, the rate of viral expansion may be increased such that larger amounts of the virus can be produced and then inactivated for use as a vaccine.

[00138] Any suitable method of inactivation known in the art may be used to inactivate the mutant viruses of the invention, such as chemical, thermal or physical inactivation or inactivation by irradiation with ionizing radiation. For example, Ilyinskii et al. have developed a physical inactivation method for HIV that utilizes gases to rupture/damage the virus structure in a way that renders it non-infective without comprising its tertiary structure
and possible immunogenicity (see Ilyinskii et al. “Development of an Inactivated HIV Vaccine” (2001) AIDS Vaccine Sep 5-8; abstract no. 192). Others have developed a method of inactivating the HIV virus chemically using 0.2% Beta-propiolactone (BPL) while retaining its immunogenicity (see Addawe et al. “Chemically inactivated whole HIV vaccine induces cellular responses in mice” (1996) Int Conf AIDS Jul 7-12; 11:4; abstract no. Mo.A.100). Whole-inactivated HIV vaccines have also been tested in human trials. For example, the therapeutic vaccine Remune® (also known as “HIV-1 Immunogen”, “Salk vaccine”, or “AG1661”) which is inactivated by a combination of chemical treatment and irradiation, has been studied as an immunotherapy in HIV-infected patients (see, Fernandez-Cruz et al. “5-year evaluation of a therapeutic vaccine (HIV-1 immunogen) administered with antiretrovirals in patients with HIV chronic infection: induction of long-lasting HIV-specific cellular immunity that impact on viral load” (2003) Second International AIDS Society Conference on HIV Pathogenesis and Treatment, Paris, abstract 486). The methods of the present invention can be used in conjunction with any of the above inactivation methods, or other viral inactivation methods known in the art.

**Viral Vector Vaccines**

[00139] The mutated viral nucleic acid sequences of the invention may also be incorporated into a viral vector suitable for administration to a subject. The viral nucleic acid may encode any lentiviral protein, including, but not limited to GAG, p17MA, p24CA, p7 and p6, GAG-POL, RT, RNAase H, PR, IN, Gp160, Gp120 ENV, Gp41, Tat, Rev, Vpu, Vif, Vpr and Nef, and fragments, variants, homologs and derivatives thereof. The advantage of using the methods of the present invention to optimize expression of an HIV, a lentiviral or a viral protein in a host tissue, a host, a host expression system, a human tissue, a human or a human expression system for the production of such vaccines, is that by optimizing one or more nucleotide sequences to optimized nucleotide sequences, the amount of protein produced and/or the rate of protein production may be substantially increased. Examples of suitable viral vectors include, but are not limited to, vaccinia viruses (such as Modified Vaccinia Virus Ankara or “MVA”, the highly attenuated strain of vaccinia used in smallpox vaccines), retroviruses, poxviruses (including canarypox, vaccinia, and fowlpox) adenoviruses and adeno-associated viruses. These viral vectors may be altered compared to their natural viral counterparts, for example they may be attenuated and/or non-replicative.
One of skill in the art can readily select a suitable viral vector and insert the mutant nucleic acids of the invention into such a vector. The mutant nucleic acid should be under the control of a suitable promoter for directing expression of the viral protein in vaccinated subjects. A promoter that is already present in the viral vector may be used. Alternatively, an exogenous promoter may be used. Examples of suitable promoters include, but are not limited to, the cytomegalovirus (CMV) promoter, the rous sarcoma virus (RSV) promoter, the HIV long terminal repeat (HIV-LTR), the HTLV-1 LTR (HTLV-LTR) and the herpes simplex virus (HSV) thymidine kinase promoter.

Techniques that can be used to insert the nucleic acid sequences of the invention into the viral expression vectors are well known to those of skill in the art. See for example Sambrook et al. (2001) Molecular Cloning: A Laboratory Manual, 3rd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

The methods of the present invention may also be used in conjunction with, or as an improvement to any type of viral vector vaccine known in the art. Examples of HIV viral vector vaccines that are currently in development include Merck’s non-replicating adenoviral vector containing HIV clade B GAG-POL Nef, Sanofi Pasteur’s canarypox vector containing clade B Env, GAG, Pro, RT, and Nef, and Therion’s MVA vector containing clade B Env and GAG. Details of these and other HIV vaccines currently in development are provided by the HIV Vaccine Trials Network at www.hvtn.org. The methods of the present invention could be used to improve the efficacy of viral vector vaccines such as these by optimizing nucleotide sequences within the lentiviral nucleic acid components to optimized nucleotide sequences, leading in turn to improved expression of the lentiviral proteins in the vaccinated subjects.

DNA Vaccines

The present invention also encompasses DNA vaccines suitable for administration to subjects. A mutated viral nucleic acid encoding any viral protein, or portion, fragment, derivative or homolog thereof, may be inserted into a DNA plasmid or expression vector in order to make a DNA vaccine according to the present invention. For example, in one embodiment, the DNA vaccine comprises a plasmid containing one or more mutated lentiviral nucleic acids encoding proteins selected from the group consisting of GAG, p17MA, p24CA, p7 and p6, GAG-POL, RT, RNAase H, PR, IN, Gp160, Gp120 ENV, Gp41,
Tat, Rev, Vpu, Vif, Vpr and Nef, and fragments, variants, homologs and derivatives thereof. The advantage of using the methods of the present invention to optimize expression of an HIV, a lentiviral or a viral protein in a host tissue, a host, a host expression system, a human tissue, a human or a human expression system for the production of such vaccines, is that by optimizing one or more nucleotide sequences to optimized nucleotide sequences, the amount of protein produced and/or the rate of protein production may be substantially increased.

[00144] One of skill in the art can readily select a suitable DNA plasmid or expression vector and insert the mutant nucleic acids of the invention into such a plasmid or expression vector. The nucleic acid encoding the viral protein should be under the control of a suitable promoter for directing expression of the mutated nucleic acid in the vaccinated subjects. A promoter that is already present in the expression vector may be used. Alternatively, an exogenous promoter may be used. Examples of suitable promoters include, but are not limited to, the cytomegalovirus (CMV) promoter, the rous sarcoma virus (RSV) promoter, the HIV long terminal repeat (HIV-LTR), the HTLV-1 LTR (HTLV-LTR) and the herpes simplex virus (HSV) thymidine kinase promoter.

[00145] Techniques that can be used to insert the nucleic acid sequences of the invention into DNA plasmids and expression vectors are well known to those of skill in the art. For example, standard recombinant DNA techniques that may be used are described in Sambrook et al. (2001) Molecular Cloning: A Laboratory Manual, 3rd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

[00146] The methods of the present invention may also be used in conjunction with, or as an improvement to, any type of viral DNA vaccine known in the art. Examples of DNA vaccines that are currently in development include an NIH DNA plasmid containing clade B Gag, Pol, Nef, and clade A, B, and C, Env, Chiron’s DNA plasmid containing clade B Gag and Env, and GENEVAX which is a DNA plasmid containing clade B Gag. Details of these and other HIV vaccines currently in development are provided by the HIV Vaccine Trials Network (HVTN) at www.hvtn.org. The methods of the present invention could be used to improve the efficacy of vaccines such as these by optimizing nucleotide sequences within the viral nucleic acid components to optimized sequences, leading in turn, to improved expression of the viral proteins in the vaccinated subjects.

Protein Vaccines
[00147] The present invention also encompasses proteinaceous vaccines. Any viral protein, or fragment, derivative, variant or homolog thereof, may be used to make a proteinaceous vaccine according to the present invention. For example, in one embodiment, the lentiviral protein is selected from the group consisting GAG, p17MA, p24CA, p7 and p6, GAG-POL, RT, RNAase H, PR, IN, Gp160, Gp120 ENV, Gp41, Tat, Rev, Vpu, Vif, Vpr and Nef, and fragments, variants, homologs and derivatives thereof. The advantage of using the methods of the present invention to optimize expression of an HIV, a lentiviral or a viral protein in a host tissue, a host, a host expression system, a human tissue, a human or a human expression system for the production of such vaccines, is that by optimizing one or more nucleotide sequences to optimized nucleotide sequences, the amount of protein produced and/or the rate of protein production may be substantially increased.

[00148] In one embodiment, a viral nucleic acid optimized by the methods of the invention is incorporated into a suitable expression vector to allow for expression of the protein in a suitable expression system. Examples of suitable expression systems include, but are not limited to, cultured mammalian, insect, bacterial, or yeast cells. The viral protein can then be expressed in the expression system, purified, and used to make a proteinaceous vaccine.

[00149] Any plasmid or expression vector may be used provided that it contains a promoter to direct expression of the viral protein in the desired expression system. For example, if the protein is to be produced in bacterial cells, a promoter capable of directing expression in bacteria should be used, if the protein is to be produced in mammalian cells, a promoter capable of directing expression in mammalian cells should be used, if the protein is to be produced in insect cells, a promoter capable of directing expression in insect cells should be used, if the protein is to be produced in yeast, a promoter capable of directing expression in yeast should be used. In one embodiment, the proteins encoded by optimized nucleic acid sequences are expressed in a mammalian expression system from a mammalian promoter. Suitable promoters include, but are not limited to, the cytomegalovirus (CMV) promoter, the rous sarcoma virus (RSV) promoter, the HIV long terminal repeat (HIV-LTR), the HTLV-1 LTR (HTLV-LTR), the herpes simplex virus (HSV) thymidine kinase promoter, and the SV40 virus early promoter. Suitable expression vectors include but are not limited to cosmids, plasmids, and viral vectors such as replication defective retroviruses, adenoviruses, adeno-associated viruses, herpes viruses, vaccinia viruses, attenuated vaccinia viruses, canary
pox viruses, lentiviruses and herpes viruses, among others. Commercially available
expression vectors which already contain a suitable promoter and a cloning site for addition
of exogenous nucleic acids may also be used.

[00150] Any suitable expression system may be used, such as bacterial, yeast, insect,
or mammalian cellular expression systems. In one embodiment, the viral proteins encoded by
optimized nucleic acid sequences are expressed in mammalian cells that have been either
stably or transiently transfected with the mutant viral nucleic acids of the invention.
Examples of suitable mammalian cells that can be used include, but are not limited to, COS,
CHO, BHK, HEK293, VERO, HeLa, MDCK, WI38, and NIH 3T3 cells. Primary or
secondary cells obtained directly from a mammal, engineered to contain the mutant nucleic
acids of the invention may also be used as an expression system.

[00151] One of skill in the art can readily select a suitable expression system, promoter
and expression vector for use in accordance with the present invention. Examples of
workable combinations of cell lines and expression vectors are described in Sambrook et al.
Cold Spring Harbor, N.Y. Techniques that can be used to insert the nucleic acid sequences of
the invention into an expression vector are well known to those of skill in the art. See, for
Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

[00152] The methods of the present invention may also be used in conjunction with, or
as an improvement to, any type of proteinaceous vaccine known in the art. Examples of
proteinaceous vaccines that are currently in development include Chiron’s protein subunit
clade B Env, and GlaxoSmithKline’s clade B Nef-Tat fusion protein and clade B Env
subunit). The methods of the present invention could be used to improve the efficacy of
production of vaccines such as these by optimizing nucleotides sequences within the nucleic
acid that encodes the various protein subunits to optimized nucleotide sequences.

Vaccine Compositions

[00153] The vaccine compositions of the invention comprise at least one virus
(including attenuated viruses, inactivated viruses, and viral vectors), nucleic acid, or protein,
such as those described above. The compositions may also comprise one or more additional
components including, but not limited to, pharmaceutically acceptable carriers, buffers,
stabilizers, diluents (such as water), preservatives, solubilizers, or immunomodulatory agents. Suitable immunomodulatory agents include, but are not limited to, adjuvants, cytokines, polynucleotides encoding cytokines, and agents that facilitate recognition by the immune system of at least one component of the vaccines of the invention. One of skill in the art can readily select suitable additives for inclusion in the vaccine compositions of the invention.

**Effective Amounts**

[00154] An immunologically effective amount of the vaccine compositions of the invention should be administered to a subject. As used herein, the term “immunologically effective amount” refers to an amount capable of inducing, or enhancing the induction of, the desired immune response in a subject. The desired response may include, inter alia, inducing an antibody or cell-mediated immune response, or both, reducing viral load, ameliorating the symptoms of infection, delaying the onset of symptoms, reducing the duration of infection, and the like. An immunologically effective amount may also be an amount sufficient to induce protective immunity.

[00155] One of skill in the art can readily determine what is an immunologically effective amount without undue experimentation. For example, an effective amount can be determined by conventional means, starting with a low dose of and then increasing the dosage while monitoring the immunological effects. Numerous factors can be taken into consideration when determining an optimal amount to administer, including the size, age, and general condition of the subject, the presence of other vaccines or drugs in the subject, the virulence of the particular virus against which the subject is being vaccinated, and the like. The actual dosage can be chosen after consideration of the results from various animal studies.

**Routes of Delivery/Administration Regimens**

[00156] The vaccine compositions of the invention may be administered in a single dose, multiple doses, or using prime-boost regimens. When prime-boost regimens are used, the vaccines of the present invention may be use as the priming agent or the boosting agent or both. The compositions may be administered by any suitable route, including, but not limited to, parenteral, intradermal, transdermal, subcutaneous, intramuscular, intravenous, intraperitoneal, intranasal, oral, or intraocular routes, or by a combination of routes. The compositions may also be administered using a gun device which fires particles, such as gold
particles, onto which compositions of the present invention have been coated, into the skin of a subject. The skilled artisan will be able to formulate the vaccine composition according to the delivery route chosen.

Other embodiments of the invention

[00157] In other embodiments, the invention is directed to methods for identifying agents that inhibit or stimulate production of viral RNA, production of viral protein or production of viral particles, or that inhibit or stimulate viral latency. In one embodiment, the method comprises providing a control cell containing at least one non optimized viral nucleic acid sequence, and a test cell containing at least one viral nucleic acid sequence containing at least one nucleotide sequence mutated to an optimized nucleotide sequence, contacting the test cell and the control cell with one or more agents, and identifying at least one agent that inhibits or stimulates production of viral RNA, production of virus protein or production of virus particles, or that inhibits or stimulates viral latency, in the test cell as compared to the control cell. In some embodiments, the agents inhibit or stimulate production of HIV or lentivirus RNA, production of HIV or lentivirus protein or production of HIV or lentivirus particles, or inhibit or stimulate HIV or lentivirus latency. For example, entire libraries of agents can be screened in this way using high throughput screening methods. One of skill in the art could readily design a high through put screening method to identify agents that inhibit or stimulate production of viral RNA, production of virus protein or production of virus particles, or that inhibit or stimulate viral latency. Methods for growing cells in multiwell plates are well known, and methods for administering different agents from a library of agents to different wells of multiwell plates are known. Several methods could be used to determine the effects of the library agents on production of viral RNA, production of viral protein or production of viral particles, or on viral latency. For example, the cells used for the high throughput screening could be engineered to encode one or more fusion proteins, such as a fusion between a viral protein and a fluorescent protein such as green fluorescent protein (GFP). In this way, production of viral proteins could be monitored by fluorescent detection methods, which would enable agents that stimulate or inhibit production of the viral protein to be detected.

[00158] In another embodiment, the invention is directed to methods for identifying agents that bind differentially to non optimized and optimized nucleotide sequences. In one embodiment, the method comprises providing a non optimized control nucleic acid and a test
nucleic acid containing at least nucleotide sequence that has been mutated to an optimized sequence, contacting the test nucleic acid and the control nucleic with one or more agents, and identifying at least one agent that binds to the control nucleic acid but does not bind the test nucleic acid, or that binds to the control nucleic acid with a higher affinity than it binds to the test nucleic acid.

[00159] In another embodiment, the method comprises providing a test nucleic acid containing an optimized nucleotide sequence and a control nucleic acid containing a random assortment and order of nucleotides, contacting the test nucleic acid and the control nucleic with one or more agents, and identifying at least one agent that binds to the test nucleic acid but does not bind the test control acid, or that binds to the test nucleic acid with a higher affinity than it binds to the control nucleic acid. There are multiple ways in which agents that bind to these constructs could be detected. For example, in one embodiment, the above test and control nucleic acids could be provided on a column or one some other suitable solid substrate, and test samples (such as cell lysates or libraries of test agents) could be passed over these substrates. Agents that bind to the test and/or control substrates could be eluted and analyzed. In other embodiments, yeast one-hybrid methods could be used to identify agents that bind to optimized nucleotide sequences. In further embodiments, electrophoretic mobility shift assays (EMSAs) could be performed to identify agents that bind to optimized nucleotide sequences. Other methods suitable for identifying nucleotide binding agents are known in the art, and any such method could be used to identify agents that bind to optimized nucleotide sequences. The present invention also encompasses optimized nucleotide sequence binding agents, such as those identified using the methods of the invention.

[00160] In yet another embodiment, the invention is directed to agents that inhibit or stimulate binding of an optimized nucleotide sequence binding agent to a nucleic acid containing at least one optimized nucleotide sequence, and to methods for identifying such agents as described above.

Other applications

[00161] The methods of the present invention have numerous other uses including, but not limited to, optimization of splice sites, optimization of exon splicing enhancers, optimization of real exons, optimization of mRNA degradation or stabilization signals,
optimization of transcription factor binding sites, and optimization of sequences associated with tissue specific expression.

[00162] The methods of the invention could be used to identify sequences that are over- or underrepresented in non optimized or optimized nucleotide sequences. For example, real exons are known to have overrepresented signals such as exon splicing enhancers. Such sequence motifs would be useful for helping to determine whether a given sequence is a real exon sequence or a confounding intronic sequence.

[00163] In another embodiment, the invention is directed to methods for identifying agents that bind differentially to optimized and non-optimized nucleotide sequences. In one preferred embodiment, the method comprises providing a non-optimized control nucleic acid sequence and a test nucleic acid sequence containing at least one optimized nucleic acid residue, contacting the test optimized nucleic acid and the control non-optimized nucleic acid with one or more agents, and identifying at least one agent that binds differentially to the control non-optimized nucleic acid but does not bind the test optimized nucleic acid, or that binds to the control non-optimized nucleic acid with different affinity than it binds to the test optimized nucleic acid. There are multiple ways in which agents that bind to these constructs could be detected. For example, in one embodiment, the above test and control nucleic acids could be provided on a column or one some other suitable solid substrate, and test samples (such as cell lysates or libraries of test agents) could be passed over these substrates. Agents that bind to the test and/or control substrates could be eluted and analyzed. In other embodiments, yeast one-hybrid methods could be used to identify agents that bind to optimized and non-optimized nucleic acids in a differential manner. In further embodiments, electrophoretic mobility shift assays (EMSAs) could be performed to identify agents that bind to optimized and non-optimized nucleic acids in a differential manner. Other methods suitable for identifying nucleotide binding agents are known in the art, and any such method could be used to identify agents that bind to optimized and non-optimized nucleic acids in a differential manner. The present invention also encompasses optimized nucleic acid binding agents, such as those identified using the methods of the invention.

[00164] These and other embodiments of the invention are further described in the following non-limiting examples. It should also be understood that numerous other variations of the embodiments described herein, including variations in the methods described herein,
are possible without departing from the spirit or scope of the invention. Such variations will be apparent to those of skill in the art.
EXAMPLES

Example 1

[00165] The method described herein was used to optimize one HIV sequence, namely that encoding Gag, for expression in human muscle. It is believed that this sequences will be more effective than a native HIV sequence when used to immunize a human subject, particularly when delivered intramuscularly. Figure 1 provides the nucleotide sequence of the non-optimized HIV Gag gene, which has the sequence identifier SEQ ID NO: 1. This amino acid and nucleotide coding sequence of this protein is publicly available in the NCBI database under the accession number BAA00992.1. Figure 2 provides the nucleotides sequence of the optimized HIV Gag gene, which has the sequence identifier SEQ ID NO: 2. The sequences of the pre-optimized HIV Gag, Pol and Nef genes can also be obtained form the Santa Cruz Genome browser (http://genome.ucsc.edu).

[00166] The ranking of the template genes for group “t” were selected on the basis of expression information in human cardiac tissue comparing RNA expression and genomic DNA via microarray analysis. To rank the genes, the median (CH11_MEAN, Mean feature pixel intensity at wavelength 532 nm.) over (CH21_MEAN, Mean feature pixel intensity at wavelength 635 nm.) was used to obtain a relative ratio for gene expression. The data used in the ranking and the parameters of the microarray platform are available through the NCBI Gene expression Omnibus (http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSM40007). In order to avoid any splice signals in the RNA transcripts, only genes containing an first exon of greater than a defined number of nucleotides were selected for ranking among the group of ranked genes “t”. This values was 1500 nucleotides for Gag.
CLAIMS

What is claimed is:

1. A method for optimizing a nucleotide sequence for expression in an organism or tissue of interest, the method comprising:

   (i) providing a nucleotide sequence “b”, which encodes an amino acid sequence “s”, for optimization,

   (ii) providing a group of one or more genes “tn” from the organism or tissue of interest use as templates to optimize the nucleotide sequence,

   (iii) ranking the group of one or more genes in “tn” according to their expression level in the organism or tissue of interest to create a group of ranked genes “t”,

   (iv) selecting an amino acid word “w0” of length “L” from the amino acid sequence “s”,

   (v) sequentially examining the ranked genes from group “t” for a nucleotide sequence, that can potentially encode the amino acid word “w0” without regard to the translation frame,

   (vi) eliminating the C-terminal amino acid from the polypeptide sequence “w0” and adding the next N-terminal amino acid in the corresponding amino acid sequence in “s” to generate amino acid sequence “w1”,

   (vii) eliminating the three nucleotides at the 5’ end of the nucleotide sequence “n0” to generate a new nucleotide sequence “n01”,

   (viii) replacing the nucleotide sequence of “n0” with the nucleotide sequence of “n1”, and

   (ix) recursively repeating the steps (v) to (viii) until the end of the amino acid sequence “s”, thereby generating the optimized nucleotide sequence “o”.

2. The method of claim 1, wherein the nucleotide sequence “b” encodes a therapeutic protein.
3. The method of claim 1, wherein the nucleotide sequence “b” encodes an immunogenic protein.

4. The method of claim 1, wherein the nucleotide sequence “b” is selected from the group consisting of: the genome of a eukaryotic organism, the genome of a prokaryotic organism, the genome of a virus, an expression vector sequence, a plasmid sequence, a cloned cDNA sequence, and the an expressed sequence tag (EST).

5. The method of claim 1, wherein 1 to 20 genes are ranked in step (iii).

6. The method of claim 1, wherein 21 to 100 genes are ranked in step (iii).

7. The method of claim 1, wherein 101 to 10000 genes are ranked in step (iii).

8. The method of claim 1, wherein more than 10000 genes are ranked in step (iii).

9. The method of claim 1, wherein the word length “l” is two amino acids.

10. The method of claim 1, wherein the word length “L” is three amino acids in length.

11. The method of claim 1, wherein the word length “L” is four amino acids in length.

12. The method of claim 1, wherein the word length “L” is five amino acids in length.

13. The method of claim 1, wherein the word length “L” is five amino acids in length.

14. The method of claim 1, wherein the word length “L” is six amino acids in length.

15. The method of claim 1, wherein the word length “L” is seven amino acids in length.

16. The method of claim 1, wherein the word length “L” is eight amino acids in length.

17. The method of claim 1, wherein the word length “L” is nine amino acids in length.

18. The method of claim 1, wherein the word length “L” is ten or more amino acids in length.

19. The method of claim 1, wherein the word length “L” is not a fixed value and is established by recursively scanning the group of ranked nucleotide sequences in group “t” for the longest possible “word size” by beginning with the longest possible “word size” in the
amino acid sequence “s” and shortening the “word size” of “w0” by removing the C-terminal amino acid of “w0” recursively until a nucleotide sequence in the group of ranked nucleotide sequences “t” is identified.

20. The method of claim 1, wherein the optimized nucleotide sequence “o” contains one or more mRNA stability signals, signals that increase the rate of transcription, signals that increase protein translation, protein binding sites, transcription factor binding sites, promoter sequences, enhancer sequences, or splice sites not present in the starting nucleotide sequence “b”.

21. The method of claim 1, wherein the optimized nucleotide sequence “o” differs from the nucleotide sequence “b” by one or more of the sequence motifs identified is an mRNA stability signal, an mRNA instability signal, a signal that increases the rate of transcription, a signal that decreases the rate of transcription, a signal involved in protein translation, a protein binding site, a transcription factor binding site, a promoter sequence, an enhancer sequence, a repressor sequence, a silencer sequence, a splice site, a restriction enzyme site, or a viral latency signal.

22. The method of claim 1, wherein step (ix) is repeated one time.

23. The method of claim 1, wherein step (ix) is repeated multiple times.

24. The method of claim 1, wherein step (ix) is repeated until the end of the amino acid sequence “s” is reached.

25. The method of claim 1 wherein steps (i) through (ix) are performed multiple times with different ranked genes or gene rankings or criteria for generating the group “t”.

26. The method of claim 1, wherein the steps (i) through (ix) are independently repeated for a nucleotide sequence “b” using two or more ranking criteria or with different genes in group “t” to generate multiple optimized nucleotide sequences “o” and the multiple optimized nucleotide sequences “o” are used to generate a consensus sequence “o”.

27. The method of claim 1, wherein the genes in group “t” comprise genes from the same host.
28. The method of claim 1, wherein the genes in group “t” comprise genes from different hosts.

29. The method of claim 1, wherein the genes in group “t” comprise artificial genes.

30. The method of claim 1, wherein the genes in group “t” comprise recombinant genes.

31. The method of claim 1, wherein the genes in group “t” comprise theoretical genes.

32. The method of claim 1, wherein the ranking is based on experimental data.

33. The method of claim 1, wherein the ranking is based on information available in a repository.

34. The method of claim 1, wherein the ranking is based on a theoretical model.

35. The method of claim 1, wherein the word “w0” is selected only from coding regions of the nucleotide sequences of genes in group “t”.

36. The method of claim 1, wherein the nucleotide sequences in step (v) exist in the same translation frame and actually encode the same amino acid sequence word “w0”.

37. The method of claim 1, wherein the genes in group “t” are ranked on a set of one or more criteria comprising gene expression, mRNA abundance, mRNA steady state levels, mRNA stability, mRNA export, mRNA translation, mRNA localization, mRNA transcription, mRNA splicing, mRNA dicing, mRNA secondary structure, mRNA tertiary structure, mRNA binding to agents, mRNA binding to proteins, mRNA binding to other RNAs, mRNA binding to DNA, sequence homology, phylogenetic analysis, the presence or absence of sequence motifs, RNA or DNA topology, RNA or DNA architecture, protein expression, the organization of sequence motifs, binding to transcription factors, binding to enhancers, binding to repressors, specificity of expression, expression patterns, functional significance, biological significance, developmental significance, pathological significance, disease significance, infection significance, virulence, infectivity, replication, gene function or any combination thereof.

38. A method for optimizing the production of a protein in an organism or tissue of interest, the method comprising:
(a) selected a nucleotide sequence to be optimized, and

(b) optimizing the nucleotide sequence using the method of claim 1.

39. The method of claim 38, wherein the protein is a therapeutic protein.

40. The method of claim 38, wherein the protein is an immunogenic protein.

41. An optimized HIV sequence for use in an HIV vaccine composition wherein the HIV sequence comprises a sequence from the optimized Gag sequence SEQ ID NO. 1, and any fragments thereof.

42. A method for optimizing the production of a protein in a host, the method comprising:

(a) obtaining a nucleotide sequence encoding a protein to be expressed in the host,

(b) optimizing the sequence of the nucleotide sequence by the method of claim 1, wherein the optimization result in improved production of the protein in the host.

43. The method of claim 42, wherein the word size used is two or more amino acids.

44. The method of claim 42, wherein the wherein the nucleotide sequence “b” selected from the group consisting of: the genome of a eukaryotic organism, the genome of a prokaryotic organism, the genome of a virus, an expression vector, a plasmid, a cloned cDNA, and an expressed sequence tag (EST).

45. The method of claim 42, wherein the wherein the genes in group “t” are selected from the group consisting of: the genome of a eukaryotic organism, the genome of a prokaryotic organism, the genome of a virus, an expression vector, a plasmid, a cloned cDNA, and an expressed sequence tag (EST).

46. The method of claim 42, wherein the amino acid sequence encoded by the nucleotide sequences “b” and “o” are identical.

47. The method of claim 42, wherein the protein is a therapeutic protein.

48. The method of claim 42, wherein the protein is an immunogenic protein.
49. The method of claim 42, wherein the protein is suitable for use in a vaccine composition.

50. The method of claim 42, wherein the nucleotide sequence encoding the protein is located in, or may be inserted into, a vector.

51. The method of claim 50, wherein the vector is an expression vector.

52. The method of claim 51, wherein the expression vector is adapted for administration to the host as a vaccine.

53. The method of claim 52, wherein the vector is a viral vector.

54. The method of claim 52, wherein the viral vector is adapted for administration to the host as a vaccine.

55. The method of claim 42, wherein the nucleotide sequence encoding the protein is located in, or may be inserted into, a recombinant virus.

56. The method of claim 55, wherein the recombinant virus is adapted for administration to the host as a vaccine.

57. The method of claim 55, wherein the recombinant virus is an attenuated virus.

58. The method of claim 42, wherein the host is a eukaryote or a eukaryotic cell.

59. The method of claim 42, wherein the host is a prokaryote or a prokaryotic cell.

60. The method of claim 42, wherein the host is a bacterium.

61. The method of claim 42, wherein the host is a yeast cell.

62. The method of claim 42, wherein the host is a mammal or a mammalian cell.

63. The method of claim 42, wherein the host is a primate or a primate cell.

64. The method of claim 42, wherein the host is a human or a human cell.

65. The method of claim 42, wherein the host is a mouse or a mouse cell.

66. The method of claim 42, wherein the host is a goat or a goat cell.
67. The method of claim 42, wherein the host is a sheep or a sheep cell.

68. The method of claim 42, wherein the host is a bird or a bird cell.

69. The method of claim 42, wherein the host is a chicken or a chicken cell.

70. The method of claim 42, wherein the host is an insect or an insect cell.

71. The method of claim 42, wherein the host is a transgenic animal or a cell from a transgenic animal.

72. The method of claim 42, wherein the host is a cell from a cultured cell line.

73. The method of claim 72, wherein the cell line is selected from the group consisting of: a Chinese hamster ovary (CHO) cell line, the mouse myeloma NS0 cell line, a baby hamster kidney (BHK) cell line, the human embryo kidney 293 (HEK-293) cell line, a chicken embryo fibroblast cell line, the human C6 cell line, a Madin-Darby canine kidney (MDCK) cell line, and the Sf9 insect cell line.

74. A method for identifying an agent that affects virus RNA production, virus protein, virus particle production or inhibition or stimulation of viral latency, the method comprising,

(a) providing a control cell containing at least one non-optimized viral nucleic acid sequence and a test cell containing at least one optimized viral nucleic acid sequence,

(b) contacting the control cell and the test cell with one or more agents,

(c) measuring virus RNA production, virus protein, virus particles production or inhibition or stimulation of viral latency, and

(d) comparing the measured virus RNA production, virus protein, virus particles production or inhibition or stimulation of viral latency in the test cell and the control cell and identifying at least one agent that affects virus RNA production, virus protein, virus particles production or inhibition or stimulation of viral latency, wherein and increase or decrease in virus RNA production, virus protein, virus particles production or inhibition or stimulation of viral latency indicates that the agent affects virus RNA production, virus protein, virus particles production or inhibition or stimulation of viral latency.
75. A method for identifying an agent that binds differentially to an non-optimized nucleotide sequence and a corresponding optimized nucleotide, the method comprising,

(a) providing at least one non-optimized nucleic acid sequence and a least one optimized nucleic acid sequence,

(b) contacting the non-optimized nucleotide sequence and the optimized nucleotide sequence with one or more agents,

(c) measuring binding of the agent to the optimized and non-optimized nucleic acid sequences, and

(d) comparing the measured binding of the agent to the optimized and non-optimized nucleic acid sequences in and identifying at least one agent that binds differentially to the optimized and non-optimized nucleic acid sequences, wherein an increase of decrease in binding of the agent with the optimized nucleic acid sequence compared to the non-optimized nucleic acid sequence indicates that the agent binds differentially to the optimized and non-optimized nucleic acid sequences.

76. The method of any of claims 74-75, wherein the agent is a small molecule, a therapeutic compound, a nucleic acid, a protein, a cell lysate or a cellular fraction.

77. An agent that inhibits or stimulates binding of an agent to an optimized nucleic acid sequence identified using any of the methods of claims 74-75.

78. A method for identifying changes in protein characteristics that arise as a result of the introduction of silent mutations during optimization of a nucleic acid encoding the protein, the method comprising,

(a) providing a protein translated by a from a non-optimized nucleic acid sequence encoding the protein and a protein translated from an optimized nucleic acid sequence,

(b) measuring the characteristics of the protein translated from the non-optimized protein sequence and the characteristic of the protein translated from the optimized protein sequence, and
(c) comparing the measured characteristics of the protein translated from the non-optimized protein sequence and the characteristic of the protein translated from the optimized protein sequence, wherein a difference in the characteristic of the protein translated from the non-optimized protein sequence and the characteristic of the protein translated from the optimized protein sequence indicates that the introduction of silent mutations during optimization of a nucleic acid change protein characteristics.

79. The method of claim 78, wherein the protein characteristics include: tertiary structure, immunogenicity, incorporation into a viral structure, protein expression, protein stability, enzymatic activity, polymerase activity, transcriptase activity, protein binding, small molecule binding, binding to an agent or protein localization.

80. The method of any of claim 4, 44, 45, 55, 56, 57 or 74, wherein the virus is from the Adenovirus family, Alfamovirus family, Allexivirus family, Allolevivirus family, Alphacryptovirus family, Alphalipotrixivirus family, Alphanodoavirus family, Alphapapillomavirus family, Alpharetrovirus family, Alphavirus family, Amdovirus family, Ampelovirus family, Aphthovirus family, Aquabirnavirus family, Aquareovirus family, Arenavirus family, Arterivirus family, Ascovirus family, Asfivirus family, Atadenovirus family, Aureusivirus family, Avastrovirus family, Avenavirus family, Aviadenovirus family, Avibirnavirus family, Avihepadnavirus family, Avipoxvirus family, Avulavirus family, Babuvirus family, Badnavirus family, Barnavirus family, Bdellicomovirus family, Begomovirus family, Betacryptovirus family, Betalipotrixivirus family, Betanodovirus family, Betapapillomavirus family, Betaretrovirus family, Betatetravirus family, Bocavirus family, Bornavirus family, Bracovirus family, Brevidensovirus family, Bromovirus family, Bymovirus family, Capillovirus family, Capripoxvirus family, Cardiovirus family, Carlavirus family, Carmovirus family, Caulimovirus family, Cavemovirus family, Chlamydiamicovirus family, Chlorovirus family, Chloriridovirus family, Chrysovirus family, Circovirus family, Closterovirus family, Coccolithovirus family, Coltivirus family, Comovirus family, Coronavirus family, Corticovirus family, Cripavirus family, Cucumovirus family, Curtovirus family, Cypovirus family, Cystovirus family, Cytomegalovirus family, Cytorhabdovirus family, Dainthovirus family, Deltapapillomavirus family, Deltaretrovirus family, Densovirus family, Dependovirus family, Ebolavirus family, Enamovirus family, Enterovirus family, Entomobirnavirus family, Entomopoxvirus family A, Entomopoxvirus family B,
Entomopoxvirus family C, Ephemeroirus family, Epsilonpapillomavirus family, Epsilonretrovirus family, Erbovirus family, Errantivirus family, Erythroivirus family, Etapapillomavirus family, Fabavirus family, Fijivirus family, Flavivirus family, Foveavirus family, Fusellovirus family, Gammalipotrichivirus family, Gammapapillomavirus family, Gammaretrovirus family, Giardiavirus family, Granulovirus family, Guttaviruses family, Gyrovirus family, Hantavirus family, Hemivirus family, Henipavirus family, Hepacivirus family, hepadnavirus family, Hepatovirus family, Hypovirus family, Ichnovirus family, Ictalurivirus family, Idenoovirus family, Ilarvirus family, Iltovirus family, Influenza A virus family, Influenza B virus family, Influenza C virus family, Inovirus family, Iotapapillomavirus family, Ipomovirus family, Iridovirus family, Isavirus family, Iteravirus family, Kappapapillomavirus family, Kobuvirus family, Lagovirus family, Lambda papillomavirus family, Leishmaniavirus family, Lentivirus family, Leporipoxvirus family, Levivirus family, Luteovirus family, Lymphocryptovirus family, Lymphocystivirus family, Lyssavirus family, Machlomovirus family, Macluravirus family, Maculavirus family, Mamastrovirus family, Mandarivirus family, Marafivirus family, Marburgvirus family, Mardivirus family, Mamnavirus family, Mastadenovirus family, Mastrevirus family, Megalocytivirus family, Metapneumovirus family, Metavirus family, Microvirus family, Mitovirus family, Molluscipoxvirus family, Morbillivirus family, Mupapillomavirus family, Muromealovirus family, Myxocovirus family, Nairovirus family, Nannovirus family, Narnavirus family, Necrovirus family, Nepovirus family, Norovirus family, Novirhabdovirus family, Nucleopolyhedrovirus family, Nucleorhabdovirus family, Nupapillomavirus family, Okavirus family, Oleavirus family, Omegatetravirus family, Omikronpapillomavirus family, Orbivirus family, Orthobunyavirus family, Orthohepadnavirus family, Orthopoxvirus family, Orthoreovirus family, Oryzavirus family, Panicovirus family, Parapoxvirus family, Parechovirus family, Parvovirus family, Pefudensovirus family, Pestivirus family, Petuvirus family, Phacovirus family, Phlebovirus family, Phytoreovirus family, Pipapillomavirus family, Plasmavirus family, Plectrovi, Pneumovirus family, Polerovirus family, Polyomavirus family, Potexvirus family, Potyvirus family, Prasinovirus family, Pymnesiovirus family, Pseudovirus family, Ranavirus family, Raphidovirus family, Respirovirus family, Rhadinovirus family, Rhinovirus family, Roseolovirus family, Rotavirus family, Rubivirus family, Rubulavirus family, Rudivirus family, Rymovirus family, Sapovirus family, Sedornavirus family, Sequivirus family, Siadenovirus family, Simplexvirus family, Soymovirus family, Spiromicrovirus family, Spumavirus family, Suipoxvirus family, Tectivirus family, Teschovirus family, Thetapapillomavirus family,
Thogotovirus family, Tombusvirus family, Topocuvirus family, Torovirus family, Tospovirus family, Totivirus family, Trichovirus family, Tritimovirus family, Tungrovirus family, Tymovirus family, Varicellovirus family, Vesiculovirus family, Vesivirus family, Vitivirus family, Waikavirus family, Whisvivirus family, Xipapillomavirus family, Yatapoxvirus family, Zetapapillomavirus family or any combination thereof.
SEQ ID NO: 1
HIV-1 GAG Non-optimized

1 ATGGGTGCGA GAGCGTCAGT ATTAAGCGGG GGAGAATTAG ATAAATGGGA AAAAAATTGG
61 TTAAGGCAG GGGGAAAGGA AAAAAATAAA TTTAAACTAA TAGTTGGGGA TCACAGGAGG
121 CTAGAACGAT TCGACGTTAA TTCTGGGCTCT TTAGAACATG CAGAGGCTCG TTAGACAACT
181 CTGGGACAGC TACAACGTC AACAAGCAGCA AGTAGACGAA TTACCTGATC TAAATTATAT
241 ACAGTAGCAG CCACTCTATGG TGCTGATCTAA AAGATAGTAG CAAATGCACG CAAACGAGG
301 TTAGACAGAA TAGAGGGAGA GCAAAACAAA ATGGAAGAAA CAGCCACGACA AGGCCAGCT
361 GCAGAGAGAA AGCAGCAGCA GGGCCGAGCA AATAACACGA TATGTGGAGA GGAAAGAGG
421 CAAATGCTTC ATCAAGCCAAT ACCTCGTAAA ACTTTTAAATG TAGGTTCAA AAATGATG
481 CAAACAGCTT TCGGCGGAGA TAGTAATACC TAGTTTTTAC CATATGACG GGAAGCCACC
541 CAAACAGATG AAACACACTC CTAACACACA CGGCGGGGAC AGCAAGCGAC CATGGCAATG
601 TTAGAAGAGA CCATCAATTG AGGAGCGTCA GAAATGGGATA TGATCTGCTA AGTGCTGACA
661 GAGCCTATTC CACAGGCGCA GATGAGAGGA CCAAGGGGAA CGGCTAGACG AGGAAGCTCT
721 AGTACCCTTC AGGCGGCAAT AGGATGGGATG ACAAATATTC CACCTATCCC AGTAGGACG
781 ATCTATAAAA GATGAGATAA TCTGGGATTT AATAAATTAG TAAAGATGTA TAGGCGTACC
841 AGCACTTTGG AGCATTGGCCA AGAACAAAGA ACAGCTTTTA TAGGCTATGT AGCACGGTTC
901 TATAAAAATC TAGAGGCTGA CAAGCGCTGTA CAGGAGGTA AAAATGGGAT GACAGAACC
961 TTGGTGGCTC AAAATGCGCA CCAAGATGTG AAGACTATTG TAAAGCATTT GGAAGGACAC
1021 GCTACACTAG AGAAGATGAT GACACGATGT CAGAGGAGAG GGGGAGGCGG CATAGAAGA
1081 AGAGTTTGGG CTGAGACGAT GAGCCCAAGTA ACAAAATCAG CTACCATAAT GATGCAGAGA
1141 GGCAATTGTTA GGAACCAAAG AAAGGCTGTG AATGTTTCTA AGCGTCGCAA AGTGGGCAAC
1201 ATAGCTAAAA ATGSCGGCCA CCGTCTAGGA AAGCGCTGTT GGAATGGTGA AAAGGAGGAA
1261 CATGAAATGA AAGATGTGAA TGAAGACAG GCTAATTTTG TAGGGAAAT CTGGCTCTAC
1321 CACAAAGGGA GGCCAGGGGA TTTCTGTCAG AGCAGACAGC AGCAGACGC CCCAGGAAA
1381 GAGAGCTTCA GGTGGGGGGA AGAGAAAACA ACTCCCTCTC AGAAGGCAAG GCGGATGAC
1441 AAGGAACTGT ATCCCTTAGG GCTCCCTAGA TCACTCTTTG GCAACGACCC CTGCTACAA
1501 TAA

FIGURE 1
SEQ ID NO: 2
HIV-1 GAG OPTIMIZED

1 ATGGGTGCAG GAGCTTCGTT GTTGTCTGGA GGGAATTGG ACAAATGGGA AAAAATAAGA
61 CTGGAGCCCA GCCGGAAGAA GAAGTACAAG ATGGAACAGC TTGTGTGAGGC CTCAAGGGAG
121 CTGGAGAGGT TCGCCGCTAA CCCCGGCCTG CTCGAGACGT CGGAGGGTGG TAGGCAAGTA
181 CTAGGAACAC TGAGCCGCAAG GCTCAGAGGC GGGAGTGCAG AATATGGATCT TCTTTACGAC
241 ACCGTTGCTA CGCTGTACTG GTGGACACCA AAGATAGATG TGGAGACAC CAAAGAAGCT
301 TTGGAGAAAA TAGAAGGAGA GCAGAACAAG AGTAAAGAAG AGATCTCAGA GGCTGCAGCC
361 GGCGACTGGAA ATAGCTTCTCA GGGTGCACAG AACTATCAAG TCTGACAAAG TCTGAGGGGC
421 CAAATGGGTC ACCAGAGCCACT TTCTCCTAGG ACAAATTTGT CCTGGGTCAGA GGTTAGTCGAA
481 GAGAAGACAT TTGAGTCCCGA GGTTGATCCCC ATGTTTCTCA GCTGTTGTGA AGGTGCCACC
541 CCCCGAGACC TGGACTTATG CCTGGAATAC GTGGGGGAC ACCAGGCGGC GATGCAATAG
601 TTGGAAAGAg GAATTAAGGC GGGCGCTGGC GAAATGGGAC GCCTGCACCC GGTGCAATGCT
661 GGACCAATAG CTGCTGACAC TTGGCGAGAG GCACTGAGGA GCACCTGTCCG TGGGACTACT
721 TCTACACTGC AAGAGCACTG TGGGTGGATG AGAGATAAGG CCTCCTATTC CTTGGCGAG
781 ATCTACCAAGA GTGGAATTAT CTTGGGCGAG GAAAGATGCT TCTGGAGATG AAGGCTACCC
841 AGATATCTGG ACTGAGCCGG GGAGCGAGCG GAACCCCTTTG GCGACTGAGT GTTGAATTTT
901 TATATACCCGT TGGAGCGCAG GGAGCGCGAC AAGAGAGTGA AGAAGCTGGT GAGGAAACCA
961 CTGTGGGTTC AGAACGCGAAA CCCGACTGCG AAGACCATTC TGAAGGCGC CTGAGCCAGCA
1021 GCAAACCCCTAG AGAATAGTATG GACCGCTGCC GGGGCAGTGG GGGCCCTGG TCACAGGCT
1081 CGTGTGTTAG CAGAGGGCAT GTCAAGGCTA ACAAAATGCAG CCACAAATTAT GATGCAACGA
1141 GCCAATTCCC GTAATCAAAG GAAAACAGTGG AAGTGCTTCA ACTGTGGGAA AGTAGGCCCT
1201 AATCGCCAAGA ACTGCTGACG TCTGGCAGAA AAGGGTTGTGG GGGTAAGGTGG TGAAGAGGCT
1261 CATCAGATGA AGGACTGTAAG CAGAAGGCGAG CCAGATTTCC TTGGAAATAT CTGGCGCGCC
1321 TATAAAGGAC GGCAGCGCGA TTTCTCTTCA TCGGTCTGGG AGCGGCAAGG CCCGGTGAGG
1381 GAGAGCTTTG CTTCCGGAGA GGAAAGGACC ACCGGAGTCC AGAACAGGA ACCAACTGAC
1441 AAGGAGCTGT ACCGCGCTGGC GCTCTCTTGA AGTTGTTTGG GAAATGACCG AAGGAGGCCG
1501 TAA

FIGURE 2
Step i) selecting a nucleotide sequence “b”, which encodes an amino acid sequence “s”, for optimization.

\[
\begin{array}{c}
\text{ATG GGT GCG AGA GCG} \\
\text{M G A R A} \\
\text{“b” (SEQ ID NO: 19)} \\
\text{“s” (SEQ ID NO: 20)}
\end{array}
\]

Step ii) selecting a group of one or more genes “tn” from the organism or tissue of interest use as templates to optimize the nucleotide sequence.

```
Gene 1
Gene 2
Gene 3
Gene 4
“tn”
```

Step iii) ranking the group of one or more genes in “tn” according to their expression level in the organism or tissue of interest.

```
Expression level
Gene 4
Gene 2
Gene 1
Gene 3
```

Step iv) selecting an amino acid word “w0” of length “L” from the amino acid sequence “s”.

\[
\begin{array}{c}
\text{ATG GGT GCG AGA GCG} \\
\text{M G A R A} \\
\text{“s”} \\
\text{“w0”}
\end{array}
\]

Step v) sequentially examining the ranked genes from group “t” for a nucleotide sequence that [can potentially encode] the amino acid word “w0” [without regard to the translation frame].

```
Gene 4
Gene 2
Gene 1
Gene 3
```

\[
\begin{array}{c}
\text{TAT GGG CGC CTA} \\
\text{M G A} \\
\text{“n0” (SEQ ID NO: 21)} \\
\text{“n0” (SEQ ID NO: 22)}
\end{array}
\]

**FIGURE 3A**
Step vi) eliminating the C-terminal amino acid from the polypeptide sequence "w0" and adding the next N-terminal amino acid in the corresponding amino acid sequence in "s" to generate amino acid sequence "w1".

\[
\begin{array}{c}
\text{ATG GGT GCG AGA GCG} \\
\text{M G A R A} \\
\text{"w0"}
\end{array}
\]

\[
\begin{array}{c}
\text{ATG GGT GCG AGA GCG} \\
\text{M G A R A} \\
\text{"w1"}
\end{array}
\]

(SEQ ID NO: 19)

Step vii) eliminating the three nucleotides at the 5' end of the nucleotide sequence "n0" to generate a new nucleotide sequence "n01".

\[
\begin{array}{c}
\text{AT GGG CGC C} \rightarrow \text{"n0"} \\
\text{AT GGG CGC C} \rightarrow \text{"n01"}
\end{array}
\]

(SEQ ID NO: 23)

(SEQ ID NO: 24)

Step viii) replacing the nucleotide sequence of "n0" with the nucleotide sequence of "n1", and,

Step ix) recursively repeating the steps (v) to (viii) until the end of the amino acid sequence "s", thereby generating the optimized nucleotide sequence "o".

\[
\begin{array}{c}
\text{TTG GCG CCA GGT} \\
\text{G A R "n1"}
\end{array}
\]

(SEQ ID NO: 25)

(SEQ ID NO: 26)

(SEQ ID NO: 27)

\[
\begin{array}{c}
\text{M G A R A} \\
\text{ATG GGT GCG AGA GCG} \\
\text{ATG GGC GCC} \\
\text{GGC GCC AGG} \\
\text{ATG GGC GCC AGG}
\end{array}
\]

(SEQ ID NO: 19)

(SEQ ID NO: 28)

(SEQ ID NO: 29)

(SEQ ID NO: 30)

Repeating recursively until the end of the end of the protein sequence "s" is reached, thereby generating the optimized nucleotide sequence "o".

FIGURE 3B