Molecular Cloning of Protein-Based Polymers

Lixin Mi*,†

Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218

Received March 3, 2005; Revised Manuscript Received May 5, 2006

Protein-based biopolymers have become a promising class of materials for both biomedical and pharmaceutical applications, as they have well-defined molecular weights, monomer compositions, as well as tunable chemical, biological, and mechanical properties. Using standard molecular biology tools, it is possible to design and construct genes encoding artificial proteins or protein-based polymers containing multiple repeats of amino acid sequences. This article reviews some of the traditional methods used for constructing DNA duplexes encoding these repeat-containing genes, including monomer generation, concatemerization, iterative oligomerization, and seamless cloning. A facile and versatile method, called modules of degenerate codons (MDC), which uses PCR and codon degeneracy to overcome some of the disadvantages of traditional methods, is introduced. Re-engineering of the random coil spacer domain of a bioactive protein, WPT2-3R, is used to demonstrate the utility of the MDC method. MDC re-constructed coding sequences facilitate further manipulations, such as insertion, deletion, and swapping of various sequence modules. A summary of some promising emerging techniques for synthesizing repetitive sequence-containing artificial proteins is also provided.

1. Introduction

Over two decades ago, powerful molecular biology and protein engineering tools emerged for the design and synthesis of biologically related protein polymers.[1−7] These standard molecular biology tools continue to be used for producing large amounts of both natural and artificial proteins and protein-based polymers. The precisely regulated transcriptional, translational, and posttranslational machinery of cells ensures a monodisperse final product and, in some cases, allows the production of proteins having well-defined posttranslational modifications. However, these standard tools were not designed for the specialized manipulations one wants to make in large, repeat-containing proteins, such as incorporating or exchanging individual modules in the repeat-containing sequences. (Such manipulations may alter existing properties or add new functional capacities to biopolymers.)

Recent examples of biologically synthesized artificial proteins containing multiple repeats of specific amino acid sequences include biomimetic proteins,[8,9] proteins that self-assemble into novel materials,[10−14] and proteins having potential environmental,[15] pharmaceutical,[16] or biomedical[17−23] applications. The biosynthesis of any artificial protein generally includes (1) constructing a synthetic gene encoding the protein of interest in a plasmid with tight transcription control; (2) transforming circularized vectors into cloning competent cells; (3) screening plasmid containing cells for ones containing the desired clones and verifying the DNA sequence; (4) transforming the chosen plasmids into expression competent host cells; (5) growing appropriate volumes of host cells and inducing protein expression; (6) purifying the protein of interest from cell lysates. The construction of stable artificial genes encoding repetitive amino acid sequences and their expression pose special problems that arise from several sources, including protein design (codon usage) and assembling a large coding sequence from multiple smaller fragments. These special problems create bottlenecks affecting whole projects, particularly the quality and yield of the final protein product.

Devising an appropriate strategy at the nucleotide level is critical for the efficient synthesis of the protein encoding sequence and for producing a uniform protein product. Most traditional methods of gene construction involve primarily concatemerization and iterative linkage of DNA sequences. Recently, seamless cloning has emerged as a promising method. However, it also has technical limitations. This review, on one hand, provides a systematic analysis of traditional and currently employed strategies for generating repetitive sequence proteins in terms of their advantages and disadvantages. This review will also highlight a facile and versatile strategy that combines the use of codon degeneracy and polymerase chain reaction (PCR) to synthesize artificial proteins with multiple identical domains containing repetitive sequences.[23] This review concludes with a brief discussion of a few novel and emerging strategies, including PCR seamless cloning and a recombination method that uses two recombination proteins (called "ET recombination" because it employs the RecE and RecT proteins) that appear capable of generating genes encoding large, artificial, repeat-sequence-containing proteins that can be efficiently expressed in bacterial cells

2. Traditional and Current Methods

2.1. Pioneering Works. In probably the first paper describing the synthesis of DNA encoding a peptide with sequence repeats and its expression, Doel et al.[1] cloned a sequence coding 150 repeats of the dipeptide aspartyl-phenylalanine (asp-phe). Their duplex DNA was constructed through consecutive oligonucleotide hybridizations. Briefly, as illustrated in Figure 1A, two-half complementary and 5′ phosphorylated oligonucleotides, 5′ *TCGAAA***TCGAAG** and 5′ *TTTCGA***CTTCGA** (italic and bold indicate complementary sequences) were mixed at a stoichio-

* To whom correspondence should be addressed.
† Present address: Lombardi Comprehensive Cancer Center, Georgetown University, 3800 Reservoir Road, Washington, DC 20057.

10.1021/bm050158h CCC: $33.50
Published on Web 06/09/2006

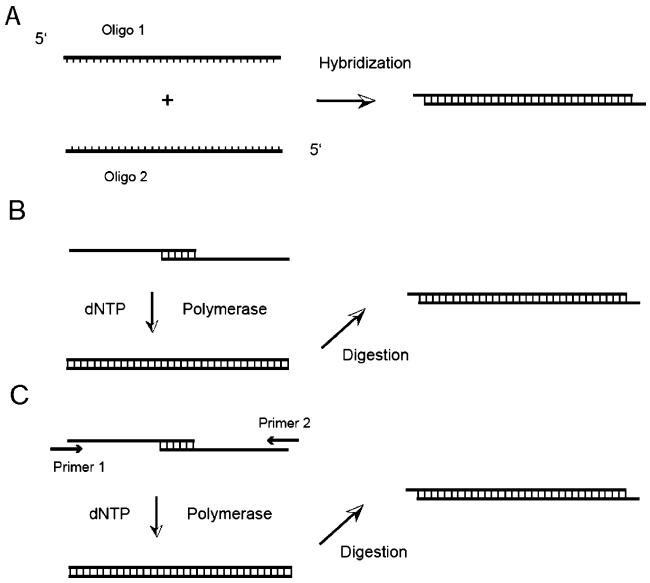

Figure 1. Monomeric sequence generation. (A) By hybridization; (B) by hybridization of partially complementary sequences followed by dNTP filling-in and restriction enzyme digestion; (C) a hybrid formed by partially complementary sequences can be used as a PCR template. Monomers with cohesive ends are created by restriction enzyme digestion of PCR products.

metric 1:1 ratio and annealed. The short six-base overlap allowed hybridization at room temperature. The 5′ phosphorylation of these synthetic oilgonucleotides was accomplished by T4 polynucleotide kinase. The nicks between juxtaposed 3′ hydroxyl termini and 5′ phosphates on both strands were closed by T4 DNA ligase.

A similar gene-fragment-assembly-through-hybridization approach was used by DeGrado's group[2] to construct a duplex encoding a tetrahelical protein (α_4) which adopted desired three-dimensional structures. The duplex encoding these four identical helices alternating with three identical loops was assembled by hybridization from eight partially complementary oligonucleotides. To avoid hybridizations producing duplexes containing mismatches between certain pairs of these eight oligonucleotides, codon degeneracy was used to design different base sequences to encode identical amino acid sequences. Their thoughtful design included several unique restriction sites embedded at convenient locations in the duplex to facilitate future cassette mutagenesis.

As the above two cases demonstrate, hybridization or annealing of partially complementary strands can be used to directly construct coding sequences; however, this strategy has limitations when applied to large proteins containing repeating sequences. For example, it is possible that single strand DNAs anneal nonspecifically when two many strands are in the mixture or sequences are similar. Also it is laborious to create a family of proteins with minor variations in amino acid sequence using this method. In addition, adapters or linkers with appropriate cohesive ends (encoding unwanted amino acids) are often required to incorporate the assembled sequences into expression plasmids. In this review, any double stranded coding sequence generated by a conventional hybridization method will be called monomeric, even if the sequence encodes repeating sequences of amino acids.

2.2. Monomeric Sequence Synthesis (Figure 1). Early work by Cappello et al.[3] and McGrath et al.[5] used the hybridization approach illustrated in Figure 1A. The resulting duplex DNA, which may encode only the desired amino acids, is flanked on both ends by appropriate restriction sites. Such duplexes can be either directly cloned into expression plasmids or can be manipulated in various ways to form oligomers encoding very large numbers of repeating units (see below).

A variation of the hybridization method for constructing a monomer encoding repeating sequences was devised by Fournier's group[25] (Figure 1B). They wanted to generate and express genes encoding spider dragline silk protein mimics containing different numbers of contiguous repeats of a 35-amino acid sequence. They first generated a monomer-a 116-base duplex DNA by (1) hybridizing two long oligonucleotides, a 84-base oligonucleotide and a 59-base oligonucleotide that contained 27-base complementary regions at their 3′ ends; (2) using Vent polymerase to make a full-length blunt-ended duplex; (3) digesting this duplex with *Xma* I (C∧CCGGG) to create cohesive ends for ligation into a plasmid. This approach solved the technical limitation (at that time) on the length of individual oligonucleotides that could be reliably synthesized.

Kostal et al.[15] employed the same approach by hybridizing a 94-base oligonucleotide with a 115-base oligonucleotide having 20-base complementary regions at their 3′ ends and then extending the duplex region by filling in with the high fidelity *pfu* polymerase. The resulting 189 bp duplex was digested by *Eco*R I (G∧AATTC) and *Bam*H I (G∧GATCC) to generate a monomer encoding an elastin-like sequence, $(VPGVG)_{10}-H_6$ that was cloned into an expression plasmid.

PCR was employed for monomer synthesis by McPherson et al.[8] (Figure 1C). Their goal was to construct a gene encoding 10 repeating units of the elastomeric pentapeptide VPGVG. Briefly, two 85-base oligonucleotides hybridized by a 20-base 3′ complimentary region were used as the template for AMV reverse transcriptase to fill-in complementary bases, generating a 150 bp duplex. This duplex DNA then served as a PCR template. The amplified blunt-end products were digested with *Xmn* I and *Eco*R I before ligating with an appropriate vector. It should be pointed out that the reverse transcription step was not required, as the hybridized oilgonucleotides can themselves be used as a PCR template.

Combining hybridization and PCR with restrictive enzyme digestion (Figure 1C) is an easy and powerful technique to generate duplex DNA monomers. Theoretically, this method can produce 2^n copies of the monomer precursors, a much greater yield than achieved by hybridizing chemically synthesized oligonucleotides, with (Figure 1B) or without (Figure 1A) complementary base fill-in (Figure 1B) (n is the thermal cycle number). As will be mentioned later, having substantial amounts of monomers is necessary for linking multiple monomers together (concatemerization). However, one downside of any approach using PCR is that it is error-prone.

To produce a large pool of monomers having higher sequence fidelity than can be achieved by PCR, monomer amplification can be carried out in vivo. Cappello et al.[3] generated small amounts of two monomer gene fragments, one encoding a silk-like peptide unit and the other a silk-elastin-like peptide unit, by hybridization. The monomers were individually ligated to linearized plasmid DNA at compatible restriction sites and then transformed into competent *E. coli* cells. Drug-resistant clones were screened for the presence of the desired fragment by restriction analysis and sequence verification. The substantial amounts of monomer needed for concatemerization were readily generated by restriction digestion of plasmids harvested from cells. Although cloning individual monomeric fragments is time-consuming, this approach has advantages in addition to the error-free production of sufficient quantities of monomers: (1) it provides an opportunity to check the monomeric sequence for

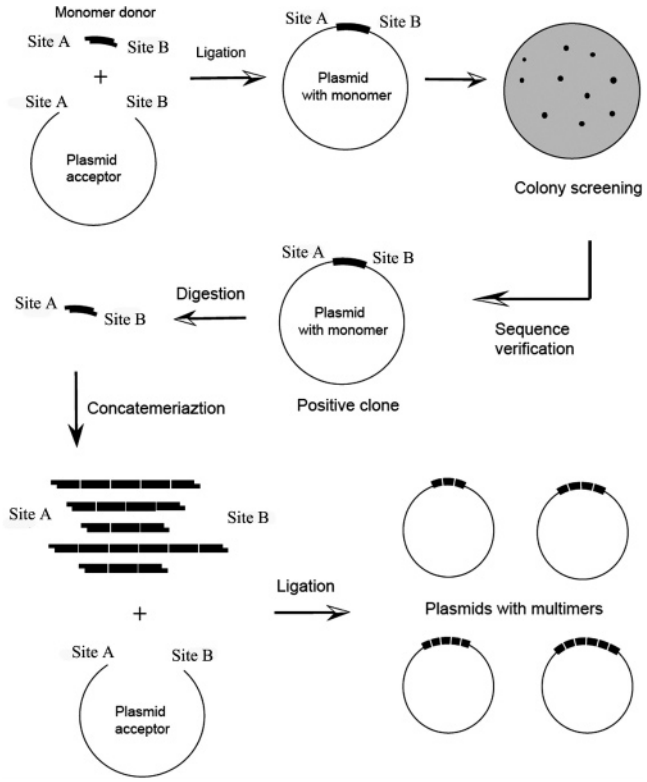

Figure 2. Illustration of concatemerization of monomers obtained by digesting plasmids containing a single monomer.

errors before oligomerization; (2) it enables further manipulating of the sequence, including generating block copolymer or site-directed mutagenesis, before proceeding to concatemerization; and (3) it preserves the phosphorylated 5′ overhangs of the monomers and avoids the time-consuming purification steps necessary when either the hybridization-only approach (Figure 1A)[5] or the base-fill-in approach (Figure 1B) is employed. As shown by Cappello and co-workers,[3] adequate amounts of several distinct proteins can be obtained by expressing block copolymers, encoding various stoichiometric ratios of silk-like blocks and elastin-like blocks, to demonstrate that these related proteins possess a range of extraordinary mechanical properties.

2.2. Concatemerization (Figure 2). The most popular and easiest method for making either homoblock or heteroblock copolymers involves concatemerization, the random ligation of monomeric fragments to one another. This approach will result in unidirectional head-to-tail linear oligomers if the single stranded regions at the two ends of each monomer are of the same lengths but different from one another and cohesive; that is, where the single stranded head of one fragment can hybridize only to a single stranded tail (of itself or another fragment). In other words, the sequences of the unmatched single strands at opposite ends of each fragment cannot be palindromic. Using concatemerization, Cappello et al.[3] were able to construct an artificial gene encoding a silk-like protein with up to 160 repeats of a six amino acid sequence (GAGAGS). The encoded protein was designed to adopt a crystalline β sheet structure typical of silk fibroin. With this approach, the distribution of concatemeric lengths is (partially) controlled by varying the monomer fragment concentration, their purity, and the ligation time.

Similarly, McGrath et al.[5] used concatemerization to produce a gene encoding 12 repeats of a nine amino acid sequence, $[(AG)_3PEG]_{12}$. Briefly, nonpalindromic cohesive ends on the repeating units, 5′-CCGGAA− on the sense strand and 5′-TTCCGG− on the antisense strand, were used to link the individual units unidirectionally (head-to-tail) into concatemers using T4 DNA ligase. Subsequently, translational start and stop codon linkers were attached to the concatemer termini. Finally, the resulting linear fragment was inserted into an expression vector by digesting with *Bam*H I and *Eco*R I. In later work,[6] this group employed the same strategy to construct genes encoding 5, 9, 14, 27, and 54 repeats of the $(AG)_3PEG$ sequence.

The nonpalindromic cohesive ends used to ensure head-to-tail ligation of monomers can be created by either hybridization or endonuclease digestion. For example, the six-base *Ban* I site (G∧GTGCC) has been extensively used.[6,12,17,26,27] In these cases, the 5′-GTGC− flanking bases prohibit head-to-head or tail-to-tail ligation. Head-to-tail ligation can also be assured by using an interrupted palindrome; for example, digesting a *Pfl*M I site (CCANNNN∧NTGG) was employed by McPherson et al[28] to synthesize a series of elastin-like protein polymers.

In general, concatemerization is excellent for making, in a single ligation, a library of linear fragments having various numbers of identical monomers. As mentioned, the length distribution is highly dependent on the concentration and purity of the monomers as well as the ligation conditions. However, a significant disadvantage of this method is that there is no way to control the linear sequence of individual monomers, a situation that arises when one wants to vary codon usage within monomer units. Therefore, this is not a good method for generating a gene fragment encoding repetitive amino acid sequences where the codon usage in individual monomers is precisely predetermined.

2.3. Iterative and Recursive Method (Figure 3). Iterative or recursive methods, a stepwise approach for generating oligomers from monomeric sequences, is able to achieve control over both the order of addition (of monomers or well-defined multimers) to a growing oligomer and the total number of monomers inserted. For example, Prince et al.[9] described a multistep strategy using plasmids and inserts containing two compatible restriction sites, *Nhe* I (G∧CTAGC) and *Spe* I (A∧CTAGT), to construct head-to-tail tandem repeats of a DNA sequence (see Figure 3A). To construct a gene encoding spider silk protein, they first generated a monomer, a 114-base pair DNA duplex having a four base 5′-CTAG− overhang on each end (and either an A or G 5′ to the CTAG overhang), by annealing four overlapping shorter single stranded oligonucleotides. This DNA duplex was 5′ phosphorylated before being ligated into a plasmid previously cut at unique *Nhe* I and *Spe* I sites (and dephosphorylated) and then transformed into cells, producing multiple clones. Most of these clones contained inserts because the vector dephosphorylation step reduces plasmid self-ligation, thus increasing insertion efficiency. Although this duplex could ligate in either orientation into the *Nhe* I and *Spe* I cut plasmid, clones having the desired orientation were identified by cutting with each of these enzymes (the desired orientation could be cut by both enzymes). They then switched to the iterative and recursive part of their strategy by opening this monomer-insert plasmid at the (upstream) *Nhe* I site and ligating this plasmid with a library of linear head-to-tail concatemers. Several head-to-tail libraries were produced by using different insert-to-vector ratios, generating various distributions of oligomers. All oligomers containing any head-to-head or tail-to tail monomers were eliminated from these libraries by digesting the ligation products with both *Nhe* I and *Spe* I. Subsequently, Winkler et al.[29] modified this method (upstream insertion at the *Nhe* I site) by doing the iterative insertion of oligomers in the downstream *Spe* I site, allowing

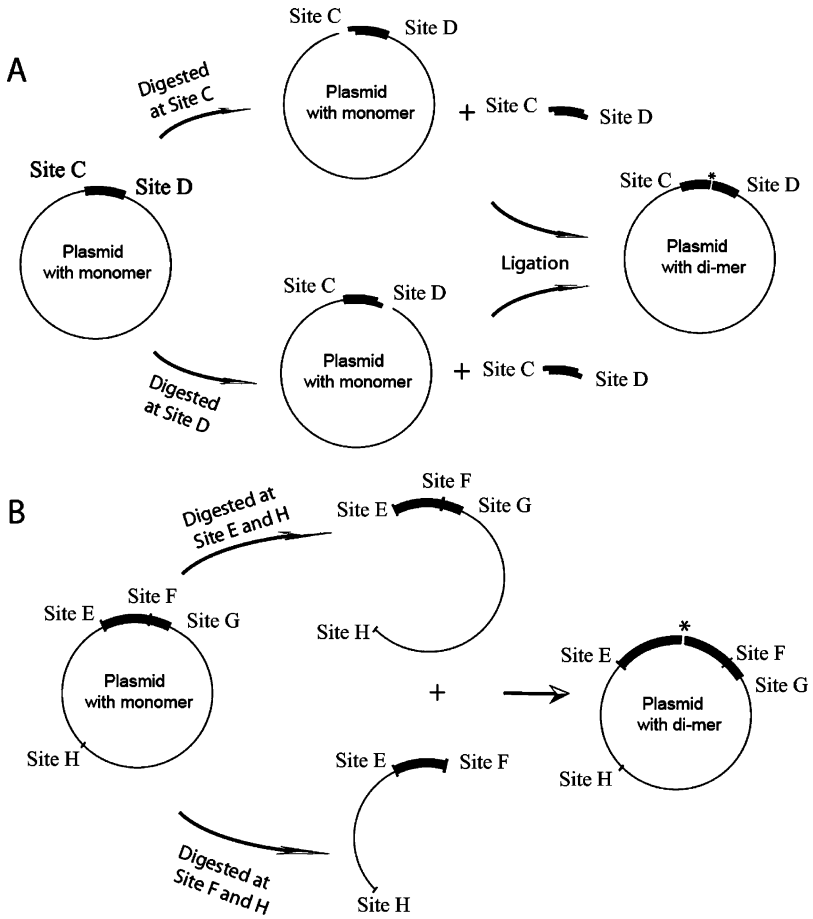

Figure 3. Iterative or recursive methods. (A) Front and back insertion (relative the orientation of the coding strand of the inserted fragment) with the help of two compatible but nonregenerate sites; (B) Uni-directional insertion with the help of a third restriction site. (Site E and G were used to clone a monomer. Site E and F are two compatible but nonregenerate sites. Site H is the third site overhanging bases are different from those generated by digestion at sites E and F.) Sign * denotes a hybrid and nonregenrated site.

them to incorporate both phosphate and redox triggers into the encoded spider silk proteins. This back-insertion approach was also used by Petka *et* al.[10,11] to construct modular block copolymeric proteins.

The iterative strategy involves several parameters: (a) two different enzymes are required to digest the plasmid and the self-ligated inserts; (b) initially the vector is cut by both enzymes for monomer insertion but each subsequent insertion requires a cut by only one or the other enzyme; (c) most pairs of restriction sites yielding compatible ends can be used, i.e., the cohesive end created by one can be ligated to that of the other; (d) when the oligomeric head-to-tail insert is in place, all of the merged *Nhe* I and *Spe* I sites, which are internal to the oligomer, will not be regenerated (no longer be recognizable by either enzyme) while the two remaining *Nhe* I and *Spe* I sites (at the ends of the inserted oligomer) are preserved for another round of insertion; and (e) only two extraneous amino acid residues are introduced per ligation joint. However, this approach does not ensure unidirectional insertion of either the first monomer or, more importantly, unidirectional insertion of subsequent insertions because all of the ends are compatible with one another. Fortunately, several strategies to overcome this problem have been devised.

The strategy adopted by Lewis et al.[25] to solve this problem was to get help from a third, incompatible restriction site on the plasmid (Figure 3B). The monomer sequence was designed to have an *Xma* I site (C∧CCGGG) at one end (site E), an internal *Bsp*E I (T∧CCGGA) site (site F), and an *Eco*RV (GAT∧ATC) site at the other end (site G). This monomer was cloned into *pBluescript* in a fixed orientation following digestion of the plasmid by *Xma* I (giving a cohesive end) and *Eco*RV (giving a blunt-end). Unidirectional insertion of a second monomer was achieved by taking advantage of a unique *Sca* I site (AGT∧ACT) in the coding sequence of the β-lactamase gene in *pBluescript*: two aliquots of the plasmid with the inserted monomer were double-digested, one with *Xma* I/*Sca* I and the other with *Bsp*E I/*Sca* I. Since both *Xma* I and *Bsp*E I create 5′-CCGG overhangs, ligation of these two sites together generates a site that cannot be digested with either enzyme, generating a plasmid containing two direct repeats of the monomer and which retains all of the features needed to repeat this insertion process (the resulting plasmid still has unique *Sca* I, *Xma* I, and *Bsp*E I sites). Not only does this approach ensure uni-directional insertion of monomers, this strategy is highly efficiency because vectors that re-circularize without the desired insert will not allow cells to survive ampicillin selection (the *Sca* I site is embedded in the β-lactamase coding sequence and cloning other fragments into this site will cause insertional inactivation).

Another way to overcome the insertion-orientation problem is to use interrupted palindromes. Recently, a new approach,[16,30] called "recursive directional ligation" (RDL), has been developed to produce elastin-like polypentapeptides with repeating VPGXG, where X is a "guest residue" X, Val, Ala, or Gly in a 5:2:3 ratio. This approach uses a pair of interrupted palindromes *Pfl*M I (CCANNNN∧NTGG) and *Bgl* I (GCCNNNN∧NGGC) as the front and back sites for oligomeric assembly. More specifically, the internal non-

specified sites in each palindrome were designed to ensure unidirectional ligation by making the three base overhangs generated by enzyme digestion to be compatible with one another but not palindromic. Thus, the *Pfl*M I site sequence was designed as CCAC*GGC*∧GTGG and the *Bgl* I sites designed as GCCG*GGC*∧GGGC. Since the enzymatic recognition sequence does not overlap with the cleavage site, the three bases 5′ to the cleavage sites (*GGC*, italicized) were designed to be compatible (allowing ligation) but nonpalindromic (forcing directionality). An additional potential benefit of inserting coding sequences into interrupted palindromic sites is that they allow flexibility in amino acid codon choices, possibly decreasing the number of extraneous amino acids at the connection joints. The authors[30] of this approach identified other interrupted palindromes that can be used in the RDL method because these sites also yield three-base, unconstrained 3′ overhangs upon cleavage (*Alw*N I, *Bsl* I, *Bst*AP I, *Dra* III, *Mwo* I, *Sfi* I, and others).

Although the use of interrupted palindromic restriction sites is a significant improvement over six-base palindromic sites, such as *Nhe* I and *Spe* I, this approach leaves a few minor problems unaddressed, including the risks associated with vector dephosphorylation and irreversibly damaged ligation sites, rendering such sites unusable for modular recombination or shuffling.

To overcome the vector self-ligation problem resulting from using pairs of compatible restriction sites, such as *Nhe* I/*Spe* I and *Pfl*M I/*Bgl* I, Kostal et al.[15] used two noncompatible restriction sites, *Sma* I (CCC∧GGG) and *Bam*H I, to elongate an elastin-based peptide unit sequence, $(VPGVG)_{10}$. In this case, the monomer unit, encoding $(VPGVG)_{10}-H_6$, was first amplified by PCR and then digested to create a *Bam*H I site on one end while keeping the other blunt. The monomer insert was ligated to a vector cut at *Sma* I and *Bam*H I sites, allowing retention of an H_6 tag on the protein construct (the H_6 tag was encoded in the *Sma* I-*Bam*H I region of the vector). The combination of a blunt-end and a cohesive end on the vector ensures not only uni-directional insertion of a monomer but also highly efficient insertion because vector self-ligation is rare. The method is also unique since it saves a digestion on one end of the insert by utilizing the *Sma* I compatible blunt-end by exonuclease activity of *pfu* polymerase during PCR.

Compared with concatemerization, which generates a distribution of insert numbers, the iterative method controls the monomeric repeat number and is able to create block copolymers with designated sequences. However, iteration can be laborious and time-consuming, especially when one wants a synthetic gene with a high repeat number of monomers.

2.4. Seamless Cloning. Uni-directional ligation can also be achieved using a technique called seamless cloning. Unlike earlier methods, seamless cloning,[31] which utilizes type IIs restriction endonucleases, permits cleavage of any DNA sequence that is at a defined distance from a recognition site. This approach has several merits: (1) it allows one to avoid inserting extraneous DNA at the ligation joints and thus can eliminate extraneous amino acid residues within the desired protein sequence; (2) it eliminates sequence constraints on the DNA fragments to be ligated, since any given DNA sequence can be inserted into any desired location without any need for an appropriate restriction site; and (3) it also ensures uni-directional insertion due to the fact that the flanking bases can be designed to be nonpalindromic. This approach liberates researchers from a limited pool of known nonpalindromic restriction endonucleases.

The earliest application of seamless cloning to the biopolymer field, McMillan et al.,[32] utilized three features of the type IIs restriction *Eam*1104 I site (CTCTTC∧1/4; where ∧1/4 indicates that the restriction site is one and four bases downstream on the sense and antisense strands, respectively): (1) the cleavage takes place only to one side of the recognition site; (2) the six-base recognition site is one base away from its cleavage site; and (3) cleavage is inhibited by 5′-methylation of a cytosine in the recognition sequence.

They first generated clones encoding two different monomers $[(VPGVG)_4VPGKG]$ and $[(VPGVG)_4VPGIG]$. These monomers then were concatemerized with the help of the nonpalindromic flanking bases created by *Eam*1104I. The recipient vector was prepared by inverse PCR[33] to ensure that the only sites digested by *Eam*1104I are the desired future ligation sites. This was accomplished by using 5′-methyldeoxycytosine (^{5m}dCTP) instead of deoxycytosine (dCTP) as a PCR substrate and by using primers containing only nonmethylated *Eam*1104 I sites. Their design also ensured that the two *Eam*1104 I recognition sites on the linearized vector were chopped off so no extraneous coding bases were added. Finally, they were able to ligate concatemers with the vector to synthesize a gene encoding a protein with over 1000 amino acids. Nevertheless, these authors, although succeeding in producing a 90 kDa protein polymer, encountered some technical challenges. For example, the cloning vector was fragmented due to incomplete incorporation of ^{5m}dCTP.

Recently, Goeden-Wood et al. [34] improved the seamless cloning approach by making two modifications; first, they removed all internal *Eam1104* I sites from the expression vector, pET-19b, by site-directed mutagenesis and by exploiting the less frequently cutting type IIs restriction enzyme, *Sap* I (GCTCTTC∧1/4), whose recognition sequence is longer but nearly identical to *Eam*1104 I.

Sap I also cleaves to the same side of and at the same distances away from its recognition site as *Eam1104* I. Thus, more efficient cloning was achieved by using or engineering unique *Sap* I sites in vectors and by avoiding troublesome incomplete incorporation of ^{5m}dCTP during inverse PCR of the recipient vector.

For constructing genes encoding repetitive sequences, modified seamless cloning, especially using unique *Sap* I sites, is one of the best approaches to date. Type IIs restriction endonucleases are highly favored in the protein polymer field because they guarantee unidirectional ligation and no extraneous sequences. However, technical disadvantages remain: (1) the vector plasmid has to be prepared by inverse PCR, which is error-prone, even when high fidelity polymerases are used, because the products are usually 4−5 kilobases long; (2) site-directed mutagenesis is usually required to remove internal type IIs sites from the cloning vectors to avoid unexpected digestion; and (3) the method is limited to being an extension of concatemerization approach.

2.5. Technical Limitations. In summery, artificial genes encoding multiple repeats of amino acid sequences can be constructed by (a) direct hybridization of oligonucleotides; (b) concatemerization or "ex plasmid" assembly of monomeric sequences; and (c) iteration or "in plasmid" sequential assembly by repeated insertion of monomers. Although widely used for constructing genes encoding repetitive sequence proteins, the reviewed approaches have several inherent limitations:

1. The concatemerization method is limited to the construction of genes where the encoding sequences in all of the monomers

are identical and where the repeat numbers may not be precisely known, especially if the repeat number is high.

2. The iterative approach can be a lengthy and time-consuming method; it usually takes multiple cycles to construct a gene with a high repeat-number.

3. There are no unique sequences embedded within the final repetitive DNA sequence that can be used to manipulate predetermined groups of the encoded sequences (domain and module recombination or shuffling) after the gene has been constructed.

4. Similarly, once generated, it is usually very difficult to mutate an individual base or amino acid in only one (predetermined) monomer or to make insertions or deletions in only one particular monomer.

5. Since the DNA sequence is highly repetitive, it is a challenge to amplify in vitro by PCR. That is, without "check points" or primer anchorage sequences (unique potential primer sequences embedded in genes longer than 1 kilobase), it is almost impossible to verify gene integrity by current DNA sequencing methods.

6. The highly biased codon usage resulting from repetitive identical monomer sequences may result in protein yields to low for biophysical characterization.

Many of the potential problems posed by current construction methods could be avoided by using an "ex plasmid" method (MDC) that assembled monomers encoding identical amino acid sequences if these monomers used different combinations of degenerate codons.

3. Modules of Degenerate Codons (MDC)

Codon degeneracy is widely exploited in molecular biology, protein design, and protein engineering. For example, in DeGrado's early work[2] on de novo design of a tetrahelix protein, degenerate codons were scrambled to avoid using identical nucleotide sequences for encoding four identical helices. Since their gene was constructed by hybridizing multiple oligonucleotides, a major rationale for their codon scrambling was to improve hybridization specificity. Although codon scrambling is clearly a powerful method for protein engineering and has a long history, it has not been widely employed by the protein biomaterials community in recent years.

Another powerful technique, PCR, is routinely used to amplify duplex DNA, and therefore, its use in artificial gene construction should be facile and efficient. However, PCR amplification of DNA containing multiple exactly repeating sequences is difficult because the process can completely fail or produce heterogeneous final products. Such problems are mainly due to nonspecific annealing and may increase if the sequence has a high GC content, which favors secondary structure formation. These problems can be minimized by reducing the number of repetitive sequence patterns in the template, thereby reducing nonspecifically annealed PCR substrates.

In contrast to the techniques reviewed, the codon scrambling strategy for repetitive sequences take advantage of the degenerate relationship between codons and amino acids. The use of diverse base sequences to encode identical amino acid sequences allows facile PCR-based construction of synthetic genes containing multiple repeated amino acid sequences. In addition, this degeneracy allows one to attach unique restriction enzymatic sites to many or even all of the different base sequence combinations (modules) used to encode a given amino acid sequence. With the help of these embedded restriction sites, a DNA fragment encoding a defined number of repeat units can be easily pre-assembled before insertion into an expression plasmid. Using this strategy, one is able to (1) eliminate repetitive base sequence patterns, increasing PCR productivity and fidelity; (2) efficiently synthesize proteins with a designated number of either homogeneous and heterogeneous amino acid repeats; (3) easily swap and combine different, predetermined modules; (4) easily mutate, insert, and delete one or a few domains in almost any module at almost any location in the synthetic protein; and (5) obtain high protein yields, presumably as a result of a more balanced, physiological codon usage.[23]

This approach was used to re-engineer an artificial gene for a well-established triblock protein design (WPT2).[10,11] This triblock design motif encodes a repetitive, soluble, disordered coil consisting of 10 repeats of the $(AG)_3$PEG sequence, flanked on each side by a leucine zipper sequence. Such triblocks proteins were previously shown to spontaneously assemble in solution into reversible hydrogel networks.[10,11,13] Here, a brief summary is given of the goals and approach used to assemble a triblock protein having one acidic and one basic leucine zipper end,[10,11] which forms robust hydrogels stabilized by favorable electrostatic interactions between the acid and basic leucine zipper domains. The goals were to (1) produce repeating amino acid sequences using scrambled codon principles; (2) incorporate three tri-peptide (RGD) sequences into a predetermined site within the predicted disordered coil of the WPT2 protein design; and (3) test this new approach for assembling the (modified) repeat units. The detailed molecular, materials, and biofunctional characterizations of the modified triblock protein produced have been reported elsewhere.[23]

The original described disordered coil $[(AG)_3PEG]_{10}$ (called C_{10}) is highly repetitive in both amino acid sequence and DNA coding sequence, consisting of 10 repeats of a coding sequence. Since exact sequence repeats were used in the original construction, there is no restriction site unique to any of the repeating DNA sequences that can be manipulated with a simple cut-and-paste approach, for example, to incorporate a new domain or to make point mutations. Our modification of the disordered coil design consists of a central region with three repeats of the RGD tri-peptide sequence dispersed within three repeats of the $(AG)_3$PEG peptide sequence (denoted as R_3) flanked on each end by three repeats of $(AG)_3$PEG (denoted as C_3). To facilitate future modifications, five unique restriction digestion sites within this module (Figure 4) were included. These restriction digestion sites also act as unique PCR priming sites for the purposes of both gene fragment amplification and DNA sequencing. Unlike the original gene for the C_{10} disordered coil, the gene fragment encoding $C_3R_3C_3$ was assembled from nonrepetitive DNA sequences.

Since the length limit of commercial oligonucleotide synthesis is about 110 bases, the sequence encoding C_3 R_3C_3 was broken into four pieces to be linked together by restriction sites (the oligonucleotides used as PCR templates and primers are listed in Table 1). As shown in Figure 4, the end modules (C3−1 and C3−2) are, after PCR and restriction enzyme digestion, terminated by unique ligation-competent restriction sites (*Kpn* I and *Eag* I for C3−1 and *Asc* I and *Sph* I for C3−2). The middle module R_3 (with the repeated RGD sequences) is divided into two submodules: R3−1 (terminated by *Eag* I and *Sac* I) and R3−2 (terminated by *Sac* I and *Asc* I). This combination of restriction sites was chosen because (1) they are unique in both recognition sequence and overhang sequence, guaranteeing unidirectional ligation and (2) they are efficiently digested, even when double digestions are performed. These four DNA

Table 1. Sequences of PCR Templates and Primers for $C_3R_3C_3$ Fragment Construction

name	oligonucleotide sequence (from 5′ to 3′)
C3−1	AGGGGTACCCCACCAGGTGCGGGTGCAGGAGCCGGTCCTGAGGGAGCAGGCGC
	TGGTGCAGGGCCAGAAGGTGCTGGTGCAGGCGCAGGTCCGGAAGGGGCAGTTC
	GGCCGATGGCTAGTGTC
primer 1	AGGGGTACCCCACCAGGTGCG
primer 2	GACACTAGCCATCGGCCGAACTGC
R-1	GCTGTTCGGCCGATGGCTAGTGTCGATCGCGGCGATTCAGGTGCCGGAGCAGGG
	GCGGGCCCCGAGGGCCGGGGTGACAGCGGCAGCGAGCTCGCCGGA
primer 3	GCTGTTCGGCCGATGGCTAGTGTCG
primer 4	TCCGGCGAGCTCGCTGCCG
R-2	GCTAGCGAGCTCGCGGGGGCTGGAGCCGGACCTGAAGGACGCGGAGATTCTGGA
	GCAGGTGCGGGTGCAGGTCCAGAGGGTGGCTCGGCGCGCCAA
primer 5	GCTAGCGAGCTCGCGGGGG
primer 6	TTGGCGCGCCGAGCCACCC
C3−2	TTGGCGCGCCAAGGAGCTGGCGCCGGCGCAGGGCCGGAAGGGGCCGGTGCAGGG
	GCGGGCCCCGAGGGCGCGGGAGCTGGAGCCGGTCCAGAAGGTAGATCTTCG
primer 7	TTGGCGCGCCAAGGAGCTGGC
primer 8	CGAAGATCTACCTTCTGGACCGGCTC

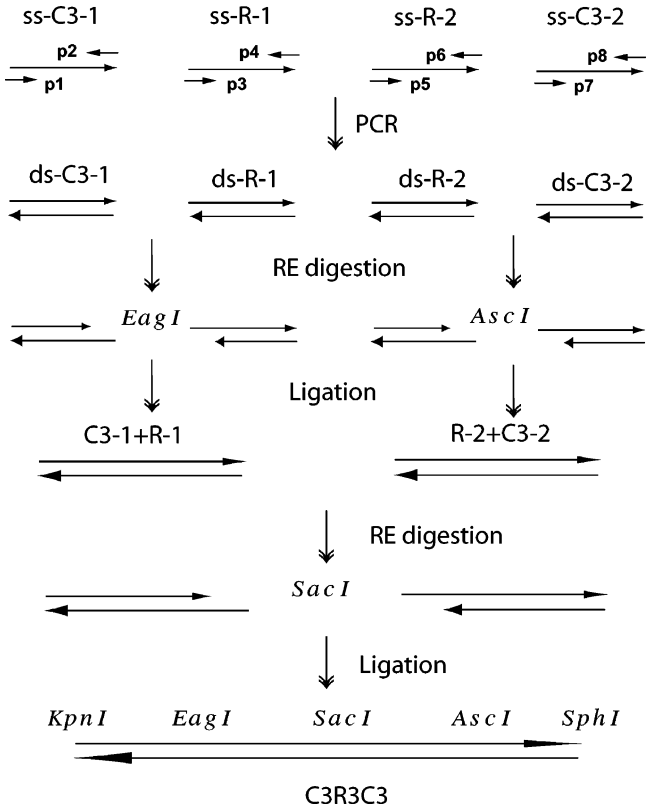

Figure 4. Module fragment assembly scheme using scrambled degenerate codons for each section (top line) used to construct the $C_3R_3C_3$ coding sequence. Each of the p1−p8 primer sequences (top line) are unique and are listed in Table 1.

fragments were assembled into C3R3C3 by ligation after only two rounds of PCR followed by restriction digestion. The $C_3R_3C_3$ fragment encoding a bioactive disordered coil domain was subsequently ligated to fragments encoding the acidic and basic leucine zipper domains of WPT2 at the unique *Kpn* I and *Sph* I sites, respectively. The final modified triblock protein was called WPT2−3R. The protein encoded in this synthetic gene was produced from the T5 *trans*-regulated pQE−9 expression vector transformed into the SG13009 *E. coli* strain. To optimize production of the repetitive amino acid blocks in this $C_3R_3C_3$ protein, a maximally diverse selection of codons were used for each amino acid residue in the blocks.

Large scale growth of WPT2−3R protein producing cells was done in a batch-mode fermentor, and the protein was purified from cell lysates using metal-ion affinity chromatography. MALDI-TOF and amino acid analysis confirmed the identity of the purified protein. A detailed description of the production and biofunctional characterization of the purified protein is presented elsewhere.[23] Here it suffices to note that WPT2−3R self-assembles reversibly into hydrogels having very similar mechanical properties to the unmodified WPT2 protein studied previously.[10,11] In addition, the RGD tripeptide repeats inserted into the spacer block endow this protein with the predicted additional mechanical properties and some unanticipated but interesting biological effects. They primarily function as ligands, giving this protein the ability to stably interact with integrin receptors on HFF cells. In addition, cell−cell adhesion, cell polarization, and the development of F-actin containing stress fibers and associated focal adhesion-like complexes localized to these newly created cell−cell contact sites have been observed. The successful application of the modular codon scrambling strategy (MCS) for re-engineering a biofunctional protein and modifying this protein to give it predictable mechanical properties having significant biological effects demonstrates the effectiveness and potential utility of this method. This method is also technically easy.

However, since this method requires restriction endonuclease sites, although not necessarily nonpalindromic sites for module linkage, the inclusion of these sites unavoidably introduces extraneous amino acids into the synthetic proteins. Fortunately, no effects assignable to the extraneous amino acids were detected in either the structural studies or functional assays with this particular protein.[23] Nevertheless, in cases where this feature causes problems, a restriction site-free method, using overlapping PCR,[36] might be developed as outlined in the following section.

4. Perspective for Future Methods

Of the 3681 biochemically characterized restriction enzymes listed in REBASE,[35] over 98% are Type II enzymes, which recognize specific double-strand DNA sequences and cleave within or close to these sites and do not require adenosine triphosphate (ATP) for their nucleolytic activity. The commercialization of nearly 600 restriction enzymes (having almost as many specificities) makes digestion with type II enzymes

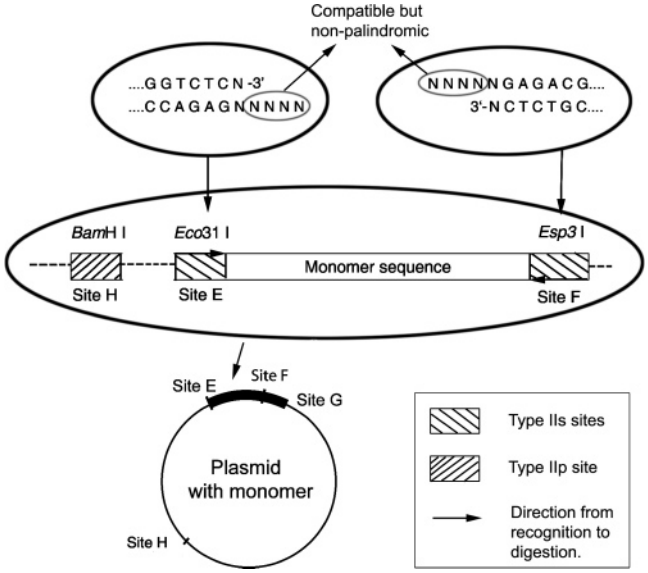

Figure 5. Illustration of restriction site manipulation for the upgraded seamless cloning method.

the preferred way for creating sites for controlling the order and directionality of duplex ligation. Other commercially available enzymes, including those which generate compatible but nonregenerate (the merged site will not be recognized by any enzymes) interrupted palindromic overhangs and type IIs enzymes, can be used for quick and efficient cloning of sequences encoding repeating amino acid sequences. For example, the pool of commercially available type IIs enzymes is getting bigger and now includes enzymes with at least six bases in their recognition sites and which cut at least once within the nearest four flanking bases, includes *Aar* I (CACCTGC∧4/8), *Acc*36 I ACCTGC∧4/8), *Bbs* I (GAAGAC∧2/6), *Bfu*A I (ACCTGC∧4/8), *Bpi* I (GAAGAC∧2/6), *Eco*31 I (GGTCTC∧1/5), and *Esp*3 I (CGTCTC∧1/5).

Since, as reviewed above, the current seamless cloning method does not work unless certain type IIs enzymatic sites in the vector are disabled or mutated, error-prone inverse PCR is required prior to every cycle of duplex insertion. One way to overcome the limits of current seamless cloning methods would be to employ the strategy from Lewis et al.,[25] i.e., to use an appropriate combination of two different type IIs sites and a unique, ordinary type IIp site in the plasmid. This approach should allow construction of the desired coding sequence from any fragments via either the concatemerization or iterative approach. For example, consider the two different type IIs enzymes (E and F in Figure 3B) that recognize different sequences and cleave at the same distance away from these sites but in the opposite directions to create equal numbers of overhang bases. As in Figures 3B and 5, two aliquots of a plasmid with monomer inserts would be digested by two enzyme sets: E/H and F/H. Enzyme H can be a commonly used type IIp enzyme such as *Bam*H I or *Eco*R I. E and F would be a pair of type IIs enzymes such as *Eco*31 I and *Esp*3 I. The flanking bases (the bases to be cut) can be chosen randomly in the monomer sequence but have to be nonpalindromic and compatible. Then, the gap created by the E/H double cut can be filled by ligation with the fragment generated by the F/H double cutting. This would link the two monomers head-to-tail, insert them into the preexisting duplex at the designated site, and retain all three unique restriction sites for another round. This method would retain all of the advantages of seamless cloning, including unidirectional ligation of monomers and would avoid having extraneous amino acid between each monomer. Additionally, it enables oligomerization of heterogeneous monomers as well as domain combinations with a predetermined sequence. Also, since the ends of both digestion fragments are not compatible, cloning efficiency will be boosted because two enzymatic steps are eliminated (phosphorylation of inserts and dephosphorylation of vector) and because problems facing the traditional seamless cloning method, such as self-ligation of individual fragments (circularization), are eliminated. In general, as a universal molecular cloning method, it enables a "clean" recombination of any sequence modules.

An alternative to most of the method reviewed above, where monomers or multiple domains are joined primarily through restriction sites, a restriction site-free approach (called overlapping PCR), has been described.[36] Briefly, a consensus sequence (around 15−25 bases) is added as an adaptor on the 3′ end of the sense strand of fragment 1 and on the 5′ end of the sense strand of fragment 2. The two fragments are then amplified once (using a primer complementary to the adapter) and then mixed and used as the template for a second round of PCR, which links the two fragments together (since the priming sites are on the 5′ end of sense fragment 1 and the 3′ end of antisense fragment 2, respectively). To make this approach successful, codon degeneracy is helpful.

PCR seamless cloning and restriction site-free-ligation-free cloning methods[37,38] have also been adapted to protein engineering. In its original form,[37] chimeric DNA/RNA primers, composed mainly of DNA (for superior annealing specificity) were used. One or more ribonucleotides were included at the 5′ ends to generate single-strand RNA overhangs flanking the double-stranded DNA. The overhangs were used to recombine DNA fragments at any location unidirectionally. This method takes advantage of polymerases such as *pfu* polymerase that lack reverse transcription activity. However, the linkage efficiency was relatively poor, due to the short single stranded RNA tails used. Recently, Donahue et al.[38] improved this method by using longer RNA overhangs and more importantly by including a single 2′-*O*-methylribonucleotide rather than a normal ribonucleotide into the primers to pause polymerase extension (hence these primers were called "terminator primers"). Circular plasmids were quickly generated in *E coli* cells transformed by a mixture of PCR products. The authors also demonstrated fast and versatile vector modular recombinations, including changing the antibiotic resistance gene, origin of replication, and gene of interest. This method is called t-PCR cloning.

Although PCR revolutionized molecular biology by enabling easy amplification and manipulation of DNA fragments, this technology also has certain limitations. For example, it is difficult to amplify fragments longer than 20 kilobases, which are often the sizes of interest in the protein polymer field. A novel, easy to use technique, ET recombination,[39,40] can be used to clone gene fragments up to 80 kb into any desired position in any plasmid vector. This method, also termed phage-mediated recombination, utilizes the 5′ to 3′ exonuclease activity of a phage-derived homlogue of the *E coli* protein RecE (Redα) and *E coli* DNA replication assisted by a single stranded binding protein (RecT or Red*β*) to specifically and precisely swap DNA sequences between two molecules. Advantages of this technique include its independence of restriction sites and indifference to DNA length; it is also easy and effective for performing mutagenesis, including insertions, deletions, and substitutions. Moreover, this method is less laborious than traditional DNA manipulation techniques and therefore saves lots of time and expense.

5. Conclusions

Several molecular cloning methods frequently used for constructing sequences encoding artificial proteins containing repeated sequences have been reviewed, highlighting some of their major advantages and limitations. Many of the limitations posed by these current methods, which generally employ restriction enzymes to digest strategically placed restriction sites in the coding sequences and the plasmid cloning vectors, can be avoided by an alternative approach: MDC. Superior efficiency and accuracy are two major advantages of this alternate approach, which combines PCR and codon degeneracy. These advantages are illustrated by applying this strategy to the synthesis of duplexes encoding protein polymers having repetitive amino acid sequence patterns. Some emerging techniques were also reviewed. The primary goal of this review has been to call attention to the importance of codon usage in the design of sequences encoding identical repeating patterns of amino acid sequences. Since codon usage issues can be a critical factor for achieving successful cloning outcomes and high protein yields, the importance of codon degeneracy consideration at both local and global levels is emphasized.

Acknowledgment. This work was supported by NASA through Grant NAG 9-1345. The author thanks his mentor Dr. James Harden for generous financial support of this project and fruitful discussions and Dr. David Tirrell at CalTech for generously providing the plasmid pQE9-L2AC$_{10}$B encoding the original triblock protein WPT2. He also thanks Chong Lee, Wendy Petka, and Jill Sakata for assistance in the early stages of this work. Finally, the author thanks Dr. Thomas Mattson at Georgetown University for manuscript editing.

References and Notes

(1) Doel, M. T.; Eaton, M.; Cook, E. A.; Lewis, H.; Patel, T.; Carey, N. H.; *Nucleic Acids Res*. **1980**, *8* (20), 4575−92.
(2) Regan, L.; DeGrado, W. F. *Science* **1988**, *241*, 976−978.
(3) Cappello, J.; Crissman, J.; Dorman, M.; Mikolajczak, M.; Textor, G.; Marquet, M.; Ferrari, F. *Biotechnol. Prog*. **1990**, *6*, 198−202.
(4) Cappello, J. *Trends Biotechnol*. **1990**, *8*, 309−311.
(5) McGrath, K. P.; Tirrell, D. A.; Kawai, M.; Mason, T. L.; Fournier, M. J. *Biotechnol. Prog.* **1990**, *6*, 188−192.
(6) McGrath, K. P.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. *J. Am. Chem. Soc*. **1992**, *114*, 727−733.
(7) *Protein-Based Materials*; McGrath, K. P., Kaplan, D., Eds.; Birkhauser: Boston, 1999.
(8) McPherson, D. T.; Morrow, C.; Minehan, D. S.; Wu, J.; Hunter, E.; Urry, D. W. *Biotechnol. Prog.* **1992**, *8*, 347−352.
(9) Prince, J. T.; McGrath, K. P.; DiGirolamo, C. M.; Kaplan, D. L. *Biochemistry* **1995**, *34*, 10879−10885.
(10) Petka, W. A. Reversible gelation of genetically engineered macromolecules. Ph.D. Thesis, Unversity of Massachusetts: Amherst, 1997.
(11) Petka, W. A.; Harden, J. L.; McGrath, K. P.; Wirtz, D.; Tirrell, D. A. *Science* **1998**, *281*, 389−392.
(12) Krejchi, M. T.; Atkins, E. D. T.; Waddon, A. J.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. *Science* **1994**, *265*, 1427−1432.
(13) Petka, W. A.; Harden, J. L.; Sakata, J. K.; Tirrell, D. A. *Mater. Res. Soc. Symp. Proc.* **1999**, *550*, 23−28.
(14) Huang, J.; Valluzzi, R.; Bini, E.; Vernaglia, B.; Kaplan, D. L. *J. Biol. Chem.* **2003**, *278*, 46117−46123.
(15) Kostal, J.; Mulchandani, A.; Chen, W. *Macromolecules* **2001**, *34*, 2257−2261.
(16) Chilkoti, A.; Dreher, M. R.; Meyer, D. E. *Adv. Drug Delivery Rev.* **2002**, *54*, 1093−1111.
(17) Panitch, A.; Yamaoka, T.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. *Macromolecules* **1999**, *32*, 1701−1703.
(18) Kobatake E.; Onoda K.; Yanagida Y.; Haruyama T.; Aizawa M., *Biotechnol. Tech.* **1999**, *13*, 23−27.
(19) Betre, H.; Setton, L. A.; Meyer, D. E.; Chilkoti, A. *Biomacromolecules* **2002**, *3*, 910−916.
(20) DiZio, K. A.; Tirrell, D. A. *Macromolecule* **2003**, *36*, 1553−1558.
(21) Heilshorn, S. C.; DiZio, K. A.; Welsh, E. R.; Tirrell, D. A. *Biomaterials* **2003**, *24*, 4245−4252.
(22) Chen, J. S.; Altman, G. H.; Karageorgiou, V.; Horan, R.; Collette, A.; Volloch, V.; Colabro, T.; Kaplan, D. L. *J. Biomed. Mater. Res. A* **2003**, *67*, 559−570.
(23) Mi, L.; Fischer, S.; Chung, B.; Sundelacruz, S.; Harden, J. L. *Biomacromolecules* **2006** *7*, 38−47.
(24) Haider, M.; Leung, V.; Ferrari, F.; Crissman, J.; Powell, J.; Cappello, J.; Ghandehari, H. *Mol Pharm*. **2005,** *2*, 139−150.
(25) Lewis, R. V.; Hinman, M.; Kothakota, S.; Fournier, M. J. *Protein Expression Purification* **1996**, *7*, 400−406.
(26) Dougherty, M. J.; Kothakota, S.; Mason, T. L.; Tirrell, D. A.; Fournier, M. J. *Macromolecules* **1993**, *26*, 1779−1781.
(27) Yoshikawa, E.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. *Macromolecules* **1994**, *27*, 5471−5475.
(28) McPherson, D. T.; Xu J.; Urry D. W. *Protein Expression Purification* **1996**, *7*, 51−57.
(29) Winkler, S.; Szela, S.; Avtgas, P.; Valluzzi, R.; Kirschner, D. A.; Kaplan, D. *Int. J. Biol. Macromol*. **1999**, *24*, 265−270.
(30) Meyer, D. E.; Chilkoti, A. *Biomacromolecules* **2002**, *3*, 357−367.
(31) Padgett, K. A.; Sorge, J. A. *Gene* **1996**, *168*, 31−35.
(32) McMillan, R. A.; Lee, T. T.; Conticello, V. P. *Macromolecules* **1999**, *32*, 3643−3648.
(33) Ochman, H.; Ajioka, J. W.; Garza, D.; Hartl, D. L. *Biotechnology* **1990**, *8*, 759−760.
(34) Goeden-Wood, N. L.; Conticello, V. P.; Muller, S. J.; Keasling, J. D. *Biomacromolecules* **2002**, *3*, 874−879.
(35) Roberts, R. J.; Vincze, T.; Posfai, J.; Macelis, D. *Nucleic Acids Res.* **2005**, *33*, D230-D232.
(36) Shuldiner, A. R.; Tanner, K.; Scott, L. A.; Moore, C. A. Roth, J. *Anal. Biochem.* **1991**, *194*, 9−15.
(37) Coljee, V. W.; Murray, H. L.; Donahue, W. F.; Jarrell, K. A. *Nat. Biotechnol*. **2000**, *18*, 789−791.
(38) Donahue, W. F.; Turczyk, B. M.; Jarrell, K. A. *Nucleic Acids Res.* **2002**, *30*, e95.
(39) Zhang, Y.; Buchholz, F.; Muyrers, J. P. P.; Stewart, A. F. *Nat. Genet.* **1998**, *20*, 123−128.
(40) Zhang, Y.; Muyrers, J. P. P.; Testa, G.; Stewart, A. F. *Nat. Biotechnol*. **2000**, *18*, 1314−1317.

BM050158H