12

RATIONALES OF GENE DESIGN AND *DE NOVO* GENE CONSTRUCTION

Marcus Graf[1], Thomas Schoedl[1], and Ralf Wagner[1,2]

[1]*GENEART AG, Josef-Engert-Str. 11, Regensburg 93053, Germany*
[2]*University of Regensburg, Molecular Microbiology and Gene Therapy, Franz-Josef-Strauss Allee 11, Regensburg 93053, Germany*

12.1 INTRODUCTION INTO THE FIELD

As per definition, synthetic biology combines science and engineering in order to build novel biological functions and systems. Genetic engineering paved the way for the development of this new field of research; for instance, in 1978, when the Nobel prize in physiology or medicine was awarded to Werner Arber, Daniel Nathans, and Hamilton O. Smith for the discovery of restriction enzymes and their application to molecular genetics, one could already read in an editorial comment in *Gene* that ''. . .The work on restriction nucleases not only permits us easily to construct recombinant DNA molecules and to analyze individual genes but also has led us into the new era of synthetic biology where not only existing genes are described and analyzed but also new gene arrangements can be constructed and evaluated'' [1]. Another cornerstone in genetic engineering and synthetic biology was developed when Genentech scientists and their academic partners in 1977 generated the first example of recombinant expression of a human protein

Systems Biology and Synthetic Biology Edited by Pengcheng Fu and Sven Panke

(somatostatin) in *Escherichia coli.* However, not many scientists are today aware of the fact that the group of Boyer, Itakura, and colleagues were not only the first to invent heterologous recombinant expression but they also did so without using a natural gene. At that time, 9 years before the polymerase chain reaction (PCR) was introduced into genetic engineering, it was easier rationally to design the 14 amino acid long somatostatin gene and synthesize it with methods of organic chemistry than to clone it using the natural template. Since then, only very limited sequence information on the *E. coli* genome was available, codons preferentially used by the MS2 phage were assumed to be beneficial for recombinant gene expression in *E. coli*. However, most of the presently used genetically engineered biotherapeutics are based on natural genes cloned by reverse transcribing message RNAs (mRNAs) into complementary DNAs (cDNAs) and subsequent cloning using restriction enzymes and other DNA modifying enzymes. From 1986 on, in particular PCR-based cloning methods started soon dominating the field either by direct amplification of genomic information or using cDNA libraries as templates for amplification and subsequent cloning and recombinant expression.

As more genetic information became available, the clearer it got that the coding region of a gene comprises more information than just the primary amino acid sequence. The genetic code provides various codon options for each of the 20 amino acids that contribute to the primary sequence of a protein except for methionine and tryptophane, which are encoded by only one codon each. However, the codon options are used in an unequal frequency in different species showing a clear tendency for certain codons, which lead to the ''genome hypothesis'' postulated in 1980 by Grantham et al. [2]. They analyzed 90 genes of 7 different species and found a nonrandom codon choice pattern, which seemed to be specific for the analyzed species and therefore established the first codon usage table named ''codon catalog'' [2]. Just 1 year later with more genes analyzed the same group correlated a certain, species-specific subset of codons with mRNA expressivity, that is, the amount of protein made by a particular messenger transcript [3]. Also Ikemura, who synthesized the first human gene to be expressed in *E. coli* found a strong positive correlation between the transfer RNA (tRNA) abundance and the choice of codons in all *E. coli* genes encoding proteins that had been sequenced completely at that time [4]. The positive correlation of codon usage and the amount of available tRNAs in a cell was also confirmed in other unicellular (*Saccharomyces cerevisiae, Salmonella typhimurium*) and multicellular (human, rat, plant, chicken, fish) organisms [5] indicating that species-specific codon bias is a general phenomenon and can be even used to perform phylogenetic analyses (see Fig. 12-1).

After it was generally accepted that species-specific codon bias exists and influences expression rates, it seemed clear that recombinant gene expression can be improved by adapting at least the codon choice of the gene of interest to the preferred codons of the target expression host. In addition to codon choice, various intragenic *cis*-acting elements heavily impact expression yields in heterologous and even autologous cell factories. However, these elements like splice sites, TATA-boxes, or ribosomal entry sites are highly species-specific and have to be taken into account when rationally designing genes.

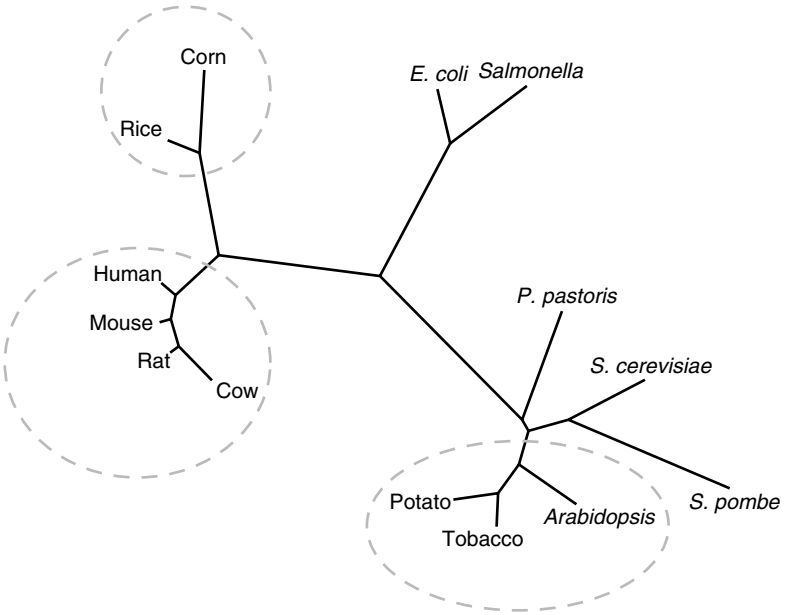

Figure 12-1 Phylogenetic tree based on codon usage. For each species pair the ''codon usage distance'' was calculated as follows: First, the difference of the frequency of each codon for each amino acid was calculated and then the total sum of frequency differences was computed. The resulting difference matrix was normalized with the highest difference to be set as 1. Phylogenetic branching of the data was performed using the Fitch–Margoliash least squares algorithm by the program ''Fitch'' (Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166). Codon usage data was obtained from the Kazusa Codon Usage Database (http://www.kazusa.or.jp/codon/). Kindly provided by M. Liss, GENEART (AG).

12.2 RATIONALS OF GENE DESIGN

12.2.1 Codon Usage Adaptation Strategies

The easiest and simplest approach to design an expression host-specific DNA sequence is to reversely translate (''backtranslate'') the amino acid sequence into a DNA sequence using the most frequently used codon. A comparably simple approach is to mimic the codon usage distribution that can be derived from the host's codon usage table. As most codon usage tables are generated on the genome-wide codon usage distribution, this approach levels out the natural differentiation between highly and low-expressed genes in terms of their codon usage. Using codon usage tables based on highly expressed genes bares several other unpredictable uncertainties: How many genes are sufficient for the compilation of the codon usage table? Is the different promoter strength of the highly expressed genes taken into account? Especially for multicellular organisms tissue and compartment specificity is often not addressed by those codon usage tables. The following chapters deal with the different codon adaptation and gene optimization strategies that should be applied for

different expression host systems. In order to be able to compare the overall codon quality of a given gene several measurements and indices were developed. The most accepted index is the codon adaptation index (CAI) introduced by Sharp and Li in 1987 [6]. The CAI is a measurement for the relative adaptiveness of the codon usage of a gene toward the codon usage of highly expressed genes. The relative adaptiveness of each codon is the ratio of the usage of each codon to that of the most abundant codon for the same amino acid. The CAI index is defined as the geometric mean of these relative adaptiveness values. Nonsynonymous codons and termination codons (dependent on genetic code) are excluded. CAI values are always between 0 (where no optimal codons are used) and 1 (where only optimal codons are used).

12.2.2 Designing Genes for Prokaryotic Expression *In Vivo*

The adverse effect of rare codons in recombinant expression in *E. coli* is known for many years. It was shown for instance that the production of β-galactosidase decreased when rare AGG codons were inserted near the translational initiation site [7]. Rare codons for arginine seem to be the most difficult to express, but other codons such as proline, glycine, or leucine are also problematic. Already very early studies revealed a strong bias in synonymous codon usage for genes encoding abundant proteins such as ribosomal proteins and elongation factors, RNA polymerase subunits, and glycolytic enzymes [3,8,9]. A strong correlation was found between copy numbers of protein and the frequency of codons whose cognate tRNAs were most abundant. Ikemura first discovered that this correlation was strongest in the most highly expressed genes which almost exclusively use optimal, that is, most frequently used codons [4,5,10] and he suggested that this bias in codon usage might both regulate gene expression and should act as an optimal strategy for recombinant gene expression. Although most frequently used codons correspond with the most abundant tRNAs, one has to keep in mind that the total tRNA composition of *E. coli* increases by 50 percent as growth rate increases to maximum [11]. However, the relative but not the absolute tRNA levels stay the same, thus the most frequently used codons still provide the largest tRNA pools feeding the translational machinery.

Therefore, it is now a widely accepted strategy for recombinant gene expression in *E. coli* to change the codons of the gene of interest or alternatively to coexpress certain tRNA-encoding genes in order to supplement tRNA pools with low abundance in *E. coli*. There are three *E. coli* strains commercially available (Novagen, Stratagene) coexpressing arginine encoding tRNA-genes specific for the nonfrequently used codons AGG, AGA, AUA, CUA, CCC, and GGA codons. Although coexpression of selected tRNAs can overcome certain expression problems due to the presence of extreme rare arginine, proline, or glycin codons, best and consistent results will be achieved only by adapting consequently the entire gene to most frequently used *E. coli* codons [12]. For instance, the sequence of the mature human IL-18 gene shows 37 rare *E. coli* codons (fraction <0.1) among 157 amino acids with an overall CAI of 0.58. Expression of nonoptimized IL-18 versus codon-optimized IL-18 was recently analyzed in depth. Although supplementation with rare tRNA genes did increase expression, Li et al. [13] found that a codon-optimized gene with a CAI of 0.84 was

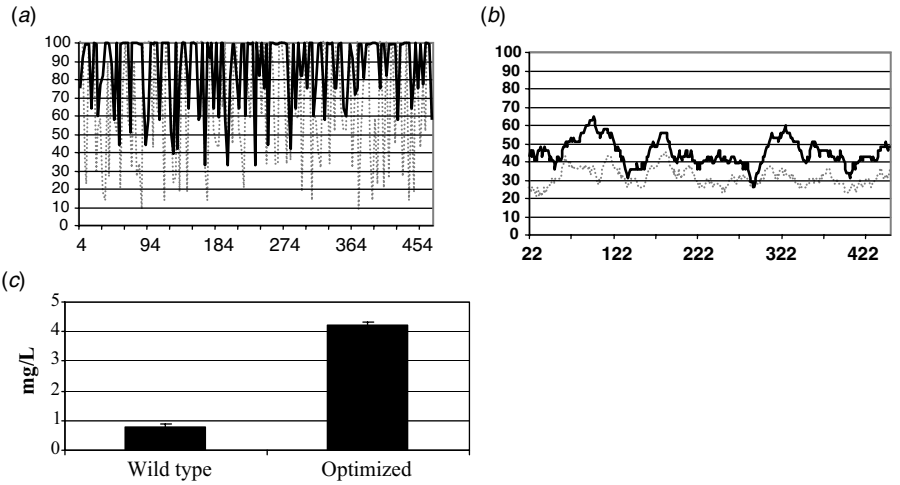

Figure 12-2 Codon, GC content, and expression analysis of human IL-18 expression in *E. coli*. (a) The plot shows the quality of the used codon at the indicated codon position. The quality value of the most frequently used codon was set to 100, the remaining synonymous codons were scaled accordingly (relative adaptiveness, see also Ref. [6]). Dotted line: wild-type gene; bold black line: optimized gene. (b) The plot shows the GC content in a 40 bp window centered at the indicated nucleotide position. Thin line: wild-type gene; bold black line: optimized gene. (c) Expression of wild-type and codon optimized human IL-18 in *E. coli* using BL-21 (DE3) strain 5 hours after induction. Modified from Ref. [6].

more beneficial and increased expression yields by a factor of five (see Fig. 12-2). Although the optimized gene shows elevated GC levels compared to the wild-type gene, which was increased from 35 percent up to 45 percent in average, the authors neither discuss nor find any correlation between elevated expression levels and GC content that could be thus solely coincidental. Interestingly, the authors also showed that the biochemical properties and protein activity of the proteins produced using wild-type gene expression and optimized gene expression were exactly the same.

In addition to codon choice, there also seems to be an influence of mRNA secondary structure on translation efficiency. Nomura et al., for instance, showed that the presence of an 8 bp stem-loop structure preceding the Shine Dalgarno ribosomal entry site may hamper expression and that destabilization of such an element could increase expression dramatically [14]. Moreover, the presence of intragenic *E. coli* ribosomal entry sites, as found in many mammalian genes, may lead to truncated products during heterologous expression and should therefore be avoided throughout the gene [15]. Hatfield et al. found by statistical analysis that certain codon pairs are seemingly overrepresented within the *E. coli* genome and reported a negative effect of such codon pairs on translational efficiency because of a translational pausing effect [16]. However, the negative activity of codon pairs could not be reproduced by another group utilizing the T7 promoter for transcription control [17].

Finally, termination efficiencies vary significantly depending on both the stop codon used and the nucleotide immediately following the stop codon. Efficiencies are ranging from 80 (TAAT) to 7 percent (TGAC) indicating that TAAT is the most

Table 12-1 Selected publications where synthetic genes were successfully used to express proteins in *E. coli* (modified from Ref. [69])

Gene Origin	Protein Name	Improvement	Reference
H. sapiens	IL2	16-fold	[19]
H. sapiens	TnT	10–40-fold	[20]
C. tetani	Fragment C	Fourfold	[21]
S. oleracea	Plastocyanin	1.2-fold	[22]
H. sapiens	Neurofibromin	Threefold	[23]
H. sapiens	M2-2	140-fold	[24]
H. sapiens	IL-6	Threefold	[25]
Pyrococcus abyssi	Phosphopantetheine adenylyltransferase	Below Detection to 15–20 mg/L	[26]

efficient translational termination sequence in *E. coli* and should be chosen to avoid undesired read through products [18].

12.2.3 Designing Genes for Prokaryotic Expression *In Vitro*

Despite coexpression of rare tRNA *in vitro* a more and more popular strategy to avoid expression problems *in vivo* is to choose *in vitro* transcription/translation systems. *In vitro* expression allows upscaling for production, which is, however, still comparably costly and more important, highly parallel screening of different protein variants for functional analysis.

To test whether gene optimization can increase protein yields of *in vitro* expression systems, differently optimized HIV-1 p24-encoding genes and wild-type p24 were translated *in vitro* under the transcriptional control of T7 in a Roche RTS 100 system using *E. coli* lysats (Table 12-1). Expression of different p24-encoding genes correlated nicely with the CAI of the respective genes. Lowest expression yields (770–831 μg/mL) were obtained using the wild-type p24 gene showing a $CAI_{E.\ coli}$ of 0.54. The mammalian-optimized gene with a $CAI_{E.\ coli}$ of 0,68 ($CAI_{H.\ sapiens} = 0.98$) showed a slight increase by 35–50 percent (1061–1170 μg/mL), but only the *E. coli*-optimized gene with a $CAI_{E.\ coli}$ of 0.98 raised protein yields by 3–3.5-fold (2565–2648 μg/mL) (Fig. 12-3). Gene optimization with a particular emphasis on codon choice improvement that results in high $CAI_{E.\ coli}$ values seems therefore beneficial to increase the yields when using *in vitro* expression systems.

Today, several companies offer cell-free expression systems that include chaperons and other proteins positively influencing correct folding of the translated protein. Alternatively, *in vitro* expression can be also performed in cells of higher eukaryotes such as rabbit reticulocytes or wheat germ lysats (Roche, Novartis, etc.).

12.2.4 Designing Genes for Yeast Expression

A very early study revealed in 1982 that an extreme codon bias is seen for the highly expressed *S. cerevisiae* genes alcohol dehydrogenase isozyme I (ADH-I) and glyceraldehyde-3-phosphate dehydrogenase. A proportion of more than 96 percent of

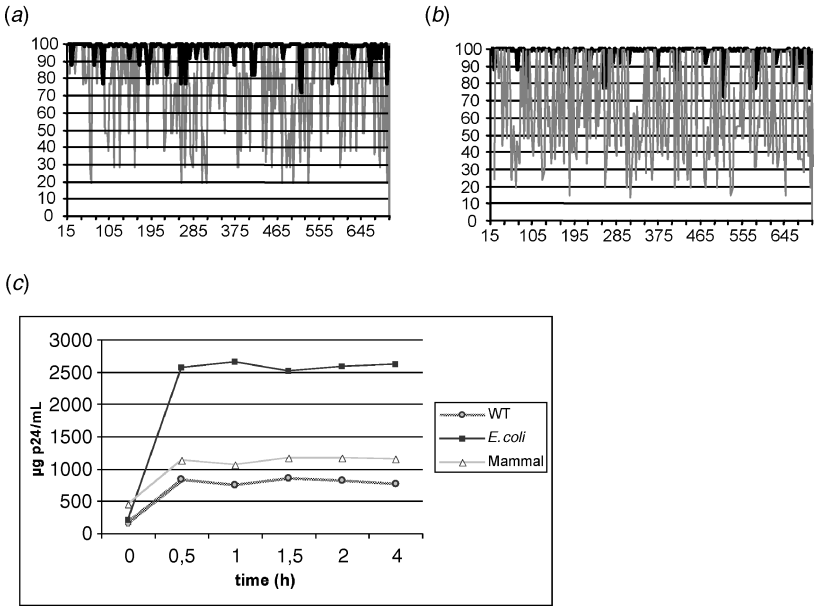

Figure 12-3 *In vitro* expression of differently optimized and wild-type HIV-1 p24 in *E. coli* lysats: (a, b) The plot shows the quality of the used codon at the indicated codon position. The quality value of the most frequently used codon was set to 100, the remaining synonymous codons were scaled accordingly (relative adaptiveness, see also Ref. [6]) gray line: wild-type gene or mammalian optimized in (a) and (b), respectively; bold black line: *E. coli*-optimized gene. (c) Expression of wild-type (WT), mammalian optimized (mammal), and *E. coli* optimized (*E. coli*) HIV-1 p24 in *E. coli* lysats under the transcriptional control of T7 using the Roche RTS system. Cell lysats (50 µL) were harvested at the indicated time points and p24 expression measured by commercial ELISA.

the 1004 amino acid residues are encoded by only 25 of the 61 possible coding triplets [27]. Only a few years later, Sharp et al. were able to show in a compiled analysis on 110 *S. cerevisiae* genes that there are two groups of genes. One group matched to highly expressed genes with a strong codon bias, which was speculated to match abundant tRNAs [28], whereas the other group corresponded to nonhighly expressed genes. More than 10 years later, Percudani et al. [29] were analyzing the now fully sequenced *S. cerevisiae* genome and were able to identify all tRNA-encoding genes. The respective 274 tRNA genes were assigned to 42 classes of distinct codon specificity. The gene copy number for individual tRNA species, which ranges from 1 to 16, correlated well with most frequently used *S. cerevisiae* codons. Moreover, they were able to show that tRNA gene copy numbers as well as codon frequencies nicely correlate with tRNA abundancy of the respective codon, thereby, confirming the previous assumptions of Sharp et al.

Other yeasts like *Pichia pastoris* or *Schizosaccharomyces pombe* also shows a highly biased, and species-specific codon choice (see Fig. 12-1) indicating that codon choice adaptation should also be beneficial for recombinant expression in these different yeasts. However, codon optimization procedures dedicated to obtain very high CAI values (> 0.95) will usually result in a significant decrease in the overall GC content due

to the relative low-GC bias of *P. pastoris* preferred codons. It was shown for instance that expression of *P. pastoris* codon-optimized human glucocerebroside fragments can be increased up to 10-fold above wild-type levels, hence almost the same levels (7.5-fold) can be achieved with a gene fragment in which rare codons were not removed, but the overall GC content was increased [30]. Consequently, it was shown by others that it is indeed beneficial not only to adapt codon choice but also to increase the GC content of the encoding gene for optimal expression in *P. pastoris* [31–33]. Therefore, an optimal balance has to be found between optimal codon choice and CAI values on the one hand and overall GC content on the other hand. To test whether this might be also true for *S. cerevisiae*, we tested expression of human MIP-1α-encoding genes with an optimal codon choice (CAI = 1.0), but low overall GC content (36 percent) and a gene with elevated GC content (47 percent), where we had to introduce several nonfrequently GC biased codons resulting in slightly lower CAI value (0.83). Protein yields were compared with gene expression resulting from nonaltered wild-type gene showing a GC content of 58 percent and a CAI of 0.56. As depicted in Figure 12-4 there seems to be no correlation between expression yields and elevated GC content in *S. cerevisiae* since highest expression was achieved with the fully codon-optimized gene showing the

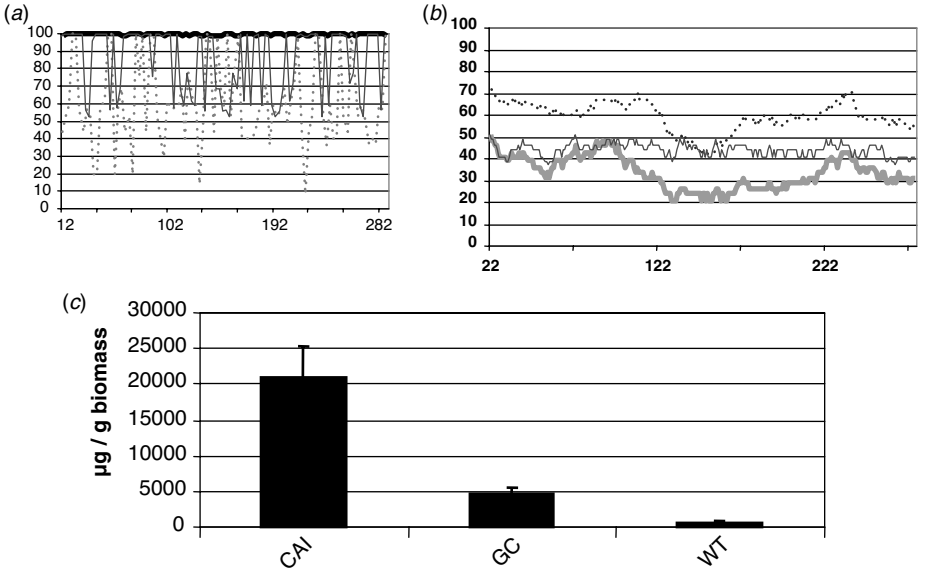

Figure 12-4 Expression of differently optimized and wild-type human MIP-1α in *S. cerevisiae*. (a) The plot shows the quality of the used codon at the indicated codon position. The quality value of the most frequently used codon was set to 100, the remaining synonymous codons were scaled accordingly (relative adaptiveness, see also Ref. [6]). Dotted line: wild-type gene; bold black line: codon-optimized gene; gray line: GC content-optimized gene. (b) The plot shows the GC content in a 40 bp window centered at the indicated nucleotide position. Dotted line: wild-type gene; bold black line: codon-optimized gene; gray line: GC content-optimized gene. (c) Transient expression of wild type (WT), GC content-optimized (GC), and codon-optimized (CAI) human MIP-1α in *S. cerevisiae* from lysed cell pellets using commercial ELISA formats.

lowest overall GC content. Expression yields exceeded GC-rich gene expression by a factor of 4.5 and by a factor of more than 35 compared to wild-type gene expression. We therefore conclude that for best expression results in *S. cerevisiae* codon choice seems to be a more dominant factor than GC content.

However, mRNA nucleotide composition itself has implications on the relative efficiency of protein expression through effects on secondary structures and stability irrespective of codon choice [34,35]. For instance, it seems advantageous to eliminate or avoid putative polyadenylation sites located in AT-rich DNA [36,37] in order to avoid premature polyadenylation.

12.2.5 Designing Genes for Plant Expression

The most prominent example and probably also the first example to express a rationally designed gene in higher plants is the insecticidal cry gene of *Bacillus thuringiensis*. Transgenic plants expressing active protein using wild-type cry genes failed to protect from insects due to poor expression. The use of different promoters, fusion proteins and leader sequences could not significantly increase protein expression either and failed to protect transgenic plants from insects [38,39]. Therefore, it was speculated that codon choice and the presence of sequence motifs not common in plants might hamper transgenic expression rates (see also Fig. 12-5). Table 12-2 indicates the differences in human, *B. thuringiensis*, rice (*Oryza sativa*), and *Arabidopsis thaliana* codon preference. Not only prokaryotic and mammalian codon choices differ greatly, but also the ones of monocots (rice) and dicots (*A. thaliana*), indicating that a general codon usage adaptation to plant genes will not reflect their phylogenetic diversity and consequently will not improve individual gene expression in the respective plant host.

Due to the preference for A or T in the third nucleotide position of preferred dicot codons (see Table 12-2), a simple reverse translation of the amino acid sequence into the respective encoding DNA sequence using the condon most frequently used by dicots will lead to a nonfavorable extremely AT-rich gene. Therefore, gene design for expression enhancement in dicots has to be carefully balanced between improving codon usage on the one hand and increasing GC content on the other hand.

Consequently, Perlak et al. tested differently modified cry-encoding genes for expression in tobacco and tomato plants. Among the plants transformed with the partially modified cry gene they identified a 10-fold higher and among plants transformed with the fully modified gene they identified 100-fold higher level of insect-control protein compared with plants expressing the wild-type gene ([40], see Table 12-3). Besides codon choice, they increased GC content, removed potential polyadenylation signals, and removed putative RNA instability elements, which have been shown previously to destabilize mRNA in other systems [41].

The very same gene design strategies were applied by others [42] to increase cryIC expression in dicots with similar positive results. Another important factor to keep in mind when designing genes for expression in plants is removal or avoidance of putative active splice motifs within the coding region. Rouwendal et al. for instance identified an 84 bp long cryptic intron within the coding region of the wild-type green fluorescent protein (GFP) encoding gene of the jellyfish *Aequorea victoria* after

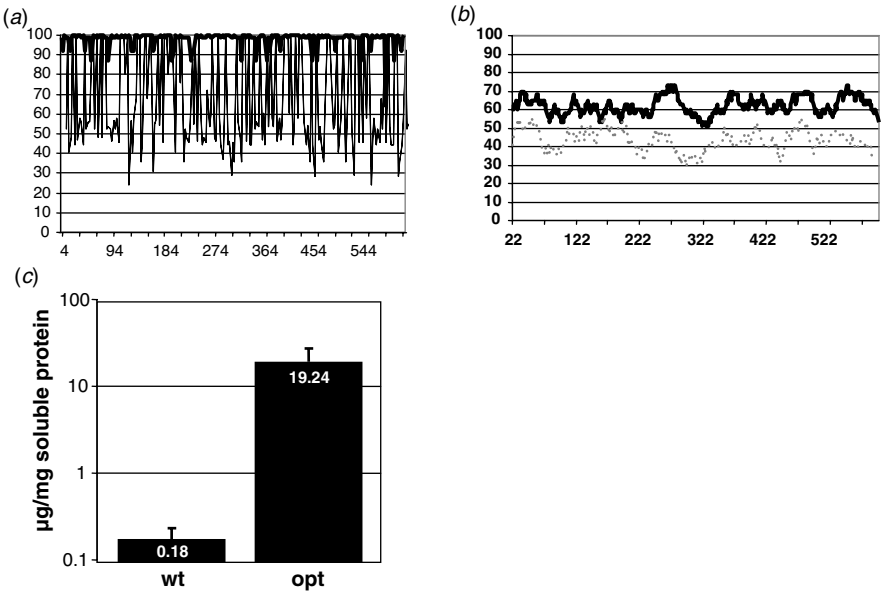

Figure 12-5 Codon, GC content, and expression analysis of (1.3–1.4)-b-glucanase tested for expression in barley. (a) The plot shows the quality of the used codon at the indicated codon position. The quality value of the most frequently used codon was set to 100, the remaining synonymous codons were scaled accordingly (relative adaptiveness, see also Ref. [6]). Dotted line: wild-type gene; bold black line: optimized gene. (b) The plot shows the GC content in a 40 bp window centered at the indicated nucleotide position. Thin line: wild-type gene; bold black line: optimized gene. (c) Expression of recombinant heat-stable (1.3–1.4)-b-glucanase under the control of the D-hordein gene (Hor3) promoter in T1 grains. For the codon-optimized gene, two individual grains were analyzed, and values are given. Modified from Refs [44,45].

transgenic expression in tobacco. Therefore, Rouwendal et al. adapted the codon usage of GFP for expression in tobacco. After codon adaptation where 42 percent of all codons were altered, the overall GC content was improved from 32 to 47 percent and the intragenic cryptic splice sites were altered to be nonfunctional. As expected, several transgenic tobacco lines containing the wild-type GFP gene contained a smaller nonfunctional protein cross-reacting with GFP antiserum, whereas only one protein of the predicted size was found in transgenics expressing the optimized GFP gene. Thus, the authors concluded that the smaller protein was probably encoded by a truncated GFP mRNA created by splicing of an 84 bp cryptic intron present only in the natural GFP-encoding gene [43].

In contrast to dicots, codon adaptation to monocots elevates the GC content automatically due to the preference for G or C in the third nucleotide position of frequently used monocot codons. Gene optimization can therefore be focused on removal of rare codons and other negatively *cis*-acting motifs as discussed above. Jensen et al., for instance, successfully expressed a synthetically engineered (1.3–1.4)-b-glucanase in barley (*Hordeum vulgare*) and were able to elevate expression levels to 107-fold compared to the wild-type gene ([44], see Fig. 12-5).

Table 12-2 Comparison of codon choice of selected amino acids for mammals, prokaryotes, and plants

Amino Acid	Codon	*B. thuringiensis*	*H. sapiens*	*O. sativa*	*A. thaliana*
	GCA	**100**	58	55	61
Ala	GCC	24	**100**	**100**	36
	GCG	41	28	88	32
	GCT	80	65	61	**100**
	AGA	**100**	**100**	61	**100**
	AGG	24	95	91	57
Arg	CGA	39	52	43	34
	CGC	17	90	**100**	20
	CGG	10	**100**	87	26
	CGT	56	38	48	49
	CTA	29	18	32	42
	CTC	8	50	**100**	65
Leu	CTG	10	**100**	82	42
	CTT	38	33	57	**100**
	TTA	**100**	18	25	54
	TTG	23	33	57	85
% GC	Total	37%	53%	56%	45%
% GC	Third	25%	59%	62%	42%

The most frequently used codon of each amino acid was set to 100 and the remaining scaled accordingly [6]. Black bold: most frequently used codon; gray bold: rare codon. GC total, overall GC content within all coding regions; GC third, GC content of the third nucleotide position of all codons.

Table 12-3 Optimization strategy for increased cry expression in tobacco and tomato

	Wild Type	Partially Modified	Fully Modified
Adapted codons	—	10%	60%
GC content	37%	41%	49%
Polyadenylation sites	18	7	1
RNA instability element	13	7	0

Modified from Ref. [40].

Taken together, there is great potential in increasing recombinant expression in transgenic plants using rationally designed genes. Codon choice, RNA instability motifs, and GC content seem to be important factors that have to be taken into account.

12.2.6 Codon Adaptation for Mammalian Expression

Synonymous codon choice also affects gene expression in mammals. In particular when nonmammalian genes are to be expressed in mammalian cells, the substitution of nonfrequently used codons with more common synonyms can significantly

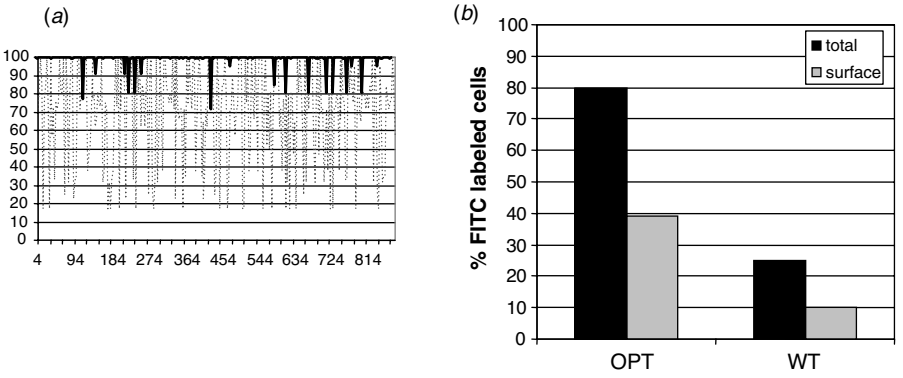

Figure 12-6 Surface and total cellular expression of HHV7 U51 in transiently transfected 293A cells was analyzed by flow cytometry. (a) The plot shows the quality of the used codon at the indicated codon position. The quality value of the most frequently used codon was set to 100, the remaining synonymous codons were scaled accordingly (relative adaptiveness, see also Ref. [6]). Dotted line: wild-type gene; bold black line: codon-optimized gene; gray line: GC-content-optimized gene. (b) Flow cytometry expression analysis of wild type and codon-optimized (OPT) HHV7 U51 in transiently transfected 293A using 2 μg plasmid DNA. Modified from Ref. [50].

increase expression [46–50]. A very prominent example is the widely used jellyfish *Aequorea victoria* green fluorescent protein. An analysis of the GFP encoding sequence showed that the codon usage frequencies of this jellyfish gene are quite different from those prevalent in the human genome [47]. Consequently, after having removed rare codons the synthetic humanized GFP allowed 5–10-fold higher expression rates compared to the wild-type cDNA in transfected mammalian cells. In another example, inhibition of expression of viral genes in mammalian cells could only be overcome by modification of the codon composition or by provision of excess tRNA [49]. A dramatic increase of 10–100-fold in gene expression was achieved when human herpesvirus (HHV) type 6- and HHV type 7-encoding genes were optimized for mammalian expression by adapting codon choice and elevating the overall GC content (see also Fig. 12-6).

More recently, Plotkin et al. [51] discovered systematic differences in synonymous codon usage between genes expressed in different human tissues. They were able to demonstrate that liver-specific genes differ in their codon choice from brain-specific genes, uterus differ from testis genes, etc. Since differences in relative tRNA abundancies in tissues of the same organisms were not reported so far, the authors suggested that codon mediated translational control might be the reason for the observed tissue-specific codon choice. Therefore, even codon choice optimization of mammalian genes (in particular nonhousekeeping genes) for autologous expression in mammalian cells can have a significant impact on the expression rates. For instance, we were able to show that expression of the human granulocyte macrophage colony stimulating factor (GM-CSF) could be increased by a factor of 2.1 by gene optimization and simultaneous removal of rare human codons and negative *cis*-acting RNA instability elements (see Fig. 12-7). In each codon position within the optimized

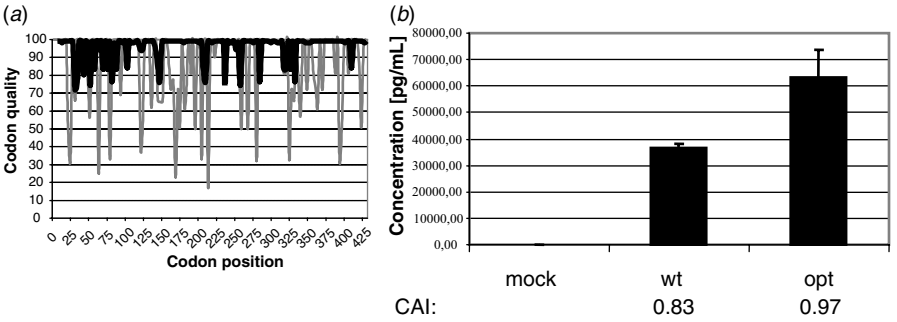

Figure 12-7 (a) The plot shows the quality of the used codon at the indicated codon position. The quality value of the most frequently used codon was set to 100, the remaining synonymous codons were scaled accordingly (relative adaptiveness, see also Ref. [6]). Gray line: wild-type gene; bold black line: optimized gene. (b) Transient transfection of human H1299 cells using wild-type (wt) and optimized (opt) human GM-CSF-encoding genes. Cytokine concentration in the supernatant was measured by commercial ELISA. At least five independent transfection experiments were performed and cytokine concentration measured. CAI, codon adaptation index.

coding region of GM-CSF where not the most frequent human codon was used, certain RNA instability elements have been avoided.

12.2.7 RNA Optimization for Mammalian Expression

Human codon usage bias seems not only to be correlating with available tRNAs but also synonymous codons are placed along coding regions in a way to minimize the number of TA and CG dinucleotides in mammalian genomes [52]. The rarity of CG dinucleotides in mammalian genes is usually ascribed to the tendency of CG to mutate to TG [53], whereas the rarity of TA in coding regions is considered adaptive because UA dinucleotides are preferably cleaved by endonucleases. Moreover, coding regions of mammalian housekeeping genes seem to have an increased GC content compared to low-expressing genes [54]. A correlation of mRNA expression levels with the third nucleotide position GC in codons of mice and rat genes was found by Konu et al., suggesting that higher GC levels may provide the rodent genes with a selective advantage for translational efficiency [55]. Consequently, we were able to show that mRNA stability of AT-rich HIV genes can be increased by orders of magnitude by elevating the GC content of the encoding mRNA from 44.1 to 62.7 percent in average thereby also reducing UA numbers from 96 to 11 (see Fig. 12-8). The increased amounts of available message lead to 100-fold higher expression of the HIV Pr55gag polyprotein compared to the nonoptimized cDNA-based gene expression [56]. This dramatic increase in expression cannot solely be explained by solely improved translational rates, rather then by the increased amounts of available optimized mRNA for cytoplasmatic expression. The optimized and wild-type RNA differ in their biochemical and physiological properties to such a great extent that they even use different nucleocytoplasmatic export pathways. Whereas expression of the wild-type HIV-1 gag genes was blocked using leptomycin B, an antibiotic known to directly interfere with exportin1-mediated nuclear export of mRNAs, the expression of the optimized RNA was not blocked [56].

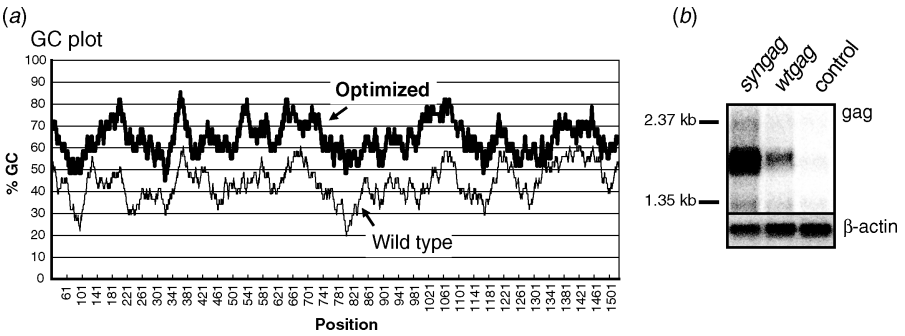

Figure 12-8 (a) The plot shows the GC content in a 40 bp window centered at the indicated nucleotide position. Thin line: wild-type gene; bold black line: optimized gene. (b) Transient transfection of human H1299 cells using wild-type (wtgag) and optimized (syngag) HIV-1 Pr55-encoding genes. Cells were lysed and total RNA was isolated and subjected to Northern Blot analysis. Pr55gag encoding transcripts were detected by a radiolabeled probe and standardized by the amount of detected β-actin RNA (lower panel). The Northern Blot analysis was repeated several times yielding comparable results. Modified from Refs [56,57].

Genes of other non-HIV related pathogens such as malaria or human papilloma virus, or certain proto-oncogenes show similar sequence properties indicating a general mode of action of these *cis*-acting RNA elements and the positive effects of RNA optimization.

When designing genes for optimal expression in mammalian cells, one therefore has to take several parameters into account. Codon choice is one parameter to influence the translation efficiency but, more importantly, the available mRNA levels necessary for translation have a much larger impact on expression yields. RNA optimization, like modification of GC content, avoidance or preferable introduction of UA/CG dinucleotides, removal of *cis*-acting RNA elements negatively influencing expression is the key for stable and high level expression in mammalian cells. Examples for these *cis*-acting RNA elements are adenine-rich elements (ARE), which are found within many cytokines, nuclear transcription factors or proto-oncogenes or within their flanking untranslated regions and are thought to be the most common determinant of RNA instability in mammalian cells [58,59]. Other, probably closely related, *cis*-acting elements are known to destabilize or retain mRNAs and should be avoided [60–63]. Despite these elements and ARE-directed mRNA degradation it is known that isolated cryptic splice sites may retain mRNA within the splicing machinery or lead to truncated transcripts [64].

12.2.8 Other Important Designing Rules for Mammalian Expression

In addition to optimizing the coding region for improved expression yields in mammalian cells, there are other sequence elements known to be beneficial for efficient and stable expression. In mammalian mRNAs, initiation sites usually align to all or part of the sequence GCCRCCAUGG referred to as the Kozak sequence [65]. A strong contribution of G directly downstream of the starting ATG was confirmed in a study that directly monitored the initiation step [66], thus negating the concern that this

conservation might simply reflect a preference for certain amino acids in the second position of eukaryotic proteins. Not only the efficiency of the translational initiation is context dependent but also the efficiency of translational termination. In order to avoid undesired read through by-products in mammalian expression, two stop codons are usually recommended to ensure efficient termination. Finally, natural mammalian transcripts always comprise and intron/exon structure and undergo complex splicing and editing before leaving the nucleus for cytoplasmatic translation. Therefore, expression of intronless cDNAs or artificial genes can be improved by adding an efficiently spliced intron 5′ of the coding region [67]. Finally, there is a recent report that by removal/avoidance of certain codon pairs gene expression can be significantly increased [68].

Taken together, gene design for mammalian expression is a sophisticated task since many different parameters have to be optimized in parallel for optimal results. Table 12-3 shows selected examples from literature where gene optimizations led to increased expression results compared to wild-type expression.

12.2.9 Methods of Analyzing Genes

Most standard software offers the possibility to analyze genes, for example, concerning their codon usage or GC content. Tables 12-4 and 12-5 list easy to handle Internet-based software for the analysis of genes with respect to codon usage, prediction of repetitive elements, and splice sites.

Further link compilations can be found on http://bioweb2.pasteur.fr/ and http://www.expasy.org/tools/. Nevertheless all those software tools just concentrate on the evaluation of one single parameter. As shown above, especially when it comes to gene optimization, several parameters have to be considered important. This complicates gene optimization as it becomes a multiparameter and multitask problem: Multiparameter as several constraints like codon choice or GC content have to be accounted for, multitask as several jobs like local and global sequence alignments have to be performed. This can only be achieved by a multiparameter process that simultaneously takes into account all constraints. State-of-the-art multiparameter optimization software (GeneOptimizer™, Geneart AG, Regensburg, Germany) allows for different weighting of the constraints and evaluates the quality of codon combinations concurrently. Approaches that are based on a successive modulation of parameters most likely spoil findings of previous optimization runs in subsequent runs (e.g., first optimization focusing on codon choice, second on GC content).

12.3 *DE NOVO* GENE CONSTRUCTION

12.3.1 Oligonucleotide Synthesis—Creating Your Template

Since rationally designed DNA comprising genes, operons, or even genomes only exist *in silico*, nature cannot provide a natural template for PCR-based amplification. Therefore, the only possible way to get access to such a rationally designed DNA is by *de novo* gene synthesis. Virtually, all gene synthesis methods are based on oligonucleotide synthesis causing many technical as well as cost implications.

Table 12-4 Selected publications where synthetic genes were used to express proteins in mammalian cells (modified from Ref. [69])

Gene Origin	Protein Name	Expression Host	Improvement	Reference
HIV	gp120	*H. sapiens*	Expression: >40-fold	[70]
Aequorea victoria	GFP	*H. sapiens*	Expression: 22-fold	[47]
Listeria monocytogenes	LLO	*M. musculus*	Expression: 100-fold	[71]
BPV1	L1, L2	Mammalian	Expression: 1.000-fold	[49]
HIV	gag	*H. sapiens*	Expression: >322-fold	[72]
HIV	gag	*H. sapiens*	Expression: >10–100-fold	[56]
Plasmodium spp.	EBA-175 region II and MSP-1	*M. musculus*	Expression: 4-fold	[73]
Tn10/HSV	rtTA	*M. musculus*	Expression: >20-fold	[74]
HPV	L1	*H. sapiens*	Expression: 1.000–10.000-fold	[75]
P1 phage	Cre	Mammalian	Expression: 1.6-fold	[76]
Schistosoma mansoni	SmGPCR	*H. sapiens*	Expression: barely detectable versus strong signal	[77]
HV-fold	U51	Mammalian	Expression: 10–100-fold	[50]
HIV	gag, pol, env, nef	*H. sapiens*	Expression: >250×, >250×, >45×, >20×, respectively	[78]
HPV	E5	Mammalian	Expression: 6–9-fold	[79]
HPV	E7	Mammalian	Expression: 20–100-fold	[80]
HIV, SIV	GagPol	*H. sapiens*	Safety in gene therapy	[80]
HIV	Gag	*H. sapiens*	Efficacy, vaccine development	[81]
HIV	GagPolNef	*H. sapiens*	Efficacy, safety for vaccine development	[82]
H. sapiens	Dopamine receptor D2	*H. sapiens*	RNA stability	[52]
Choristoneura fumiferana	GE, GEvy, VR	*H. sapiens*	Increase in reporter induction: 2.7- and 1.7-fold	[83]
Coronavirus	Spike and nucleocapsid proteins	*H. sapiens*	Safe and efficient expression for studying antibody binding	[84]

Table 12-5 Selected web-based software tools for gene analysis

Codon usage distribution and codon usage table compilation
http://www.kazusa.or.jp/codon/countcodon.html
http://www.bioinformatics.org/sms2/codon_usage.html
http://gcua.schoedl.de/
http://www.bioinformatics.org/sms2/codon_plot.html
Repetitive or secondary structure elements
http://www.genebee.msu.su/services/rna2_reduced.html
http://bioweb.pasteur.fr/seqanal/interfaces/dottup.html
Splice site prediction
http://www.cbs.dtu.dk/biolinks/pserve2.php
http://genes.mit.edu/GENSCAN.html
http://www.fruitfly.org/seq_tools/splice.html
Reverse translation
http://www.bioinformatics.org/sms2/rev_trans.html

The chemical synthesis of short single strand DNA dates back to 1955 and provided the directed chemical synthesis of a dithymidinyl nucleotide [85]. Over the decades oligonucleotide chemistry was improved, but it took until the beginning of the 1980s when the phosphoramidite approach and solid phase synthesis allowed the development of routine oligonucleotide synthesis, as we know it today. The oligonucleotide synthesis starts on a solid phase (controlled pore glass CPG) in 3′ to 5′ direction. The first 3′ nucleotide is already immobilized on the CPG solid phase and a synthesis cycle comprising four steps builds up the desired oligonucleotide. Briefly, the first step is deblocking which removes a protective group of the 5′-end thereby exposing a reactive hydroxyl group at which the next nucleotide (phosphoramidite) is added (coupling step) after activation. An overview of the cycle is depicted in Fig. 12-9.

The capping step prevents all oligonucleotides that were not elongated in the previous coupling step from subsequent coupling steps. Finally, oxidation step stabilizes the phosphate bond of the newly added amidite and the cycle may commence from the beginning until the full-length oligonucleotide is synthesized. During synthesis, there are several steps of this cycle that can lead either to deletions or nucleotide substitutions within the desired oligonucleotide for instance due to inefficient capping, deblocking, or removal of purines in an acidic solution. No matter what gene synthesis method is used to build up long DNA fragments, the sequence identity and the respective error frequency of the used oligonucleotides has a major impact on costs, efforts, and duration of subsequent *de novo* gene production.

12.3.2 *De Novo* Gene Synthesis Methods

The field of gene synthesis was pioneered by Khorana et al. with the groundbreaking work to synthesize a full-length tRNA encoding sequence, which took them several years [86]. Koester et al. synthesized the first protein-encoding gene (angiotensin II)

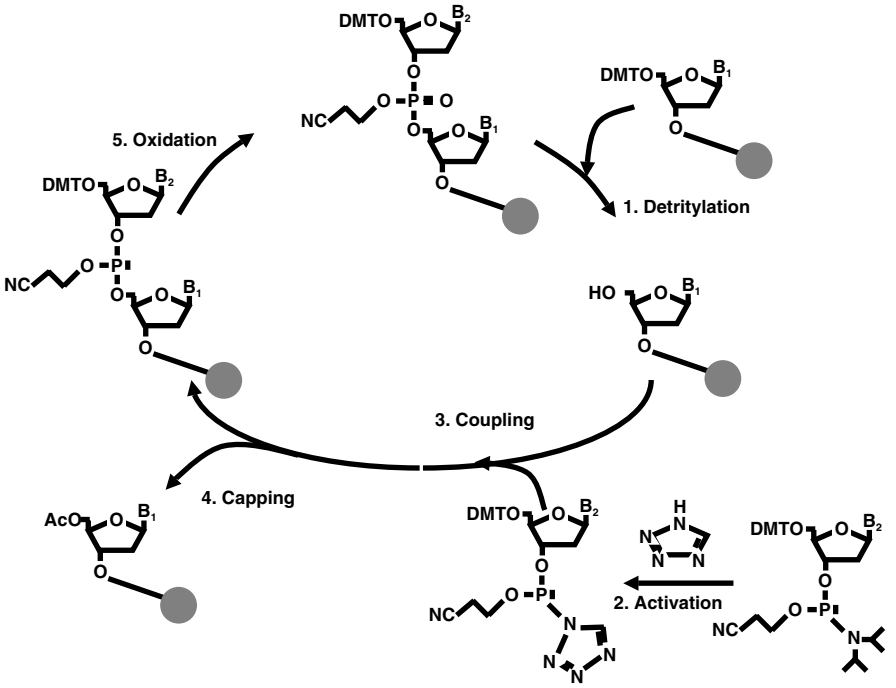

Figure 12-9 Oligonucleotide synthesis cycle description (see text). CPG, controlled pore glass, solid phase. B_1, base 1; B_2, base 2; DMTO, dimethoxytrityl; Ac, acid.

just 5 years later and Itakura et al. revolutionized the field in 1977 with the synthesis of a human gene in just 6 months [87,88]. Although Itakura et al.'s work gave birth to the field of genetic engineering and recombinant DNA technology, gene synthesis was restricted these days by the limited availability of synthetic oligonucleotides. DNA synthesis methods and molecular biology methods needed to coevolve to allow high-throughput gene synthesis, as we know it today from specialized gene synthesis laboratories, some of which are capable to produce several hundreds of genes per month.

In principle, all gene synthesis methods used so far rely on the elongation of hybridized oligonucleotides with long overlaps or the ligation of the phosphorylated oligonucleotides. The latter technique was used already by Koester et al. who phosphorylated oligonucleotides by T4 polynucleotide kinase and joined them using T4 ligase giving rise to a double-stranded DNA consisting of 33 bp. A similar approach was used by Edge et al. to synthesize long DNA fragments, which were assembled and ligated to a 514 bp long human leukocyte interferon encoding synthetic gene in 1981 [89]. After Francis Barany introduced the ligase chain reaction (LCR) using a heat stable ligase [90], it was possible to anneal oligonucleotides for *de novo* gene synthesis at high temperatures making ligation of phosphorylated oligonucleotides a very robust, but yet labor-intensive, time consuming, and expensive gene synthesis strategy [91].

In 1982, 3 years before the PCR was invented by Kary B. Mullis, the gene synthesis pioneer Itakura introduced a primer extension method enabling him to build up short genetic sequences *de novo*. A subsequently filed patent application was granted many years later in 1997 and is thus still active within the United States for many years to come. After the introduction of PCR into genetic engineering in 1985, several PCR-based oligonucleotide assembly methods emerged but all of these are based on one or more primer extension steps with subsequent amplification. This has to be kept in mind when using PCR-based assembly methods and resulting genes for commercial purposes or for clinical testing. By applying PCR-/primer-extension-based methods soon the 1000 bp size barrier was broken in 1990 by the synthesis of a 2.1 kb long fully synthetic plasmid [92]. Stemmer et al. used 132 oligonucleotides in a single primer extension reaction of overlapping complementary oligonucleotides with subsequent PCR amplification to assemble a 2.7 kb sized plasmid [93]. Similar approaches were used by Withers-Martinez et al. in 1999, who assembled an apparently difficult to construct AT-rich malaria gene [94]. In Figure 12-10 a schematic drawing shows the principles of ligase-based and PCR-/primer-extension-based gene synthesis methods.

Surely, one major cost factor in gene synthesis is still the bulk of oligonucleotides needed for *de novo* gene construction. To reduce oligonucleotide costs,

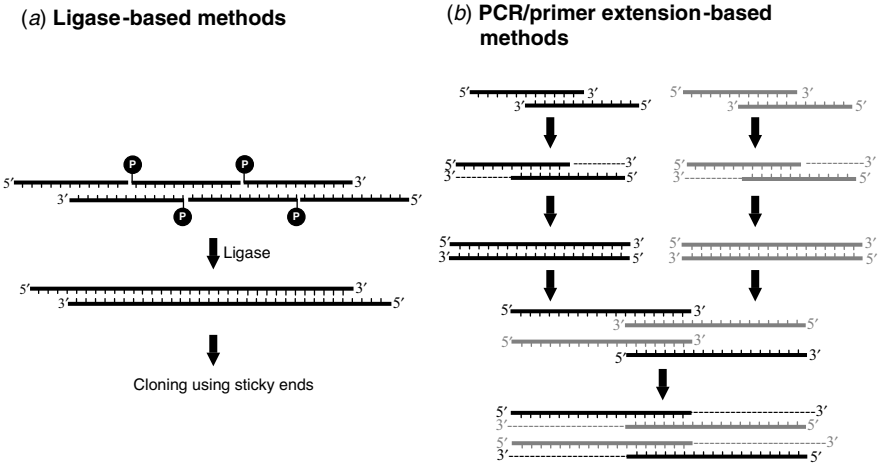

Figure 12-10 Assembly of complementary overlapping oligonucleotides. (a) The inner oligonucleotides need to be phosphorylated (P) and no gaps should separate the oligonucleotides. To allow subsequent cloning, 5′ overhangs should be added to the flanking oligonucleotides thereby creating sticky ends which are compatible with the cohesive ends created after restriction digestion of a suitable cloning vector. After annealing the oligonucleotides can be ligated, the ligation product can be purified and cloned in a suitable cloning vector. (b) PCR cycling allows the annealing of overlapping oligonucleotides and their elongation. The assembled sequence may be subsequently amplified using flanking PCR primer oligonucleotides and cloned into a suitable cloning vector.

George Church et al. recently combined photo-programmable microfluids chips with classic PCR-based gene synthesis techniques. Although use of chemicals can be dramatically reduced, the oligonucleotides are present after release into solution only in femtomolar concentrations per sequences that are insufficient to allow bimolecular priming necessary for *de novo* gene construction. Therefore, the oligonucleotides released from the chip need to be reamplified by PCR using universal primers, endonuclease treatment to remove the universal priming region, and final purification. These extensive post synthesis treatments in addition to the setup costs of the oligonucleotide chip add up to the overall costs of this new and visionary gene synthesis method [95].

All gene synthesis methods so far have in common that they totally rely on quality and sequence identity of the oligonucleotides used in the assembly process. Due to the nature of the chemical synthesis of oligonucleotides there will be a certain fraction of oligonucleotides present showing deviations (like deletions, duplications, substitutions) from the desired end product. To test the quality and sequence correctness of oligonucleotides, we analyzed different batches and differently long oligonucleotides from a commercial source for gene synthesis. Oligonucleotides of 24mer, 44mer, and 64mer were used to assemble a 513 bp long fully artificial gene encoding all epitopes of the HIV-1 tat gene. A ligase-based gene assembly technique was used to avoid introduction of PCR-based mutations. After gene assembly, cloning and transformation 161 different clones were established, sequenced, and analyzed (see Fig. 12-11a). Apparently, there are huge deviations in oligonucleotide quality that differ from batch to batch even when the same oligonucleotide sequence is ordered. There seems to be a clear tendency that the longer the oligonucleotides are the higher the likelihood of mutations will be within the final clones (see Fig. 12-11b). This is most probably due to the fact that an increasing number of chemical reactions

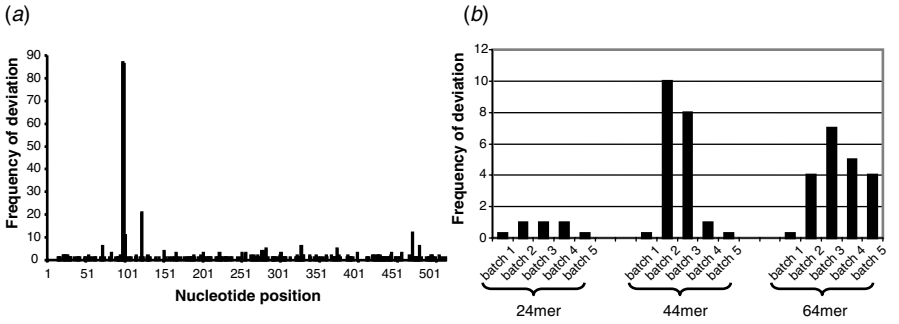

Figure 12-11 (a) Sequence analysis of a 513 bp long synthetic gene. The gene was assembled using 24 different overlapping, phosphorylated oligonucleotides using a ligase-based approach. The synthetic genes were cloned and 161 colonies sequenced. Frequency of deviations was plotted along the sequence. (b) Within the 513 long synthetic gene distinct oligonucleotides were replaced with a 24mer, 44mer and 64mer to test the influence of oligonucleotide length error frequency found in clones. Five different batches of each oligonucleotide sequence were tested and cloned. Twenty clones of each batch were analyzed by sequencing and frequency of deviations analyzed.

take place during synthesis leading to an increase of undesired side products. As deduced from this example it is most likely to get a statistical clustering of mutations within a synthetic gene. In our case, almost 90 percent of all clones showed mutations in one specific oligonucleotide, resulting in a deletion frequency of approximately 55 percent at position 97 and 98, respectively. Consequently, only 2 of 161 sequenced clones showed no deviation from the desired sequence design. Obviously, it would have been cheaper and easier to replace the low-quality oligonucleotide batch with new one resulting in a much higher likelihood to identify a 100 percent correct clone. In addition, gene length has clearly also influence on the chance to identify 100 percent correct clones by sequencing screening. When 50mer oligonucleotides with a given deviation frequency of approximately 10 percent would be used to build up a 500 bp long synthetic gene the probability of identifying a correct clone will be 12 percent (0.90^{20}, 20 oligonucleotides used in assembly). One would therefore need to screen at least 12–24 colonies to identify a 100 percent correct clone. However, to assemble a 1000 bp gene one would need at least 40 oligonucleotides to build up a synthetic gene. Subsequently, only 1.5 percent (0.90^{40}) of all screened clones would be correct. By doubling the gene length the screening effort increased by a factor of 10!

Consequently, one should therefore use only oligonucleotides with highest possible sequence identy, which are unfortunately not easy to access from commercial oligonucleotide suppliers. Alternative strategies were published by several groups trying to eliminate the fraction of mutated genes from the initial oligonucleotide assembly product in order to reduce the sequencing screening workload in gene synthesis. By denaturation and annealing of the crude oligo assembly product, heteroduplexes are most probably formed by DNA products showing mutations and homoduplexes are formed with the highest probability only from such DNA sequences showing no mutations. In a subsequent step, the heteroduplex fraction is removed by either enzymatic treatment, for example, T4 endonuclease VII cleavage [91,96] or by addition of proteins binding specifically to heteroduplex DNA and subsequent homoduplex enrichment [97].

More recently, Cello et al. reported the synthesis of a 7.5 kb cDNA poliovirus by a combination of PCR-based oligonucleotide assembly/PCA (polymerase cycling assembly) and ligation methods [98]. Being one of the first genomes to be fully created synthetically, the generation of an infectious and pathogenic virus based entirely on a *de novo* constructed genome will surely be considered as a hallmark in synthetic biology and even raised further ethical questions regarding biosafety of such synthetic genomes. Using similar approaches, only 2 years later, Kodumal et al. of Kosan Biosciences reported the synthesis of a contiguous 32 kb polyketide synthase gene cluster being the longest synthetic DNA assembled from synthetic oligonucleotides so far [99]. The functionality of the gene cluster was demonstrated by successfully expressing the polyketide synthase and producing its polyketide product in *E. coli*, being the first report of a functionally operon synthesized and assembled *de novo*. Just 2 years later, the same laboratory reported the synthesis of a redesigned polyketide synthase gene cluster expressing significantly more protein that the wild-type cluster did [100]. The last remarkable cornerstone was set here by Craig Venter who managed

to synthesise and artificially construct the first bacterial genome of more than 500.000 bp in size [101].

Today, it is therefore not unrealistic to predict that in the years to come more and more synthetic operons, artificial chromosomes, and synthetic genomes will be synthesized being the major enabling technology for the new field of synthetic biology.

REFERENCES

1. Szybalski W, Skalka A. Nobel prizes and restriction enzymes. *Gene* 1978;4:181–182.
2. Grantham R, Gautier C, Gouy M, Mercier R, Pave A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* 1980;8:r49–r62.
3. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* 1981;9:r43–r74.
4. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 1981;151: 389–409.
5. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 1985;2:13–34.
6. Sharp PM, Li WH. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987;15: 1281–1295.
7. Chen GF, Inouye M. Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. *Nucleic Acids Res* 1990;18:1465–1473.
8. Grosjean H, Fiers W. Preferential codon usage in prokaryotic genes: the optimal codon–anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* 1982;18:199–209.
9. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* 1982;10:7055–7074.
10. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* 1982;158:573–597.
11. Emilsson V, Naslund AK, Kurland CG. Growth-rate-dependent accumulation of twelve tRNA species in *Escherichia coli*. *J Mol Biol* 1993;230:483–491.
12. Zhou Z, Schnake P, Xiao L, Lal AA. Enhanced expression of a recombinant malaria candidate vaccine in *Escherichia coli* by codon optimization. *Protein Expr Purif* 2004;34:87–94.
13. Li A, Kato Z, Ohnishi H, Hashimoto K, Matsukuma E, Omoya K, Yamamoto Y, Kondo N. Optimized gene synthesis and high expression of human interleukin-18. *Protein Expr Purif* 2003;32:110–118.
14. Nomura M, Ohsuye K, Mizuno A, Sakuragawa Y, Tanaka S. Influence of messenger RNA secondary structure on translation efficiency. *Nucleic Acids Symp Ser* 1984;(15): 173–176.

15. Ivanov IG, Alexandrova R, Dragulev B, Leclerc D, Saraffova A, Maximova V, Abouhaidar MG. Efficiency of the 5′-terminal sequence (omega) of tobacco mosaic virus RNA for the initiation of eukaryotic gene translation in *Escherichia coli*. *Eur J Biochem* 1992;209:151–156.
16. Irwin B, Heck JD, Hatfield GW. Codon pair utilization biases influence translational elongation step times. *J Biol Chem* 1995;270:22801–22806.
17. Cheng L, Goldman E. Absence of effect of varying Thr-Leu codon pairs on protein synthesis in a T7 system. *Biochemistry* 2001;40:6102–6106.
18. Poole ES, Brown CM, Tate WP. The identity of the base following the stop codon determines the efficiency of *in vivo* translational termination in *Escherichia coli*. *EMBO J* 1995;14:151–158.
19. Williams DP, Regier D, Akiyoshi D, Genbauffe F, Murphy JR. Design, synthesis and expression of a human interleukin-2 gene incorporating the codon usage bias found in highly expressed *Escherichia coli* genes. *Nucleic Acids Res* 1988;16:10453–10467.
20. Hu X, Shi Q, Yang T, Jackowski G. Specific replacement of consecutive AGG codons results in high-level expression of human cardiac troponin T in *Escherichia coli*. *Protein Expr Purif* 1996;7:289–293.
21. Makoff AJ, Oxer MD, Romanos MA, Fairweather NF, Ballantine S. Expression of tetanus toxin fragment C in *E. coli*: high level expression by removing rare codons. *Nucleic Acids Res* 1989;17:10191–10202.
22. Ejdeback M, Young S, Samuelsson A, Karlsson BG. Effects of codon usage and vector–host combinations on the expression of spinach plastocyanin in *Escherichia coli*. *Protein Expr Purif* 1997;11:17–25.
23. Hale RS, Thompson G. Codon optimization of the gene encoding a domain from human type 1 neurofibromin protein results in a threefold improvement in expression level in *Escherichia coli*. *Protein Expr Purif* 1998;12:185–188.
24. Johansson AS, Bolton-Grob R, Mannervik B. Use of silent mutations in cDNA encoding human glutathione transferase M2-2 for optimized expression in *Escherichia coli*. *Protein Expr Purif* 1999;17:105–112.
25. Li Y, Chen CX, von Specht BU, Hahn HP. Cloning and hemolysin-mediated secretory expression of a codon-optimized synthetic human interleukin-6 gene in *Escherichia coli*. *Protein Expr Purif* 2002;25:437–447.
26. Nalezkova M, de Groot A, Graf M, Gans P, Blanchard L. Overexpression and purification of *Pyrococcus abyssi* phosphopantetheine adenylyltransferase from an optimized synthetic gene for NMR studies. *Protein Expr Purif* 2005;39:296–306.
27. Bennetzen JL, Hall BD. Codon selection in yeast. *J Biol Chem* 1982;257:3026–3031.
28. Sharp PM, Tuohy TM, Mosurski KR. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* 1986;14:5125–5143.
29. Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol* 1997;268:322–330.
30. Sinclair G, Choy FY. Synonymous codon usage bias and the expression of human glucocerebrosidase in the methylotrophic yeast, *Pichia pastoris*. *Protein Expr Purif* 2002;26:96–105.
31. Woo JH, Liu YY, Mathias A, Stavrou S, Wang Z, Thompson J, Neville DM Jr. Gene optimization is necessary to express a bivalent anti-human anti-T cell immunotoxin in *Pichia pastoris*. *Protein Expr Purif* 2002;25:270–282.

32. Outchkourov NS, Stiekema WJ, Jongsma MA. Optimization of the expression of equistatin in *Pichia pastoris*. *Protein Expr Purif* 2002;24:18–24.
33. Gurkan C, Ellar DJ. Expression in *Pichia pastoris* and purification of a membrane-acting immunotoxin based on a synthetic gene coding for the *Bacillus thuringiensis* Cyt2Aa1 toxin. *Protein Expr Purif* 2003;29:103–116.
34. Oliveira CC, van den Heuvel JJ, McCarthy JE. Inhibition of translational initiation in *Saccharomyces cerevisiae* by secondary structure: the roles of the stability and position of stem-loops in the mRNA leader. *Mol Microbiol* 1993;9:521–532.
35. Seffens W, Digby D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res* 1999;27:1578–1584.
36. Romanos MA, Makoff AJ, Fairweather NF, Beesley KM, Slater DE, Rayment FB, Payne MM, Clare JJ. Expression of tetanus toxin fragment C in yeast: gene synthesis is required to eliminate fortuitous polyadenylation sites in AT-rich DNA. *Mol Biol* 1991;19:1461–1467.
37. Scorer CA, Buckholz RG, Clare JJ, Romanos MA. The intracellular production and secretion of HIV-1 envelope protein in the methylotrophic yeast *Pichia pastoris*. *Gene* 1993;136:111–119.
38. Hofte H, Seurinck J, Van Houtven A, Vaeck M. Nucleotide sequence of a gene encoding an insecticidal protein of *Bacillus thuringiensis* var. *tenebrionis* toxic against Coleoptera. *Nucleic Acids Res* 1987;15:7183.
39. Muller BC, Raphael AL, Barton JK. Evidence for altered DNA conformations in the simian virus 40 genome: site-specific DNA cleavage by the chiral complex lambda-tris(4,7-diphenyl-1,10-phenanthroline)cobalt(III). *Proc Natl Acad Sci USA* 1987;84: 1764–1768.
40. Perlak FJ, Fuchs RL, Dean DA, McPherson SL, Fischhoff DA. Modification of the coding sequence enhances plant expression of insect control protein genes. *Proc Natl Acad Sci USA* 1991;88:3324–3328.
41. Shaw G, Kamen R. A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell* 1986;46:659–667.
42. Strizhov N, Keller M, Mathur J, Koncz-Kalman Z, Bosch D, Prudovsky E, Schell J, Sneh B, Koncz C, Zilberstein A. A synthetic cryIC gene, encoding a *Bacillus thuringiensis* delta-endotoxin, confers *Spodoptera* resistance in alfalfa and tobacco. *Proc Natl Acad Sci USA* 1996;93:15012–15017.
43. Rouwendal GJ, Mendes O, Wolbert EJ, 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.
44. Jensen LG, Olsen O, Kops O, Wolf N, Thomsen KK, von Wettstein D. Transgenic barley expressing a protein-engineered, thermostable (1,3-1,4)-beta-glucanase during germination. *Proc Natl Acad Sci USA* 1996;93:3487–3491.
45. Horvath H, Huang J, Wong O, Kohl E, Okita T, Kannangara CG, von Wettstein D. The production of recombinant proteins in transgenic barley grains. *Proc Natl Acad Sci USA* 2000;97:1914–1919.
46. Levy JP, Muldoon RR, Zolotukhin S, Link CJ Jr. Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. *Nat Biotechnol* 1996;14:610–614.
47. Zolotukhin S, Potter M, Hauswirth WW, Guy J, Muzyczka N. A ''humanized'' green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J Virol* 1996;70:4646–4654.

48. Wells KD, Foster JA, Moore K, Pursel VG, Wall RJ. Codon optimization, genetic insulation, and an rtTA reporter improve performance of the tetracycline switch. *Transgenic Res* 1999;8:371–381.
49. Zhou J, Liu WJ, Peng SW, Sun XY, Frazer I. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. *J Virol* 1999; 73:4972–4982.
50. Bradel-Tretheway BG, Zhen Z, Dewhurst S. Effects of codon-optimization on protein expression by the human herpesvirus 6 and 7 U51 open reading frame. *J Virol Methods* 2003;111:145–156.
51. Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. *Proc Natl Acad Sci USA* 2004;101:12588–12591.
52. Duan J, Antezana MA. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. *J Mol Evol* 2003;57:694–701.
53. Salser W. Globin mRNA sequences: analysis of base pairing and evolutionary implications. *Cold Spring Harb Symp Quant Biol* 1978;42(Part 2): 985–1002.
54. Lercher MJ, Hurst LD. Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? *Gene* 2002; 300:53–58.
55. Konu O, Li MD. Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. *J Mol Evol* 2002;54:35–41.
56. Graf M, Bojak A, Deml L, Bieler K, Wolf H, Wagner R. Concerted action of multiple *cis*-acting sequences is required for Rev dependence of late human immunodeficiency virus type 1 gene expression. *J Virol* 2000;74:10822–10826.
57. Graf M, Deml L, Wagner R. Codon-optimized genes that enable increased heterologous expression in mammalian cells and elicit efficient immune responses in mice after vaccination of naked DNA. *Methods Mol Med* 2004;94:197–210.
58. Chen CY, Xu N, Shyu AB. mRNA decay mediated by two distinct AU-rich elements from c-fos and granulocyte-macrophage colony-stimulating factor transcripts: different deadenylation kinetics and uncoupling from translation. *Mol Cell Biol* 1995;15: 5777–5788.
59. Chen CY, Shyu AB. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci* 1995;20:465–470.
60. Maldarelli F, Martin MA, 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.
61. Nasioulas G, Zolotukhin AS, Tabernero C, Solomin L, Cunningham CP, Pavlakis GN, Felber BK. 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.
62. Olsen HS, Cochrane AW, Rosen C. Interaction of cellular factors with intragenic *cis*-acting repressive sequences within the HIV genome. *Virology* 1992;191:709–715.
63. Schwartz S, Felber BK, Pavlakis GN. Distinct RNA sequences in the gag region of human immunodeficiency virus type 1 decrease RNA stability and inhibit expression in the absence of Rev protein. *J Virol* 1992;66:150–159.
64. Chang DD, Sharp PA. Regulation by HIV Rev depends upon recognition of splice sites. *Cell* 1989;59:789–795.
65. Kozak M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol* 1987;196:947–950.

66. Kozak M. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* 1987;15:8125–8148.
67. Chapman BS, Thayer RM, Vincent KA, Haigwood NL. Effect of intron A from human cytomegalovirus (Towne) immediate-early gene on heterologous expression in mammalian cells. *Nucleic Acids Res* 1991;19:3979–3986.
68. Trinh R, Gurbaxani B, Morrison SL, Seyfzadeh M. Optimization of codon pair use within the $(GGGGS)_3$ linker sequence results in enhanced protein expression. *Mol Immunol* 2004;40:717–722.
69. Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. *Trends Biotechnol* 2004;22:346–353.
70. Haas J, Park EC, Seed B. Codon usage limitation in the expression of HIV-1 envelope glycoprotein. *Curr Biol* 1996;6:315–324.
71. Uchijima M, Yoshida A, Nagata T, Koide Y. Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I-restricted T cell responses against an intracellular bacterium. *J Immunol* 1998;161:5594–5599.
72. zur Megede J, Chen MC, Doe B, Schaefer M, Greer CE, Selby M, Otten GR, Barnett SW. Increased expression and immunogenicity of sequence-modified human immunodeficiency virus type 1 *gag* gene. *J Virol* 2000;74:2628–2635.
73. Narum DL, Kumar S, Rogers WO, Fuhrmann SR, Liang H, Oakley M, Taye A, Sim BK, Hoffman SL. Codon optimization of gene fragments encoding *Plasmodium falciparum* merozoite proteins enhances DNA vaccine protein expression and immunogenicity in mice. *Infect Immun* 2001;69:7250–7253.
74. Valencik ML, McDonald JA. Codon optimization markedly improves doxycycline regulated gene expression in the mouse heart. *Transgenic Res* 2001;10:269–275.
75. Leder C, Kleinschmidt JA, Wiethe C, Muller M. Enhancement of capsid gene expression: preparing the human papillomavirus type 16 major structural gene L1 for DNA vaccination purposes. *J Virol* 2001;75:9201–9209.
76. Shimshek DR, Kim J, Hubner MR, Spergel DJ, Buchholz F, Casanova E, Stewart AF, Seeburg PH, Sprengel R. Codon-improved Cre recombinase (iCre) expression in the mouse. *Genesis* 2002;32:19–26.
77. Hamdan FF, Mousa A, Ribeiro P. Codon optimization improves heterologous expression of a *Schistosoma mansoni* cDNA in HEK293 cells. *Parasitol Res* 2002;88: 583–586.
78. Gao X, Yo P, Keith A, Ragan TJ, Harris TK. Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Res* 2003;31:e143.
79. Disbrow GL, Sunitha I, Baker CC, Hanover J, Schlegel R. Codon optimization of the HPV-16 E5 gene enhances protein expression. *Virology* 2003;311:105–114.
80. Cid-Arregui A, Juarez V, zur Hausen H. A synthetic E7 gene of human papillomavirus type 16 that yields enhanced expression of the protein in mammalian cells and is useful for DNA immunization studies. *J Virol* 2003;77:4928–4937.
81. Barouch DH, Pau MG, Custers JH, Koudstaal W, Kostense S, Havenga MJ, Truitt DM, Sumida SM, Kishko MG, Arthur JC, Korioth-Schmitz B, Newberg MH, Gorgone DA, Lifton MA, Panicali DL, Nabel GJ, Letvin NL, Goudsmit J. Immunogenicity of recombinant adenovirus serotype 35 vaccine in the presence of pre-existing anti-Ad5 immunity. *J Immunol* 2004;172:6290–6297.

82. Didierlaurent A, Ramirez JC, Gherardi M, Zimmerli SC, Graf M, Orbea HA, Pantaleo G, Wagner R, Esteban M, Kraehenbuhl JP, Sirard JC. Attenuated poxviruses expressing a synthetic HIV protein stimulate HLA-A2-restricted cytotoxic T-cell responses. *Vaccine* 2004;22:3395–3403.

83. Karzenowski D, Potter DW, Padidam M. Inducible control of transgene expression with ecdysone receptor: gene switches with high sensitivity, robust expression, and reduced size. *Biotechniques* 2005;39:191–192, 194, 196.

84. van den Brink EN, Ter Meulen J, Cox F, Jongeneelen MA, Thijsse A, Throsby M, Marissen WE, Rood PM, Bakker AB, Gelderblom HR, Martina BE, Osterhaus AD, Preiser W, Doerr HW, de Kruif J, Goudsmit J. Molecular and biological characterization of human monoclonal antibodies binding to the spike and nucleocapsid proteins of severe acute respiratory syndrome coronavirus. *J Virol* 2005;79:1635–1644.

85. Michelson AM, Todd AR. Synthesis of a dithymidine dinucleotide containing a 3′:5′-internucleotidic linkage. *J Chem Soc* 1955;2632.

86. Agarwal KL, Buchi H, Caruthers MH, Gupta N, Khorana HG, Kleppe K, Kumar A, Ohtsuka E, Rajbhandary UL, Van de Sande JH, Sgaramella V, Weber H, Yamada T. Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature* 1970;227:27–34.

87. Koster H, Blocker H, Frank R, Geussenhainer S, Kaiser W. Total synthesis of a structural gene for the human peptide hormone angiotensin II. *Hoppe Seylers Z Physiol Chem* 1975;356:1585–1593.

88. Itakura K, Hirose T, Crea R, Riggs AD, Heyneker HL, Bolivar F, Boyer HW. Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin. *Science* 1977;198:1056–1063.

89. Edge MD, Green AR, Heathcliffe GR, Meacock PA, Schuch W, Scanlon DB, Atkinson TC, Newton CR, Markham AF. Total synthesis of a human leukocyte interferon gene. *Nature* 1981;292:756–762.

90. Barany F, Gelfand DH. Cloning, overexpression and nucleotide sequence of a thermostable DNA ligase-encoding gene. *Gene* 1991;109:1–11.

91. Young L, Dong Q. Two-step total gene synthesis method. *Nucleic Acids Res* 2004;32:e59.

92. Mandecki W, Hayden MA, Shallcross MA, Stotland E. A totally synthetic plasmid for general cloning, gene expression and mutagenesis in *Escherichia coli*. *Gene* 1990;94:103–107.

93. Stemmer WP, Crameri A, Ha KD, Brennan TM, Heyneker HL. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 1995;164:49–53.

94. Withers-Martinez C, Carpenter EP, Hackett F, Ely B, Sajid M, Grainger M, Blackman MJ. PCR-based gene synthesis as an efficient approach for expression of the A + T-rich malaria genome. *Protein Eng* 1999;12:1113–1120.

95. Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, Church G. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 2004;432:1050–1054.

96. Greger B, Kemper B. An apyrimidinic site kinks DNA and triggers incision by endonuclease VII of phage T4. *Nucleic Acids Res* 1998;26:4432–4438.

97. Smith J, Modrich P. Removal of polymerase-produced mutant sequences from PCR products. *Proc Natl Acad Sci USA* 1997;94:6847–6850.

98. Cello J, Paul AV, Wimmer E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* 2002;297:1016–1018.

99. Kodumal SJ, Patel KG, Reid R, Menzella HG, Welch M, Santi DV. Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc Natl Acad Sci USA* 2004;101:15573–15578.

100. Menzella HG, Reisinger SJ, Welch M, Kealey JT, Kennedy J, Reid R, Tran CQ, Santi DV. Redesign, synthesis and functional expression of the 6-deoxyerythronolide B polyketide synthase gene cluster. *J Ind Microbiol Biotechnol* 2006;33:22–28.

101. Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA, Merryman C, Young L, Noskov VN, Glass JI, Venter JC, Hutchison CA 3rd, Smith HO. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genilalium genome. *Science* 2008; 29:319 (5867): 1215–1220. Epub 2008 Jan.