UpGene: Application of a Web-Based DNA Codon Optimization Algorithm

Wentao Gao,[†] Alexis Rzewski,[‡] Huijie Sun,[‡] Paul D. Robbins,[†] and Andrea Gambotto*,[†,‡]

Department of Molecular Genetics and Biochemistry and Department of Surgery, University of Pittsburgh, Pittsburgh, Pennsylvania

Although DNA codon optimization is a standard molecular biology strategy to overcome poor gene expression, to date no public software exists to facilitate this process. Among the uses of codon optimization, human immunodeficiency virus (HIV) vaccine development represents one of the most difficult challenges. A key obstacle to an effective DNA-based vaccine is the low-level expression of HIV genes in mammalian cells, which is due primarily to the instability of HIV mRNAs resulting from AU-rich elements and rare codon usage. In this report we describe the development of a DNA optimization algorithm integrated with a PCR primer design program to redesign specific coding sequences for maximal gene expression. Using this algorithm combination, together with PCR-based gene assembly, we have successfully optimized gene sequences for simian immunodeficiency virus (SIV) strain mac239 structural antigenic proteins *gag* and *env*, resulting in high-level gene expression in eukaryotic cells. Our findings demonstrate that our user-friendly algorithm is a valuable tool for DNA-based HIV vaccine development. Moreover, it can be used to optimize any other genes of interest and is freely available online at http://www.vectorcore.pitt.edu/upgene.html.

Introduction

The expression of a gene in mammalian cells, in general, depends on a variety of factors, including gene copy number, transcription control elements, mRNA stability, and translational efficiency. Human immunodeficiency virus (HIV) gene expression is tightly regulated at multiple levels. At the transcriptional level, gene expression is controlled mainly by the viral transactivator *tat*. The binding of *tat* to the RNA element, TAR, which is located at the 5′ end of the HIV transcript, stimulates transcription from the HIV long terminal repeat. At a posttranscriptional step, the binding of HIV *rev* to an RNA element, the *rev*-responsive element (RRE), which is found in all unspliced and singly spliced HIV mRNAs, results in the nuclear export of the latter and the translation of *gag-pol*, *env*, *vif*, *vpr*, and *vpu* (*1*). Although not all HIV instability elements contain the AUUUA pentanucleotide, studies have indicated that the high AU content (AU-rich elements, ARE) of the viral RNA is one of the major factors affecting the instability of HIV mRNAs (*2*−*4*). In addition, the presence of AUUUA sequences in human genes such as *c-myc* or certain cytokines is associated with high RNA instability (*5*−*6*). Several cis-acting repressive sequences, or instability sequences (INSs), have been identified within HIV. These include sequences in the *gag-pol* and *env* genes, and the RRE (*4*,*7*). When the AU-rich INSs within *gag-pol* were mutated, increased steady-state mRNA levels correlated positively with protein production.

At the translational level, the unusual HIV codon bias, which is markedly different from the one used by highly expressed human genes (*8*−*9*), may also affect the expression of HIV genes because of the abundance of various tRNA isoacceptor species (*10*−*11*). The pausing of the ribosomes at rare codons may lead to enhanced RNA turnover (*12*). Furthermore, studies have shown that changing codon usage has a significant effect on translation (*13*−*14*). We have developed an algorithm that computerizes codon optimization, ARE exclusion, and Kozak consensus sequence introduction within target genes without interrupting the primary amino acid sequence. The program also integrates with a PCR primer design program based on the output-optimized DNA sequence.

To demonstrate the utility of the optimization algorithm, we used the simian immunodeficiency virus (SIV) mac239 strain as a model of HIV gene expression; this strain has been used extensively in nonhuman primates, and infection with it results in an AIDS-like illness. Using the algorithm together with a PCR-based gene assembly technique, we generated a series of optimized DNA sequences encoding SIVmac239 *gag* p17 and *gag* p45 (which represent the first and the second half of SIV *gag*, respectively) and SIVmac239 *env1/3* (which represents the first third segment of SIV *env*). After infection of mammalian cells with adenoviral vectors bearing the optimized genes, the expression levels of the SIV mac239 structural proteins *gag* and *env* were substantially increased.

* Department of Molecular Genetics and Biochemistry and Department of Surgery, University of Pittsburgh School of Medicine, Center for Biotechnology and Bioengineering, 300 Technology Drive, Suite 211, Pittsburgh, PA 15219. Phone: (412) 383-7684. Fax: (412) 383-9760. E-mail: agamb@imap.pitt.edu.

[†] Department of Molecular Genetics and Biochemistry.

[‡] Department of Surgery.

10.1021/bp0300467 CCC: $27.50
Published on Web 12/24/2003

Materials and Methods

Cell Lines. HEK293 cells (ATCC) and CRE8 cells were maintained in DMEM with 10% FBS, 2 mM L-glutamine, 100 IU /mL of penicillin, and 100 μg/mL of streptomycin as previously described.

Codon Optimization Algorithm. A software implementation for codon optimization was developed using the IBM VisualAge for SmallTalk software (IBM, Armonk, NY). Using object-oriented analysis and design methodologies, areas of responsibilities were designed that identified and described a sequence, an enzyme, a codon, the task of optimizing a sequence, the task of formatting both an optimized sequence and primers, the interface to external file system, and the different graphical user interface components used to display data. For each area of responsibility, a corresponding SmallTalk class was implemented, and its functionality was tested. Enzyme and codon probability distribution data was stored in files.

Algorithm for Preparation of the Altered Sequence. The content of the original sequence was first validated by verifying that it contained only characters in the set {a c g t} and that its number of characters was a multiple of three so that the sequence could be segmented into triplets, i.e., subsequences of three characters each. Then, occurrences of "fixed", or "unalterable", subsequences were marked on the original sequence; "fixed" sequences are subsequences found in the original sequence that remain unaltered. Having then subdivided the original sequence into triplets, for each triplet that does not contain any character that was marked as "fixed", an amino acid was found from a table that correlates amino acids with triplets. A random number between 0 and 100 was generated and given a specified codon source, a new triplet was found from another table that correlates amino acids, triplets, and the probability of occurrence of the triplet for that codon source. The new triplet replaced the one currently being altered. The process was continued until the last triplet of the original sequence was replaced. The altered sequence was then validated by verifying that it did not contain any of the selected restriction enzymes or any of the user-specified "excluded" sequences, i.e., subsequences that were not to be found in the altered sequence. If the validation failed, the process of random alteration was repeated until a valid altered sequence was found or until the number of attempts reached the maximum level allowed. The subsequences of the selected restriction enzymes to be added on the 5′ end of the altered sequence were chained together and verified that the resulting sequence did not contain any "excluded" sequence. Then, it was added to the 5′ end of the altered sequence. The above process was repeated for the subsequences of the selected restriction enzymes to be appended on the 3′ end of the altered sequence.

Algorithm for Preparation of the Primers. If the altered sequence with concatenated enzymes was even, then the primer divisor 1 was set to 21. If it was odd, the primer divisor 1 was set to 20. Then, an offset map was built for key values of 38, 40, 42, 44, and 46. For each of the keys, a value was found for which the sequence length minus the primer divisor 1 plus an offset was divisible by the key with no remainder. The offset was incremented until a value was found. Once the map was built, the key with the lowest value was found and assigned to primer divisor 2. A stuffer of randomly generated characters was created and split into two subsequences with one prefixed to the head of the altered sequence and the other appended to the tail. The sequence was validated for exclusion of undesired subsequences, inverted, complemented, and formatted into rows with length of primer divisor 2, except for the first row, with length of primer divisor 1.

Codon-Optimized Gene Synthesis. Oligonucleotide primers were synthesized using standard phosphoramidite chemistry. The PCR protocol for gene synthesis consists of two steps: gene assembly and gene amplification. Because single-stranded ends of complementary oligonucleotide DNA fragments are filled in by PCR, cycling with DNA polymerase results in increasingly larger DNA fragments until the full-length gene is obtained (Figure 1). Briefly, in the gene assembly step, equal volumes from each oligonucleotide solution (100 pM) were combined, and the mixture was diluted 40-fold in 50 μL of PCR mix containing 10 mM Tris-HCl pH 9.0, 2.2 mM $MgCl_2$, 50 mM KCl, 0.2 mM each dNTP, 0.1% Triton X-100, and 1 u of Taq polymerase/0.02 u of Pfu polymerase. The combination of enzymes was used to maximize efficiency and proofreading (*15*). The PCR program consisted of 55 cycles at 94 °C for 30 s, 52 °C for 30 s, and 72 °C for 30 s. In the amplification step, the gene assembly reaction mixture was diluted 40-fold in 100 μL of PCR mix containing 10 mM Tris-HCl pH 9.0, 2.2 mM $MgCl_2$, 50 mM KCl, 0.2 mM each dNTP, 0.1% Triton X-100, and 5 u Taq polymerase/0.1 u of Pfu polymerase and combined with two outside primers at a concentration of 1 μM. The outside primers are the same as the two oligonucleotides representing the 5′ ends of the plus and minus strands. The PCR program consisted of 25 cycles at 94°C for 30 s, 50°C for 30 s, and 72°C for 60 s. The PCR products for the optimized SIVmac239 structural genes *gag* p17, *gag* p45, or *env*1/3 were separated on a 1% agarose gel. Each appropriate band was excised and purified using a Gel Extraction Kit (Qiagen, Valencia, CA) and then separately cloned into the pcDNA3.1 TOPO T/A cloning plasmid (Invitrogen, Carlsbad, CA). Each correct sequence was confirmed by sequencing before further cloning. The optimized sequences of SIVmac239 *gag* p17 (153 aa), *gag* p45 (367 aa), and *env1/3* (343 aa) contained 72, 189, and 268 codon replacements, respectively. Corresponding wild-type genes were synthesized by standard PCR using SIVmac239 cDNA as the template and cloned as described above.

Construction of Recombinant Adenoviruses. E1/E3-deleted adenoviral vectors expressing the optimized *gag* p17, *gag* p45, or *env1/3* genes of SIVmac239 (Ad*gag* p17, Ad*gag* p45, and Ad*env1/3*) were constructed using Cre-lox recombination. Reagents were generously provided by Dr. S. Hardy (Somatix, Alameda, CA). Briefly, a Sal I−Not I fragment containing the optimized *gag* p17, *gag* p45, or *env1/3* from the corresponding pcDNA3.1 plasmids was inserted into a modified version of the shuttle vector pAdlox (GenBank U62024). E1/E3-substituted recombinant adenovirus was generated by cotransfection of the Sfi I-digested pAdlox/*gag* p17, *gag* p45, or *env1/3* with Ψ5 helper virus DNA into the adenoviral packaging cell line CRE8. Adenoviruses were propagated in CRE8 cells, purified by cesium chloride density gradient centrifugation and dialysis and stored at −70°C.

Protein Expression. To test the codon-optimized SIVmac239 antigen expression, HEK293 cells were transfected with either the expression plasmid containing the optimized *gag* p17(pcDNA3.1*gag* p17 optimized) or a control plasmid containing the wild-type gene using the DNA−calcium phosphate coprecipitation method (5 μg of plasmid DNA per 60 mm plate). In addition, HEK293 cells were infected with Ad*gag* p45, Ad*env1/3*, or control

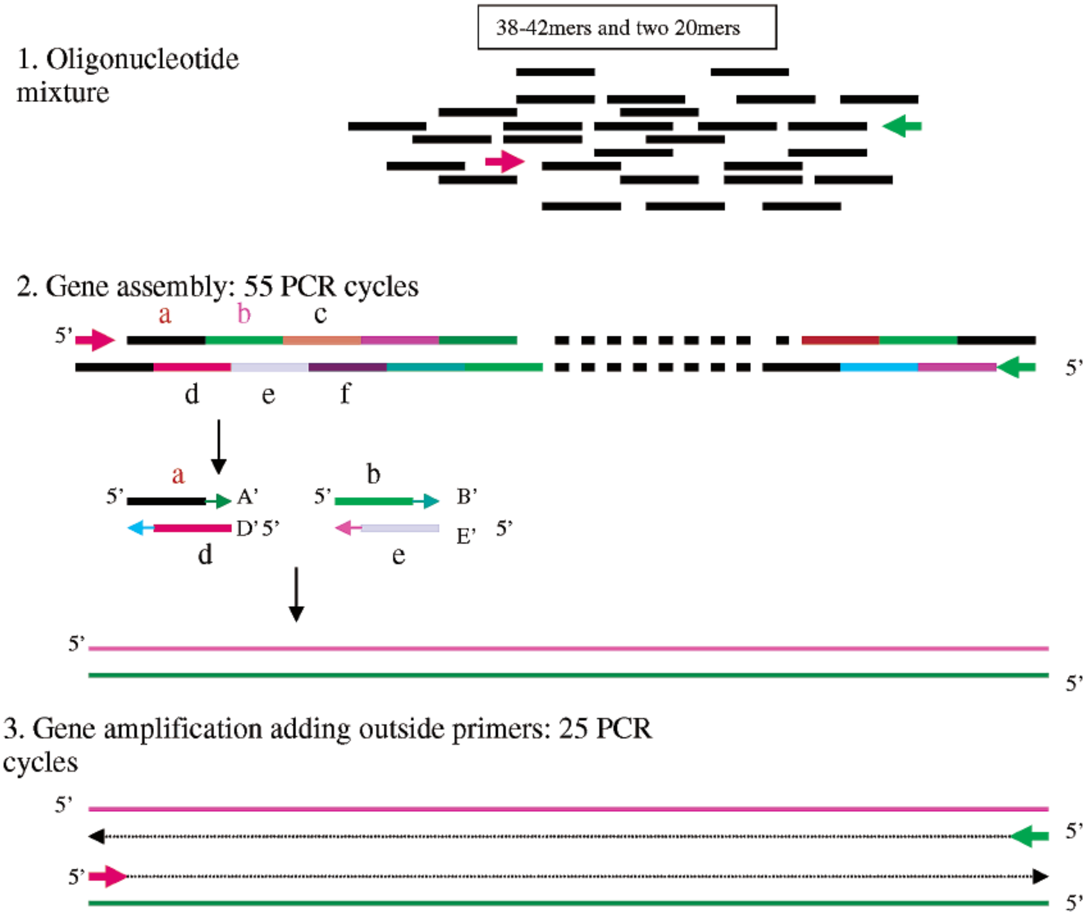

38-42mers and two 20mers
1. Oligonucleotide mixture
2. Gene assembly: 55 PCR cycles
a
b
c
d
e
f
5'
A'
B'
D' 5'
E'
3. Gene amplification adding outside primers: 25 PCR cycles

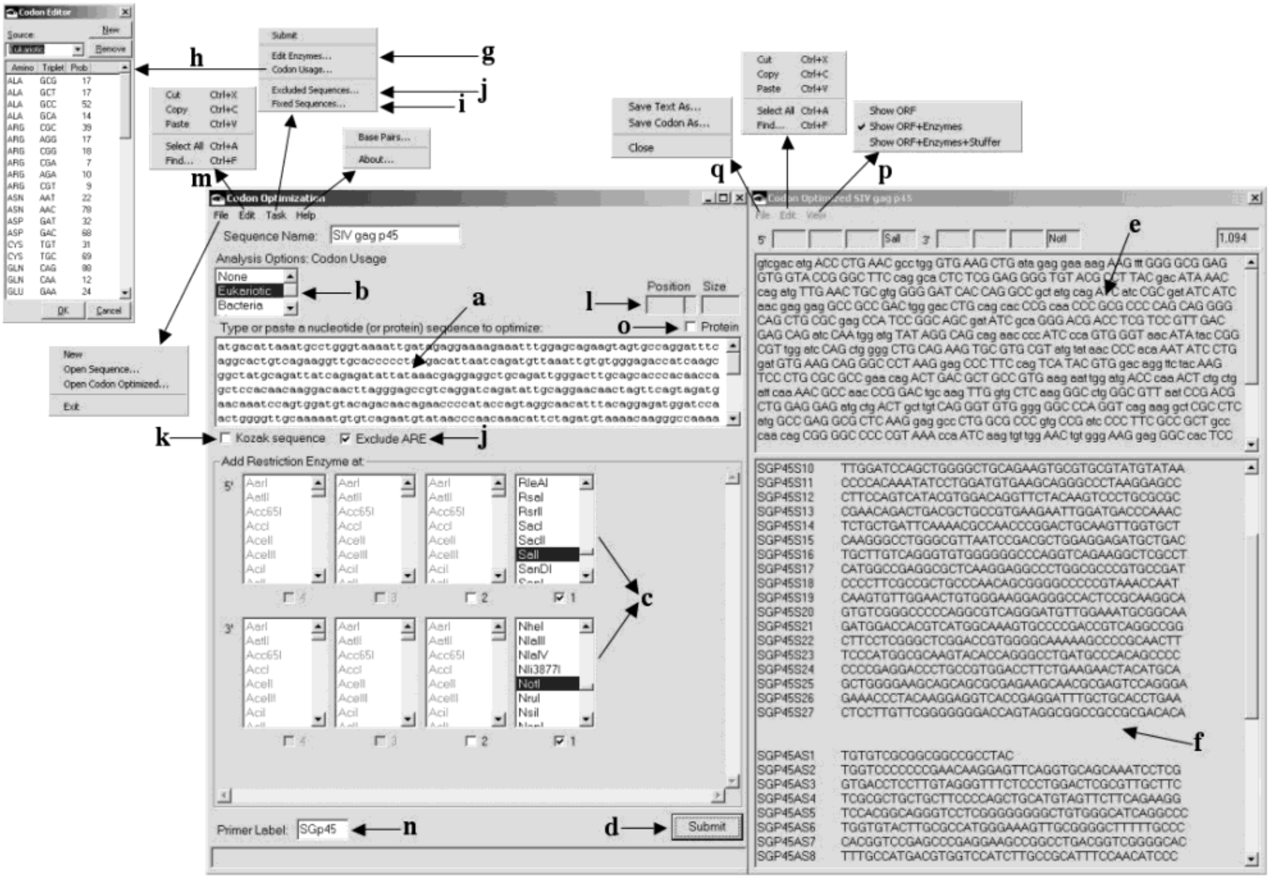
Figure 1. Schematic representations of overlapping oligonucleotide PCR-based gene synthesis. (1) Mixture of oligonucleotides includes 38−42mers and 20mers. (2) In the gene assembly step, oligonucleotide **a** extends using d as a template, forms A′, and then stops. Oligonucleotide **b** forms B′, **d** forms D′, and **e** forms E′. During the next cycle, A′ extends using E′ as a template, continuing until synthesis of the entire length of the gene is complete. (3) In the gene amplification step, outside primers are added. Standard PCR is performed to amplify the full-length gene.

adenovirus Ψ5 containing no gene insert at MOI = 10. The cells were harvested 72 h later. Western blot analysis was used to detect protein expression. The cell pellets were suspended in lysis buffer (100 μL per 1 × 10^6 cells, 15 mM HEPES pH7.5, 1 mM EGTA, 1.5 mM $MgCl_2$, 1 mM DTT, 10 mM KCl) with proteinase inhibitor cocktail (Novagen, Madison, WI) and lysed during three freeze/thaw cycles in a dry ice/ethanol bath. The cell lysates (20 μL) were separated on SDS polyacrylamide gels (10% or 12%) and transferred to a PVDF membrane (BioRad, Hercules, CA). The SIV *gag* p17, *gag* p45, and *env*1/3 proteins were detected by immunoprobing with primary monoclonal antibodies KK59, 55-2F12, and KK8, respectively, followed by HRP-conjugated anti-mouse IgG and detection using an ECL kit (Amersham Pharmacia Biotech, Piscataway, NJ). All monoclonal antibodies and the recombinant SIV gag p55 protein were obtained from the NIH AIDS Research & Reference Reagent Program. To confirm uniform sample loading, a rabbit polyclonal antibody to human GAPDH was used to detect GAPDH in cell lysates (Abcam, Ltd., Cambridge, UK).

Results

Codon Optimization Algorithm. We have developed a GUI software application implementing an optimization algorithm that breaks a given DNA open reading frame (ORF) into triplets and replaces them with synonymous ones drawn from a probability distribution based on the optimal frequency of codon usage in the human genome. Figure 2 shows the algorithm work area, including the menu at the top of the screen and the image window containing a variety of tools and palettes for editing and adding elements. The user specifies a sequence in the textbox of the Codon Optimization window (2a), chooses a codon usage source such as “Eukaryotic” in a listbox (2b), selects one or more restriction enzymes to attach to either the 5′ or 3′ ends of the ORF (2c), and presses the “Submit” button (2d). The Optimized Codon window displays the results as an “optimized” sequence (2e) and a primer sequence (2f). Triplets that have been optimized are displayed in uppercase letters.

The user selects restriction enzymes by clicking on the checkbox below a listbox (2c). Selecting an enzyme also enables the left-adjacent checkbox, thus allowing the enzyme selection process to flow from right to left, i.e., from position 1 to position 4. Displayed restriction enzymes can be added and removed using an Enzyme Editor that can be opened as a dialogue (2g). The application contains three predefined codon distributions for Eukaryotic, Bacteria, and Yeast codon sources, as well as an option to define custom sources using the Codon Editor (2h). This editor allows the user to name a codon in a textfield and specify its codon probability distribution values in a table. The entered data are automatically saved to a file that is read when the application is launched; the added codon source appears in the Codon Optimization window listbox.



Source: Eukariotic

Amino	Triplet	Prob
ALA	GCG	17
ALA	GCT	17
ALA	GCC	52
ALA	GCA	14
ARG	CGC	39
ARG	AGG	17
ARG	CGG	18
ARG	CGA	7
ARG	AGA	10
ARG	CGT	9
ASN	AAT	22
ASN	AAC	78
ASP	GAT	32
ASP	GAC	68
CYS	TGT	31
CYS	TGC	69
GLN	CAG	88
GLN	CAA	12
GLU	GAA	24

Sequence Name: SIV gag p45

Analysis Options: Codon Usage: None / Eukariotic / Bacteria

Type or paste a nucleotide (or protein) sequence to optimize: □ Protein

```
atgacattaaatgcctgggtaaaattgatagaggaaaagaaatttggagcagaagtagtgccaggatttc
aggcactgtcagaaggttgcaccccctaagacattaatcagatgttaaattgtgtgggagaccatcaagc
ggctatgcagattatcagagatattataaacgaggaggctgcagattgggacttgcagcacccacaacca
gctccacaacaaggacaacttagggagccgtcaggatcagatattgcaggaacaactagttcagtagatg
aacaaatccagtggatgtacagacaacagaaccccataccagtaggcaacatttacaggagatggatcca
actggggttgcaaaaatgtgtcagaatgtataacccaacaaacattctagatgtaaaacaagggccaaaa
```

□ Kozak sequence ☑ Exclude ARE

Primer Label: SGp45

5' SalI 3' NotI 1,094

```
gtcgac atg ACC CTG AAC gcc tgg GTG AAG CTG ata gag gaa aag AAG ttt GGG GCG GAG
GTG GTA CCG GGC TTC cag gca CTC TCG GAG GGG TGT ACG CCT TAC gac ATA AAC
cag atg TTG AAC TGC gtg GGG GAT CAC CAG GCC gct atg cag ATC atc CGC gat ATC ATC
aac gag gag GCC GCC GAC tgg gac CTG cag cac CCG caa CCC GCG CCC CAG CAG GGG
CAG CTG CGC gag CCA TCC GGC AGC gat ATC gca GGG ACG ACC TCG TCC GTT GAC
GAG CAG atc CAA tgg atg TAT AGG CAG cag aac ccc ATC cca GTG GGT aac ATA tac CGG
CGT tgg atc CAG ctg ggg CTG CAG AAG TGC GTG CGT atg tat aac CCC aca AAT ATC CTG
gat GTG AAG CAG GGC CCT AAG gag CCC TTC cag TCA TAC GTG gac agg ttc tac AAG
TCC CTG CGC GCC gaa cag ACT GAC GCT GCC GTG aag aat tgg atg ACC caa ACT ctg ctg
att caa AAC GCC aac CCG GAC tgc aag TTG gtg CTC aag GGC ctg GGC GTT aat CCG ACG
CTG GAG GAG atg ctg ACT gct tgt CAG GGT GTG ggg GGC CCA GGT cag aag gct CGC CTC
atg GCC GAG GCG CTC AAG gag gcc CTG GCG CCC gtg CCG atc CCC TTC GCC GCT gcc
caa cag CGG GGC CCC CGT AAA cca ATC aag tgt tgg AAC tgt ggg AAG gag GGC cac TCC
```

```
SGP45S10    TTGGATCCAGCTGGGGCTGCAGAAGTGCGTGCGTATGTATAA
SGP45S11    CCCCACAAATATCCTGGATGTGAAGCAGGGCCCTAAGGAGCC
SGP45S12    CTTCCAGTCATACGTGGACAGGTTCTACAAGTCCCTGCGCGC
SGP45S13    CGAACAGACTGACGCTGCCGTGAAGAATTGGATGACCCAAAC
SGP45S14    TCTGCTGATTCAAAACGCCAACCCGGACTGCAAGTTGGTGCT
SGP45S15    CAAGGGCCTGGGCGTTAATCCGACGCTGGAGGAGATGCTGAC
SGP45S16    TGCTTGTCAGGGTGTGGGGGGCCCAGGTCAGAAGGCTCGCCT
SGP45S17    CATGGCCGAGGCGCTCAAGGAGGCCCTGGCGCCCGTGCCGAT
SGP45S18    CCCCTTCGCCGCTGCCCAACAGCGGGGCCCCCGTAAACCAAT
SGP45S19    CAAGTGTTGGAACTGTGGGAAGGAGGGCCACTCCGCAAGGCA
SGP45S20    GTGTCGGGCCCCCAGGCGTCAGGGATGTTGGAAATGCGGCAA
SGP45S21    GATGGACCACGTCATGGCAAAGTGCCCCGACCGTCAGGCCGG
SGP45S22    CTTCCTCGGGCTCGGACCGTGGGGCAAAAAGCCCCGCAACTT
SGP45S23    TCCCATGGCGCAAGTACACCAGGGCCTGATGCCCACAGCCCC
SGP45S24    CCCCGAGGACCCTGCCGTGGACCTTCTGAAGAACTACATGCA
SGP45S25    GCTGGGGAAGCAGCAGCGCGAGAAGCAACGCGAGTCCAGGGA
SGP45S26    GAAACCCTACAAGGAGGTCACCGAGGATTTGCTGCACCTGAA
SGP45S27    CTCCTTGTTCGGGGGGGACCAGTAGGCGGCCGCCGCGACACA

SGP45AS1    TGTGTCGCGGCGGCCGCCTAC
SGP45AS2    TGGTCCCCCCCGAACAAGGAGTTCAGGTGCAGCAAATCCTCG
SGP45AS3    GTGACCTCCTTGTAGGGTTTCTCCCTGGACTCGCGTTGCTTC
SGP45AS4    TCGCGCTGCTGCTTCCCCAGCTGCATGTAGTTCTTCAGAAGG
SGP45AS5    TCCACGGCAGGGTCCTCGGGGGGGGCTGTGGGCATCAGGCCC
SGP45AS6    TGGTGTACTTGCGCCATGGGAAAGTTGCGGGGCTTTTTGCCC
SGP45AS7    CACGGTCCGAGCCCGAGGAAGCCGGCCTGACGGTCGGGGCAC
SGP45AS8    TTTGCCATGACGTGGTCCATCTTGCCGCATTTCCAACATCCC
```

Figure 2. Screens of the Codon Optimization Algorithm application. In the left window, the wild-type sequence is specified and the options are selected. In the right window, the output is represented by the optimized sequence and the list of overlapping oligonucleotides. Details on the application are given in Results.

The user can also specify sequences within the original sequence that should remain fixed, i.e., not replaceable by the user (2i). When such sequences are found during the optimization, the application guarantees that they will remain unchanged in the resulting optimized sequence. Similarly, the user can specify sequences that should be excluded (Excluded Sequences dialogue, 2j). When such "excluded sequences" are found within the optimized sequences, the application must repeat the optimization process until are they are removed. If after 1000 attempts the removal of excluded sequences is unsuccessful, no resulting sequence is produced and an informative message is displayed to the user.

The user can also choose to introduce a Kozak Sequence by selecting the appropriate checkbox (2k). Highlighting text in the sequence textbox displays the starting and ending positions within the sequence, the frame, and the length in a series of text fields above the sequence textbox (2l). Simply positioning the insertion cursor within the text displays its frame position. Selecting the "Find menu item" in the Edit menu opens a Search dialogue in which text can be specified and used to search the sequence; when found, the text in the sequence textbox is highlighted with the Search dialogue still open (2m). Repeating the Search locates the next occurrence in the sequence. The Primer Prefix textfield (2n) is used to specify a prefix for each primer. The Protein checkbox (2o) informs the application that the sequence specified in the textbox represents an amino acid sequence.

In the Optimized Codon result window, the user can select from three possible viewing options: (1) ORF only, (2) ORF + restriction enzymes, or (3) ORF + enzymes + stuffer (2p). The stuffer is a sequence of randomly generated nucleotides whose length allows the primer to be formatted in a certain size and shape. Care has been taken to ensure that the stuffer does not contain the ATG start codon or any stop codons.

For storage, the resulting optimized codon and primer text can be saved to a text file (2q). Also, the optimized codon, along with all its attached values (such as list of restriction enzymes, codon source, and Kozak sequence usage) can be saved to a proprietary-formatted binary data file. This proprietary file can be used to launch the application when double-clicking it in a Windows operating system. A Web-based version of the software is available at http://www.vectorcore.pitt.edu/upgene.html.

Expression of Codon-Optimized SIV/HIV Genes. Using the algorithm and PCR-based gene assembly technique, we generated a series of optimized DNA sequences encoding SIVmac239 *gag* p17, SIVmac239 *gag* p45, and SIVmac239 *env1/3* (Eukaryotic codon usage). The codon-optimized genes were tested for expression in HEK293 cells. Separately, both expression plasmids and recombinant adenovirus containing either the codon-optimized or the wild-type form were introduced to these mammalian cells; levels of gene expression were compared (Figure 3). Western blot analysis revealed high levels of the optimized antigens and undetectable expression of the wild-type genes. These results demonstrated that the appropriate modification of SIV sequences, including codon optimization, exclusion of ARE, and introduction of a Kozak consensus sequence, can significantly increase the expression of SIV/HIV genes.

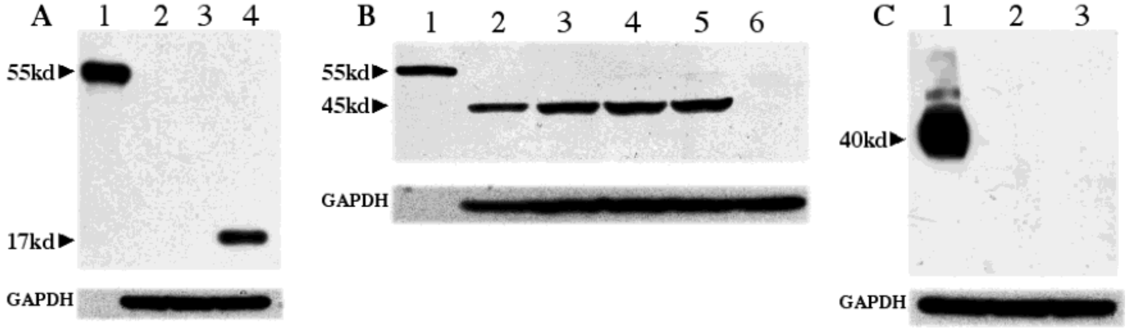


Figure 3. Western blot analysis of optimized SIVmac239 *gag* and *env* genes. (A) Lane 1, recombinant SIVmac239 *gag* p55 protein control; lane 2, HEK293 cell lysate control; lane 3, lysate of HEK293 cells transfected with pcDNA3.1*gag* p17 wild type; lane 4, lysate of HEK293 cells transfected with pcDNA3.1*gag* p17 optimized. (B) Lane 1, recombinant SIVmac239 *gag* p55 protein; lanes 2−5, increasing amounts of lysed HEK293 cells infected with recombinant adenovirus Ad-SIV*gag* p45 optimized; lane 6, HEK293 cell lysate control. (C) Lane 1, lysate of HEK293 cells infected with recombinant adenovirus Ad-SIV*env1/3* optimized; lane 2, lysate of HEK293 cells infected with adenovirus ψ5 control containing no gene insert; lane 3, HEK293 cell lysate. To confirm consistency of sample loading, GAPDH was detected in all corresponding cell lysates.

Discussion

The relentless expansion of the HIV pandemic highlights the desperate need for a vaccine. One hurdle in the development of an effective anti-HIV DNA vaccine is the low expression of HIV genes in mammalian cells, which is due primarily to the instability of HIV mRNAs resulting from ARE and rare codon usage. In this report we described our development of a codon optimization algorithm to redesign SIV/HIV coding sequences in order to maximize their expression in eukaryotic cells. By combining this algorithm with PCR-based gene synthesis, we have generated a series of optimized gene sequences for SIV mac239 antigens; transfection of eukaryotic cells with adenoviral shuttle plasmids containing these sequences yielded a significant increase in SIV mac239 antigen expression without the requirement for Rev/RRE.

Although DNA codon optimization is a standard molecular biology strategy to overcome poor gene expression, to date no public software exists to facilitate this process. In this report, we describe the development of the first publicly available software for DNA codon optimization. This user-friendly application provides a multilevel optimization of given gene sequences that overcomes the tedious manual processes used previously, and our refined PCR-based technique to synthesize modified genes is rapid and relatively inexpensive. Using these combined methods, an optimized gene can be generated in less than a week.

The extraordinary degree of genetic diversity among HIV isolates immeasurably complicates the development of an HIV vaccine. It is likely that a successful vaccine against HIV will induce an immune response of both neutralizing antibodies and cytotoxic T cells. Published studies elucidating the replication of HIV and the immunopathogenesis of AIDS suggest that HIV is not amenable to control by immune responses elicited through traditional vaccine modalities. In fact, experiments in nonhuman primates and early-phase human studies bear out this supposition, providing convincing evidence that live attenuated virus vaccines, inactivated virus vaccines, and recombinant protein vaccines are all likely to be ineffective in preventing HIV infection and AIDS. Recognition of the limitations of these traditional immunization strategies for preventing HIV infection has inspired researchers to explore a plethora of novel vaccine designs. The most promising approaches involve the use of DNA-based vaccines, which include plasmid DNA immunogens and live recombinant viral vectors; the best of these appear to be the recombinant adenoviral vectors that have demonstrated impressive immunogenicity in both murine and nonhuman primate studies (*16*).

Because HIV isolates that infect humans and cause AIDS include a genetically diverse population of viruses, a vaccine will be required to induce broad-spectrum immunity covering most, if not all, HIV subtypes. It is likely that a combination of vaccines (rather than a single vaccine) will achieve success. Therefore, the rapid generation of high-level expression genes coding for HIV antigens represents an important step in the process of effective vaccine development for a large human population. The algorithm and technique we have devised and made publicly available may prove to be valuable for DNA-based HIV vaccine development.

Acknowledgment

We thank Mrs. Kaori Okada for excellent technical assistance. This work was supported by NIH grants AE41870 (A.G.) and AI43664 (P.D.R.) and by the NIH AIDS Research & Reference Reagent Program.

References and Notes

(1) Pollard, V.; Malim, M. The HIV-1 Rev protein. *Annu. Rev. Microbiol.* **1998**, *52*, 491−532.

(2) Afonina, E.; Neumann, M.; Pavlakis, G. Preferential binding of poly(A)-binding protein I to an inhibitory RNA element in the HIV-1 *gag* mRNA. *J. Biol. Chem*. **1997**, *272*, 2307−2311.

(3) Maldarelli, F.; Martin, M. A.; Strebel, K. Identification of posttranscriptionally active inhibitory sequences in human immunodeficiency virus type 1 RNA: novel level of gene regulation. *J. Virol.* **1991**, *65*, 5732−5743.

(4) Nasioulas, G.; Zolotukhin, A. S.; Tabernero, C.; Solomin, L.; Cunningham, C. P.; Pavlakis, G. N.; Felber, B. K. Elements distinct from human immunodeficiency virus type 1 splice sites are responsible for the Rev dependence of env mRNA. *J. Virol.* **1994**, *68*, 2986−2993.

(5) Maurer, F.; Tierney, M.; Medcalff, R. An AU-rich sequence in the 3′ UTR of PAI-2 mRNA promotes PAI-2 mRNA decay and provides a binding site for nuclear HuR. *Nucleic Acids Res.* **1999**, *27*, 1664−1673.

(6) Rabbits, P.; Forster, A.; Stinson, M.; Rabbits, T. Truncation of exon 1 from the c-myc gene results in prolongued c-myc mRNA stability. *EMBO J*. **1985**, *4*, 3727−3733.

(7) Brighty, D.; Rosenberg, M. A cis-acting repressive sequence that overlaps the Rev responsive element of HIV-1 regulates nuclear retention of env mRNAs independently of known splice signals. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 8314−8318.

(8) Kypr, J.; Mrazek, J. Unusual codon usage of HIV. *Nature* **1987**, *327*, 20.

(9) Kypr, J.; Mrazek, J.; Reich, J. Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1 and 2. *Biochim. Biophys. Acta* **1989**, *1009*, 280−282.

(10) Berg, O.; Kurland, C. Growth-rate optimised tRNA abundance and codon usage. *J. Mol. Biol.* **1997**, *270*, 1705−1711.

(11) Varenne, S.; Buc, J.; Lloubes, R..; Lazdunski, C. Translation is a nonuniform process: effect of transfer RNA availability on the rate of elongation of nascent polypeptide chains. *J. Mol Biol.* **1984**, *180*, 549−576.

(12) Hentze, M.; Kulozik, A. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell* **1999**, *96*, 307−310.

(13) Haas, J.; Park, E.-C.; Seed, B. Codon usage limitation in the expression of HIV-1 envelope glycoprotein. *Curr. Biol.* **1996**, *6*, 315.

(14) Rouwendal, G. J.; Mendes, O.; Wolbert, E. J.; Douwe de Boer, A. Enhanced expression in tobacco of the gene encoding green fluorescent protein by modification of its codon usage. *Plant Mol. Biol.* **1997**, *33*, 989−999.

(15) Stemmer, W. P.; Crameri, A.; Ha, K. D.; Brennan, T. M.; Heyneker, H. L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **1995**, *164*, 49−53..

(16) Shiver, J. W.; Fu, T. M.; Chen, L.; et al. Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. *Nature* **2002**, *415*, 331−335.

Accepted for publication October 28, 2003.

BP0300467