# 3

# *De novo* Proteins From Binary-Patterned Combinatorial Libraries

**Luke H. Bradley, Peter P. Thumfort, and Michael H. Hecht**

**Summary**

Combinatorial libraries of well-folded *de novo* proteins can provide a rich source of reagents for the isolation of novel molecules for biotechnology and medicine. To produce libraries containing an abundance of well-folded sequences, we have developed a method that incorporates both rational design and combinatorial diversity. Our method specifies the "binary patterning" of polar and nonpolar amino acids, but allows combinatorial diversity of amino acid side chains at each polar and nonpolar site in the sequence. Protein design by binary patterning is based on the premise that the appropriate arrangement of polar and nonpolar residues can direct a polypeptide chain to fold into amphipathic elements of secondary structures, which anneal together to form a desired tertiary structure. A designed binary pattern exploits the periodicities inherent in protein secondary structure, while allowing the identity of the side chain at each polar and nonpolar position to be varied combinatorially. This chapter provides an overview of the considerations necessary to design binary patterned libraries of novel proteins.

**Key Words:** Protein design; binary patterning; combinatorial library; *de novo* proteins; library design.

## 1. Introduction

The amino acid sequences in a combinatorial library can be drawn from an enormous number of possibilities. For example, for a chain of 100 residues composed of the 20 standard amino acids, there are $20^{100}$ possible sequences. Because sequence space is so enormous, neither nature nor laboratory studies can explore all possibilities.

Although the quantity of sequences in a randomly generated library may be enormous, the quality of those sequences is likely to be low. Indeed, libraries of randomly generated sequences yield proteins with desired properties only very rarely ***(1–5)***. Powerful methods for screening and selection can enable the

From: *Methods in Molecular Biology, vol. 340: Protein Design: Methods and Applications*
Edited by: R. Guerois and M. López de la Paz © Humana Press Inc., Totowa, NJ

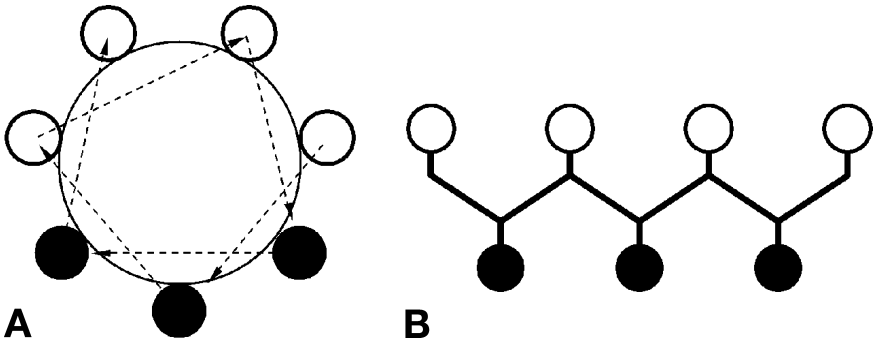


Fig. 1. The designed binary pattern of polar (open circles) and nonpolar (closed circles) amino acids for α-helices and β-strands exploits the inherent periodicities of the respective secondary structures. (**A**) α-helices have a repeating periodicity of 3.6 residues per turn. By placing a nonpolar amino acid at every third or fourth position, an amphipathic helix can be encoded in which one face is polar and the opposite face is nonpolar. Note that this figure shows the positioning of seven amino acids. For longer α-helices, the binary pattern can be adjusted to allow the periodicity of 3.6 residues per turn to maintain the amphipathic nature through the entire length of the helix. (**B**) β-strands have an alternating periodicity of polar and nonpolar amino acids. This pattern would cause one face of the strand to be polar and the opposite face to be nonpolar.

isolation of rare "winners" from vast libraries of inactive candidates; however, the success of these methods depends on the quality of the library being screened or selected. To enhance the likelihood of success, combinatorial libraries must be focused into regions of sequence space that are most likely to yield well-folded proteins. Thus, the numerical power of the combinatorial approach must be tempered by features of rational design.

To focus a library into the most productive regions of sequence space, we are guided by the observation that natural proteins typically fold into structures that (1) contain abundant secondary structures and (2) expose polar side chains to solvent while burying nonpolar side chains in the interior. Our strategy for protein design draws on these two features to rationally design focused libraries of *de novo* sequences in ways that favor folded structures.

The strategy—called the binary code strategy—is based on the premise that the appropriate patterning of polar (P) and nonpolar (N) residues can direct a polypeptide chain to form amphiphilic secondary structure ***(6–9)***. A designed binary pattern exploits the periodicities inherent in protein secondary structure: α-helices have a periodicity of 3.6 residues per turn, whereas β-strands have an alternating periodicity (**Fig. 1**). Thus, a binary patterned sequence designed to form amphipathic α-helices would place nonpolar residues at

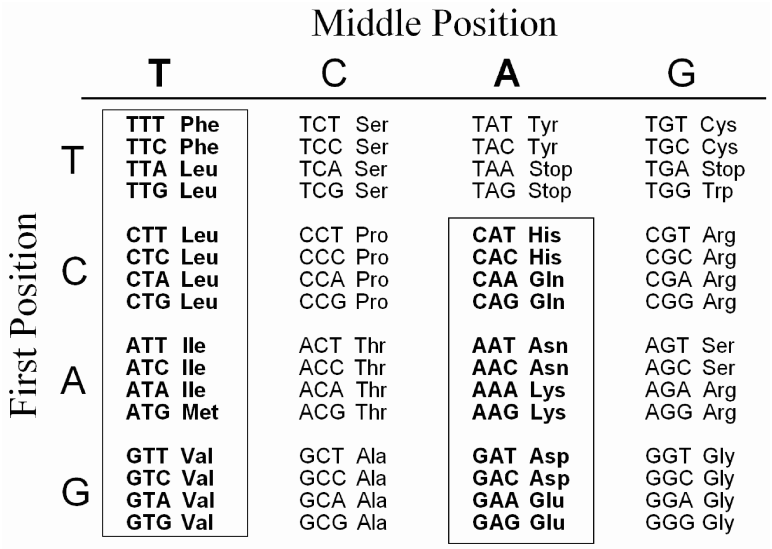


Fig. 2. The organization of the genetic code allows sequence degeneracy for polar and nonpolar amino acids to be incorporated into a combinatorial library of synthetic genes by defining the middle position of the codon. For a nonpolar amino acid position, the degenerate codon NTN would encode Phe, Leu, Ile, Met, or Val. For positions requiring polar amino acids, the degenerate codon VAN would encode His, Gln, Asn, Lys, Asp, or Glu. (N represents an equimolar mixture of A, C, T, and C; V represents an equimolar mixture A, C, or G.).

every third or fourth position. In contrast, the binary pattern for an amphipathic β-strand would alternate between polar and nonpolar residues (*see* **Note 1**). In the binary code strategy, the precise three-dimensional packing of the side chains is not specified *a priori*. Consequently, within a library of binary patterned sequences, the identity of the side chain at each polar and nonpolar position can be varied, thereby facilitating enormous combinatorial diversity.

A combinatorial library of binary patterned proteins is expressed from a combinatorial library of synthetic genes. Each gene encodes a different amino acid sequence, but all sequences within a given library have the same patterning of polar and nonpolar residues. This sequence degeneracy is made possible by the organization of the genetic code (**Fig. 2**). The degenerate codon NTN encodes nonpolar amino acids, whereas the degenerate codon VAN encodes polar amino acids. (V = A, G, or C; N = A, G, C, or T; *see* **Subheading 3.2.** on codon usage.) With these degenerate codons, positions requiring a nonpolar amino acid are filled by phenylalanine, leucine, isoleucine, methionine, or valine, whereas positions requiring a polar amino acid are filled by glutamate, aspartate, lysine, asparagine, glutamine, or histidine (**Fig. 2**; *see* **Note 2**).

This chapter outlines the methodology for using binary patterning to design libraries of *de novo* proteins. Using examples from our laboratory, we describe the design and construction of both α-helical and β-sheet proteins.

## 2. Materials

Oligonucleotides were purchased from commercial vendors (e.g., IDT; www.idtdna.com). All oligonucleotides should be PAGE purified (*see* **Note 3**). For overlap extensions, we use a thermostable polymerase that leaves blunt ends (such as Deep Vent [New England Biolabs] or Pfu polymerase). All other enzymes and reagents are commercially available.

## 3. Methods

### *3.1. Design of a Structural Template*

Binary patterning can be applied to any amphipathic α-helical or β-stranded segment of a protein. Although our laboratory has focused on *de novo* proteins, the binary code strategy can also be applied to local areas of existing proteins such as the active site, part of the core, or an interface ***(10)***. For the design of *de novo* proteins, the success of the strategy depends primarily on how well the template is designed. Several factors important for template design are described.

#### *3.1.1. Binary Patterned Regions*

##### 3.1.1.1. α-Helical Designs

Binary patterning exploits the periodicities inherent in secondary structures. α-Helices have a repeating periodicity of 3.6 residues per turn (**Fig. 1A**). To design an amphipathic segment of α-helical secondary structure, a binary pattern of P-N-P-P-N-N-P is used. Our initial α-helical design focused on the four-helix bundle motif (**Fig. 3**). In this structure, the hydrophobic face of each helix is oriented toward the central core of the bundle, whereas the hydrophilic faces of the helices are exposed to aqueous solvent. The P-N-P-P-N-N-P pattern favors formation of an amphiphilic α-helical structure that can bury all nonpolar amino acids upon formation of the desired tertiary structure. From our designed four-helix bundle libraries, more than 60 proteins have been purified and characterized. All have shown typical α-helical circular dichroism spectra. Additionally, the collection yielded several proteins with native-like properties, including nuclear magnetic resonance (NMR) chemical shift dispersion, co-operative chemical and thermal denaturations, and slow hydrogen/deuterium exchange rates ***(11–15)***. Recently, the structure of protein S-824 from a second-generation binary patterned library was determined by NMR spectroscopy, and shown to be a four-helix bundle as specified by the binary code design ***(16)***.

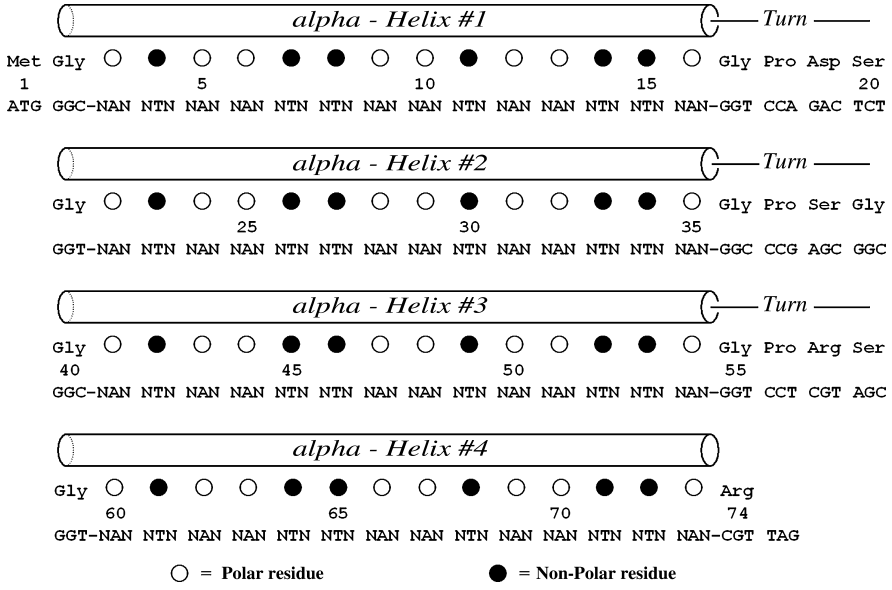


Fig. 3. The design template for the initial four-helix bundle library. Nonpolar positions (closed circles) were encoded by the degenerate NTN codon, and polar positions (open circles) were encoded by the degenerate VAN codon. The defined residue positions were located at the N- and C-terminal regions as well as the interhelical turns of the designed protein.

### 3.1.1.2. β-Sheet Design

Amphiphilic β-strands have an alternating periodicity of …P-N-P-N… (**Fig. 1B**). Based on this periodicity, a combinatorial library of synthetic genes can be created to encode β-sheet structures in which polar residues comprise one face and nonpolar residues comprise the opposing face of the resulting β-sheet. The sequences in our first β-sheet library were designed to have six β–strands with each strand having the binary pattern P-N-P-N-P-N-P *(7)*. Proteins from this library were expressed from synthetic genes cloned into *Escherichia coli*. All proteins from this collection that have been analyzed thus far indeed form β-sheet secondary structure, displaying circular dichroism spectra with the characteristic minimum at approx 217 nm. In aqueous solution, the β-sheet proteins from this initial library self-assembled into amyloid-like fibrils, with nonpolar side chains forming a hydrophobic core and polar side chains exposed to solvent *(7)*.

When these same β-sheet sequences are placed in a heterogeneous environment with a polar/nonpolar interface, they form a different structure. For example,

at an air/water interface, they self-assemble into flat β-sheet monolayers with the nonpolar residues pointing up toward air and polar side chains pointing down toward water ***(17)***. Alternatively, at an interface between water and the nonpolar surface of graphite, binary patterned β-sheet sequences undergo template-directed assembly on the graphite surface to yield highly ordered structures ***(18)***.

The formation of fibrils in aqueous solution, and monolayers at polar/nonpolar interfaces is consistent with the inherent tendency of β-strands to assemble into oligomeric structures ***(19,20)***. Designed β-strands with sequences that adhere rigorously to the alternating polar/nonpolar binary pattern are especially prone to aggregate because of their need to bury their "sticky" hydrophobic face ***(8,21)***. To favor monomeric β-sheet proteins, the alternating binary pattern must be modified: The pattern P-N-P-N-P-N-P on the edge strand of a β-sheet can be changed to P-N-P-**K**-P-N-P (where K denotes lysine). The four methylene groups of the lysine side chain can substitute for the replaced nonpolar residue, whereas the charged amine at the end of the lysine side chain will seek solvent and thereby prevent aggregation. This strategy has been used successfully to convert binary patterned *de novo* proteins from amyloid-like structures to monomeric β-sheet proteins ***(21)***.

### *3.1.2. Fixed Regions*

In practice, it is often necessary to keep part of the protein sequence fixed (i.e., not combinatorially diverse), especially when the target sequence is long. When assembling a library of synthetic genes, these constant regions serve as sites for single-stranded synthetic oligonucleotides to anneal together and prime the enzymatic synthesis of complementary strands (**Fig. 4**). (Assembly of full-length genes from single-stranded oligonucleotides is discussed in **Subheading 3.3.**).

Single-stranded oligonucleotides are typically used to encode the binary patterning of individual segments of secondary structure. Nondegenerate fixed regions on the 5' and 3' termini of these oligonucleotides are typically used to encode fixed turn regions between units of secondary structure (**Figs. 3** and **4**) ***(6,7)***. The amino acid sequences chosen to occupy these turn regions are based on statistical and rational design criteria outlined as follows.

1. Sequences in the turn regions are chosen based on positional preferences. For example, in the initial four-helix bundle library, glycine residues were placed at N-cap and C-cap positions at the termini of the helices (**Fig. 3**) ***(6)***. Glycine residues are frequently found at these positions in natural proteins ***(22)***. At the position after the C-cap, proline residues were used because they are strong helix breakers. In some situations, however, proline may be undesirable because *cis*/*trans* isomerism could lead to multiple (rather than unique) conformations. For the β-sheet library ***(7)***, design of the turn regions was based on the positional turn potentials of various amino acids in the known structures of natural proteins ***(23)***.

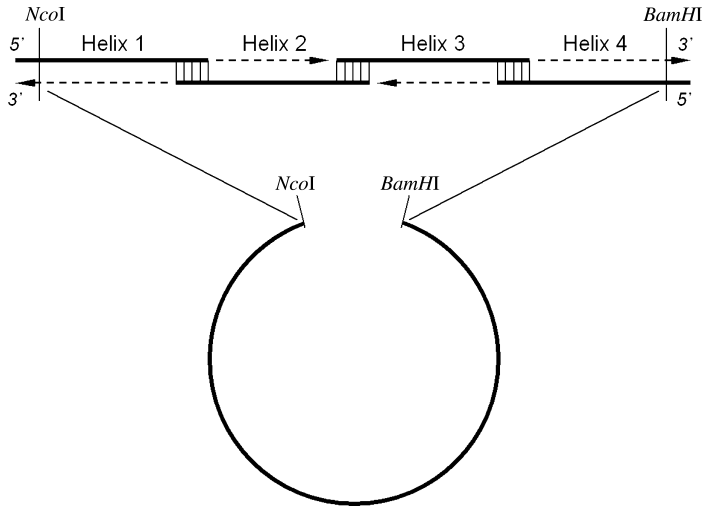


Fig. 4. Assembly of full-length genes from four single-stranded oligonucleotides. Constant regions at the 5' and 3' ends of the single-stranded oligonucleotides serve as sites for annealing and for priming enzymatic synthesis (using DNA polymerase) of complementary strands. The full-length gene is then ready for insertion into an expression vector.

2. Sequences of the turns can be designed to incorporate restriction sites, which can facilitate the assembly of full-length genes ***(6)*** (*see* **Subheading 3.3.**).
3. The lengths of constant regions must be sufficient to promote sequence-specific annealing. Pairs of oligonucleotides with overlaps of 12 to 15 nucleotides are typically used for annealing. To further enhance annealing, one or two nucleotides in the codons immediately preceding and following the turn regions may also be held constant. For example, the synthetic oligonucleotide (5'…NAN-NTN-NTN-N**AN**-GGT-CCT-CGT-AGC-3') has a constant gene segment (underlined) in which 12 nucleotides encode a four residue turn. The previous codon (NAN) encodes a polar amino acid residue. By defining the third position, for example with a G, the codon now becomes NAG, thereby yielding two additional constant bases (bold) for sequence-specific annealing (5' …NAN-NTN-NTN-N**AG**-GGT-CCT-CGT-AGC-3'). By defining only the second and third (but not the first) positions of the codon, amino acid diversity is maintained.

Besides the turn regions, the N- and C- termini of the *de novo* sequences can also be held constant. Fixed sequences in these regions are typically necessary for cloning into expression vectors. Some design criteria for the termini are:

1. An initiator methionine is placed at the N-terminus of the *de novo* sequence. This is required for expression in vivo.
2. An aromatic chromophore (tyrosine or tryptophan) can be incorporated at a constant site in the sequence to aid in protein purification and concentration deter-

mination *(7,15)*. This aromatic residue could be placed either in a constant turn or at one of the chain termini. In some of our libraries, we inserted a tyrosine immediately after the initiator methionine. This provides a chromophore, while also preventing cleavage of the initiator methionine in vivo *(24–27)*.

3. The C-terminal residue of the designed proteins should be charged and polar. The C-terminal sequence of a protein can affect its rate of intracellular proteolysis, and the presence of a charged residue at the C-terminus can extend half-life in vivo *(28–30)*. In addition, an exhaustive statistical analysis of protein amino acid sequences found that polar and charged amino acids are overrepresented at the C-termini of proteins *(31)*. For the four-helix bundle libraries, an arginine residue was designed to occupy this terminal position *(6,15)*. Moreover, positively charged side chains at the C-termini of α-helices (and negatively charges residues at the N-termini) can enhance stability by interacting with the helix dipole *(32)*.

### 3.1.3. Considerations for the Design of Tertiary Structure

A successful binary patterned template must be long enough to encode well-folded structures, but at the same time short enough to be accessible to strategies for assembling large libraries of error-free genes. Many proteins from our first generation (74-residue) four-helix bundle library formed dynamic structures resembling molten globules *(6,11–14)*. To investigate the potential of the binary code strategy to encode collections of native-like tertiary structures, a second-generation library of binary-patterned α-helical proteins was prepared *(15)*. This new library was based on protein #86, a preexisting sequence from the original 74-residue library. The major change to protein #86 was the addition of six combinatorially diverse residues to each of the four helices. These residues continued to follow the binary patterning. Overall, the second-generation proteins were 102 amino acids in length, which is similar to a number of natural four-helix bundles.

Characterization of five sequences chosen arbitrarily from this second-generation library showed that all were substantially more stable than the parental protein #86 *(15)*. In addition, most of them yielded NMR spectra that were well dispersed and exhibited well-resolved nuclear Overhauser effect cross peaks, indicative of well-folded tertiary structures *(15)*. Recently, the solution structure of a protein (S-824) from this second-generation library was solved *(16)*. The structure was indeed a four-helix bundle in accordance with its binary patterned design. Moreover, the protein was not a molten globule: the interior side chains were well ordered—even by the standards of natural proteins *(16)*.

## 3.2. Codon Usage

### 3.2.1. The VAN (Polar) Codon

1. The first base of the VAN codon is occupied by an equimolar mixture of G, C, and A. By excluding T, two stop codons (TAG and TAA) and two tyrosine codons are eliminated (*see* **Note 4**).

2. The mixture of nucleotides at the third base of the VAN codon can be altered to favor some amino acids over others. An equimolar mixture of G, C, A, and T would yield an equal likelihood of histidine, glutamine, asparagine, lysine, aspartate, and glutamate. However, some of the residues of the VAN codon have higher intrinsic propensities than others to form α-helices ***(33–36)***. By omitting T from the third position (i.e., VAV), glutamine, lysine, and glutamate are favored over histidine, asparagine, and aspartate. This increases the percentage of residues with high propensities to form α-helices ***(33–36)***.

### *3.2.2. The NTN (Nonpolar) Codon*

Equimolar mixtures of all four bases at the first and third positions of the NTN codon would encode six times as many leucines as methionines. In addition, an equimolar mixture would encode protein sequences in which one quarter of the hydrophobic residues would be valine. Because valine has a relatively low α-helical propensity ***(33–36)***, this may be undesirable for some designs. By altering the molar ratio of the mixture at the first and third N positions, the relative abundance of hydrophobic residues can be altered. For example, in the initial four-helix bundle library, the first base of the NTN codon contained A:T:C:G in a molar ratio of 3:3:3:1 and the third base mixture contained an equimolar mixture of G and C ***(6)***.

### *3.2.3. Codon Usage of the Host Expression System*

The DNA sequences in both the constant and the combinatorial regions of the library should be biased to favor those codons used most frequently in the host expression system. For example, including only C and G (rather than all four bases) in the third position of a degenerate codon favors those codons preferred by *E. coli* ***(37)***. Codons that are used only rarely in the host expression system, such as CGA, AGA, and AGG (arginine); CTA (leucine); CCC (proline); and ATA (isoleucine) in *E. coli* should be avoided wherever possible, because genes containing rare codons may express poorly ***(38)***. Other (non–*E. coli*) expression systems have different codon preferences, and these must be considered in the design.

### *3.2.4. Restriction Digest Analysis* in Silico

To ensure that those restriction sites used for ligating the library into the host vector occur only at the ends of the genes and not within the (degenerate) gene sequences, restriction digests are performed *in silico* before finalizing the design of the gene library. We have developed a program that reads the binary patterned sequence—including the degenerate wildcard bases (*see* **Note 5**)—and randomly generates a large number (typically $10^4$) of gene sequences from the library master template. This pool of gene sequences is subsequently analyzed against a restriction enzyme database, and the data for all restriction sites

are sorted by cutting frequency. This enables a quick comparison among different design choices and provides a list of sites that will be absent from the library and can therefore be used for cloning into the appropriate vectors.

For example, the NdeI restriction site, CATATG, includes the ATG initiator codon and is commonly used as a cloning site at the 5' end of genes. To assess whether this site will appear in a combinatorial library, one must consider the degenerate codons used to construct the library. For example, the combinatorial sequence VAN.NTN, which encodes a polar residue followed by a nonpolar residue, would contain the NdeI site in 1 of 192 cases. In contrast, the combinatorial sequence VAV.NTS, which also encodes a polar residue followed by a nonpolar residue, will never encode an NdeI site. (N denotes an equimolar mixture of all four bases; V, an equimolar mixture of A, G, and C; and S, an equimolar mixture of G and C.)

### *3.3 Assembly of Full-Length Genes*

Full-length genes are typically assembled from smaller single-stranded oligonucleotides (**Fig. 4**). This is done for two reasons: (1) to minimize the inherent errors (mostly deletions and frameshifts) associated with the synthesis of long degenerate oligonucleotides and (2) to increase the diversity of full-length sequences via combinatorial assembly (*see* **Note 6**).

When synthesizing the semirandom oligonucleotides, some are made as coding (sense) strands and others as noncoding (antisense) strands (*see* **Fig. 4**). Typically, each oligonucleotide is designed to encode an individual segment of secondary structure. Assembly of full-length genes from such segments allows individual α-helices or β-strands to be designed and manipulated as independent modules, thereby enhancing the versatility of the binary code strategy.

In the design of our initial library of four-helix bundles, four synthetic oligonucleotides were used to construct the full-length gene. Each oligonucleotide was designed to encode a single helix and turn. As described above, the turn regions were defined precisely (i.e., not degenerate), thereby allowing them to serve as priming sites for DNA polymerase to synthesize complementary strands (**Fig. 4**) ***(6)***.

Various methods can be used to assemble the full-length genes. In some cases, we have made two libraries of half genes, and then ligated them together to produce a library of genes encoding full-length proteins ***(6)***. To ensure correct head to tail ligation, nonpalindromic restriction sites, which produce directional "sticky ends" for ligation, can be designed into the constant regions ***(6)***. Other methods for assembling full-length genes include various polymerase chain reaction strategies (e.g., overlap extension), and these have been used in constructing several of our libraries ***(7,15,39)***.

### 3.4. Optimization of Gene Assembly

#### 3.4.1. Avoidance of Incorrect Annealing

The correct assembly of full-length genes (**Subheading 3.3.**) can be hindered by the presence of alternate annealing sites in the synthetic oligonucleotides. Such sites would result from internal repeats or inverted repeats and thus give rise to problems such as hairpins or mispriming. To minimize these undesired sequences, the gene template sequence (containing both fixed and degenerate bases), should be analyzed computationally, and optimized prior to synthesis.

Although algorithms (e.g., dotplots) are available to search DNA sequences for alternate annealing sites or possible secondary structure ***(40–43)***, the standard methods are not suitable for analyzing the degenerate sequences required to assemble combinatorial libraries. Therefore we devised a new algorithm, called *Designer DotPlot* to analyze degenerate DNA sequences. In contrast to previous methods, which search DNA sequences containing only the four standard bases A, T, C, and G, *Designer DotPlot* also handles sequences containing the wildcards N, B, D, H, V, M, S, W, Y, H, and K (*see* **Note 5**). *Designer DotPlot* calculates and displays probability-weighted matches between all compatible degenerate bases. In addition to standard base-pairings (A-T, G-C), an optional feature of *Designer DotPlot* also allows G-T base pairs. For example, the degenerate bases R (A/G) and N (A/C/G/T) have a match probability of 0.25. Allowing G-T base pairing increases it to 0.375. *Designer DotPlot* uses a $15 \times 15$ lookup table containing the probabilities for all 225 combinations of the 4 standard DNA bases and all 11 wildcards.

**Figure 5A** shows the dotplot for a sequence encoding a binary patterned α-helix bracketed by constant sequences encoding the helix caps. The darkness of each dot shows the probability of a match. For example, an A-T match has a probability of 1 and is shaded black, whereas an A-B match has a probability of 0.33 and is shown in light gray. Two such plots are generated: one forward (**Fig. 5A**), analyzing the interstrand pairings between the upper and lower strand; and one reverse (not shown), analyzing the intrastrand pairings within the upper strand itself. Each plot is then filtered (window size, probability) and overlaid. Probabilities above a threshold are transformed into a log-scale and rescaled to gray values with the highest value being plotted for every position.

Stretches of sequence that contain a high probability of complementary bases are plotted as lines in **Fig. 5B**, with the problem areas for misannealing and hairpins outlined. Based on the analysis of such plots, the base sequence (including both fixed and degenerate regions) of each design can be improved manually by simply changing the problem regions to alternate codons and replotting, or automatically via a genetic algorithm. The optimized sequence (**Fig. 5C**)

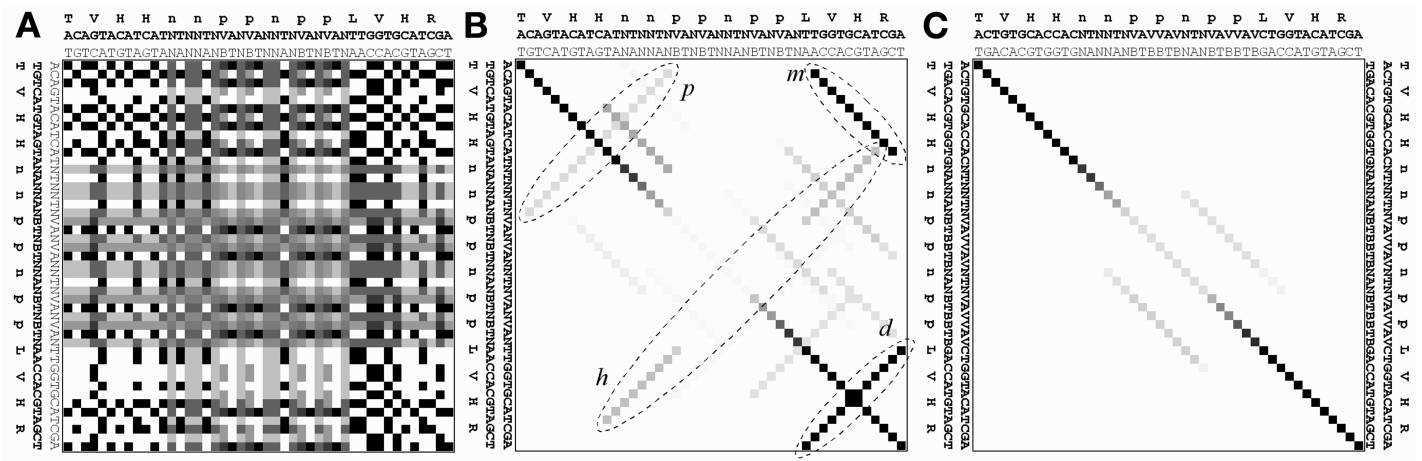


Fig. 5. A sequence analysis and optimization of a binary patterned library using *Designer Dotplot*. For a desired amino acid sequence, the upper (sense) strand and the complementary lower (anti-sense) strand gene templates are labeled on the *x* and *y* axes, respectively. (**A**) *Designer dotplot* shows single base matches between the upper and lower strand. The gray scale corresponds to base-pairing probability (darker = greater probability). (**B**) After filtering (window size = 8; cutoff probability = 0.01) and overlaying the *Designer Dotplot* with intrastrand base pairings, problem regions within the sequence are identified as lines. These highlighted assembly problem areas are from palindromic (p), hairpin (h), mispriming (m), and primer-dimer sites (d) present in the DNA library template sequence. (**C**) The highlighted problem areas identified in (B) have been minimized following sequence optimization. By using different codon choices while maintaining same binary pattern, the library template has been optimized for gene assembly.

minimizes possible cross-annealing among oligonucleotides, misannealing or hairpin formation within the same oligonucleotide, and reverse complementarity during assembly of the oligonucleotides (*see* **Note 7**).

### *3.4.2. Preselection for Open Reading Frames*

Constructing large and diverse libraries that are free of frameshifts or stop codons can be extremely challenging. The presence of interrupted sequences complicates screening strategies by increasing the burden on the selection system, while failing to provide additional valid candidate genes for evaluation. It is therefore important to remove incorrect sequences from these libraries before screening for function.

To construct a high-quality combinatorial library of *de novo* genes, we have developed a method to screen libraries of gene segments for open reading frames before assembly of the full-length genes ***(39)***. In our system, synthetic DNA from a library is inserted upstream from a selectable intein/thymidylate synthase (TS) fusion (*see* **Note 8**). Gene segments that are in-frame and devoid of stop codons produce a tripartite precursor protein. Subsequent cleavage by the intein releases and activates the TS enzyme. This enables TS deficient *E. coli* host cells to survive in selective medium. Libraries of error-free gene segments are isolated from the surviving cells, and the individual segments are subsequently assembled combinatorially into full-length genes (**Subheading 3.3.**; *also see* **Note 6**). This preselection system has recently enabled the construction of a large library of 102 amino acid sequences in which virtually all sequences are free of frameshifts and deletions ***(39)***.

The availability of large, diverse, and error-free libraries of binary patterned sequences encoding well-folded and native-like structures sets the stage for experiments aimed at the isolation of novel proteins with functions that may ultimately find use in biotechnology and medicine.

## 4. Notes

1. The overall amino acid compositions are similar for the α-helical and β-sheet binary patterned libraries. Therefore, the different properties of the resulting proteins ***(6,7)*** are *not* from differences in amino acid composition ***(8)***. Moreover, the observed differences are *not* from differences in sequence length: irrespective of length, sequences with the P-N-P-P-N-N-P periodicity form α-helical proteins, whereas those with the P-N-P-N-P-N-P periodicity form β-sheet proteins. Thus, it is the binary patterning itself that dictates whether a *de novo* protein forms α or β structure.
2. The binary codons, VAN and NTN, encode six polar amino acids (glutamate, aspartate, lysine, asparagine, glutamine, and histidine), and five nonpolar amino acids (valine, methionine, isoleucine, leucine, and phenylalanine), respectively. In addition to these 11 variable amino acids, a variety of other residues can be incorporated into the constant regions of the sequences. For example, our recent

library of 102 residue four-helix bundles contains 17 of the 20 amino acids ***(15)***. Only alanine, proline, and cysteine were omitted. In natural proteins, alanine occurs both in surface and core positions. Thus, its role in the binary code as polar or nonpolar is somewhat ambiguous. Proline is a special case because its restricted *phi* angle makes it useful only in certain well-defined regions of structure. Cysteine should be used only in designs wherein a disulfide bond or metal binding is planned.

3. PAGE purification of synthetic oligonucleotides is essential. This reduces the likelihood of truncated oligonucleotides being incorporated into the library. Although this purification step reduces the quantity of DNA (and potentially the diversity), the quality of the genes, and resulting libraries is enhanced significantly.
4. By excluding T from the first position of the VAN (polar) codon, tyrosine codons are avoided. This is desirable because tyrosine is not a completely polar residue and frequently occurs in the hydrophobic cores of natural proteins. Therefore only the most polar residues (histidine, glutamine, asparagine, lysine, aspartate, and glutamate) are incorporated into the designed surface positions.
5. According to the International Union of Biochemistry, the degenerate base symbols and their nucleotide base compositions are as follows ***(44)***: K = G/T, M = A/C, R = A/G, S = C/G, W = A/T, Y = C/T, B = C/G/T, D = A/G/T, H = A/C/T, V = A/C/G, N = A/C/G/T.
6. Our method for constructing full-length *de novo* gene libraries relies on combinatorial assembly using libraries of shorter fragments ***(6,7)***. Among the advantages of this approach is an increase in the diversity of the full-length library. For example, four individual segment libraries containing only $10^4$ sequences per library can be combined to yield a full-length library with a theoretical diversity of $10^{16}$ sequences ***(39)***.
7. The purpose of the *Designer Dotplot* analysis of library sequences is not merely to eliminate all undesired base-pairings, because this may not be possible, but rather to identify problem regions and design assembly protocols accordingly. For example, library oligonucleotides with significant cross-annealings should be assembled separately and joined by restriction digest and ligation (*see* **Subheading 3.3.**).
8. The preselection for gene fragments that are in frame must be independent of the structure and solubility of the encoded polypeptide fragments. Because binary patterned segments of secondary structure are designed to fold only in the context of the full tertiary structure ***(8)***, it is crucial that these segments not be weeded out of a library before assembly of the full-length genes. The intein-thymidylate synthase system (**Subheading 3.4.2.**) selects for in-frame gene segment sequences, regardless of the structure or solubility of the expressed polypeptide. This is made possible by a poly-asparagine linker designed to separate the inserted gene segment from the intein-TS fusion ***(39)***. This linker permits the reporter TS enzyme to fold and function independently of the polypeptide encoded by the inserted gene segment.

## Acknowledgments

Supported by NIH R01-GM62869 (MHH). LHB was supported by a postdoctoral fellowship from the Princeton University Council on Science and Technology.

## References

1. Mandecki, W. (1990) A method for construction of long randomized open reading frames and polypeptides. *Protein Eng*. **3,** 221–226.
2. Davidson, A. R. and Sauer, R. T. (1994) Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA* **91,** 2146–2150.
3. Davidson, A. R., Lumb, K. J., and Sauer, R. T. (1995) Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol*. **2,** 856–864.
4. Prijambada, I. D., Yomo, T., Tanaka, F., Kawama, T., Yamamoto, K., Hasegawa, A., et al. (1996) Solubility of artificial proteins with random sequences. *FEBS Lett*. **382,** 21–25.
5. Keefe, A. D. and Szostak, J. W. (2001) Functional proteins from a random sequence library. *Nature* **410,** 715–718.
6. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M., and Hecht, M. H. (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262,** 1680–1685.
7. West, M. W., Wang, W., Patterson, J., Mancias, J. D., Beasley, J. R., and Hecht, M. H. (1999) *De novo* amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. USA* **96,** 11211–11216.
8. Xiong, H., Buckwalter, B. L., Shieh, H. M., and Hecht, M. H. (1995) Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Natl. Acad. Sci. USA* **92,** 6349–6353.
9. Hecht, M. H., Das, A., Go, A., Bradley, L. H., and Wei, Y. (2004) De novo proteins from designed combinatorial libraries. *Protein Sci*. **13,** 1711–1723.
10. Taylor, S. V., Walter, K. U., Kast, P., and Hilvert, D. (2001) Searching sequence space for protein catalysts. *Proc. Natl. Acad. Sci. USA* **98,** 10596–10601.
11. Roy, S., Ratnaswamy, G., Boice, J. A., Fairman, R., McLendon, G., and Hecht, M.H. (1997) A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties. *J. Am. Chem. Soc*. **119,** 5302–5306.
12. Roy, S., Helmer, K. J., and Hecht, M. H. (1997) Detecting native-like properties in combinatorial libraries of *de novo* proteins. *Folding Des*. **2,** 89–92.
13. Roy, S. and Hecht, M. H. (2000) Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids. *Biochemistry* **39,** 4603–4607.
14. Rosenbaum, D. M., Roy, S., and Hecht, M. H. (1999) Screening combinatorial libraries of *de novo* proteins by hydrogen-deuterium exchange and electrospray mass spectrometry. *J. Am. Chem. Soc*. **121,** 9509–9513.
15. Wei, Y., Liu, T., Sazinsky, S. L., Moffet, D. A., Pelczer, I., and Hecht, M. H. (2003) Stably folded and well-ordered structures from a designed combinatorial library of *de novo* proteins. *Protein Sci*. **12,** 92–102.

16. Wei, Y., Kim, S., Fela, D., Baum, J., and Hecht, M. H. (2003) Solution structure of a de novo protein from a designed combinatorial library. *Proc. Natl. Acad. Sci. USA* **100,** 13270–13273.
17. Xu, G., Wang, W., Groves, J. T., and Hecht, M. H. (2001). Self-assembled monolayers from a designed combinatorial library of *de novo* β-sheet proteins. *Proc. Natl. Acad. Sci. USA* **98,** 3652–3657.
18. Brown, C .L., Aksay, I. A., Saville, D. A., and Hecht, M. H. (2002) Template-directed assembly of a *de novo* designed protein. *J. Am. Chem. Soc.* **124,** 6846–6848.
19. Hecht, M. H. (1994) *De novo* design of β-sheet proteins. *Proc. Natl. Acad. Sci. USA* **91,** 8729–8730.
20. Richardson, J. S. and Richardson, D. C. (2002) Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci. USA* **99,** 2754–2759.
21. Wang, W. and Hecht, M. H. (2002) Rationally designed mutations convert *de novo* amyloid-like fibrils into soluble monomeric β-sheet proteins. *Proc. Natl. Acad. Sci. USA* **99,** 2760–2765.
22. Richardson, J. S. and Richardson, D. C. (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science* **240,** 1648–1652.
23. Hutchinson, E. G. and Thornton, J. M. (1994) A revised set of potentials for β-turn formation in proteins. *Protein Sci.* **3,** 2207–2216.
24. Hirel, P. H., Schmitter, M. J., Dessen, P., Fayat, G., and Blanquet, S. (1989) Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. *Proc. Natl. Acad. Sci. USA* **86,** 8247–8251.
25. Dalboge, H., Bayne, S., and Pedersen, J. (1990) *In vivo* processing of N-terminal methionine in *E. coli*. *FEBS Lett.* **266,** 1–3.
26. Tsunasawa, S., Stewart, J. W., and Sherman, F. (1985) Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase. *J. Biol. Chem.* **260,** 5382–5391.
27. Huang, S., Elliott, R. C., Liu, P. S., Koduri, R. K., Weickmann, J. L., Lee, J. H., et al. (1987) Specificity of cotranslational amino-terminal processing of proteins in yeast. *Biochemistry* **26,** 8242–8246.
28. Bowie, J. U. and Sauer, R. T. (1989) Identification of C-terminal extensions that protect proteins from intracellular proteolysis. *J. Biol. Chem.* **264,** 7596–7602.
29. Parsell, D. A., Silber, K. R., and Sauer, R. T. (1990) Carboxy-terminal determinants of intracellular protein degradation. *Genes Dev.* **4,** 277–286.
30. Milla, M. E., Brown, B. M., and Sauer, R. T. (1993) P22 Arc repressor: enhanced expression of unstable mutants by addition of polar C-terminal sequences. *Protein Sci.* **2,** 2198–2205.
31. Berezovsky, I. N., Kilosanidze, G. T., Tumanyan, V. G., and Kisselev, L. L. (1999) Amino acid composition of protein termini are biased in different manners. *Protein Eng.* **12,** 23–30.
32. Shoemaker, K. R., Kim, P. S., York, E. J., Stewart, J. M., and Baldwin, R. L. (1987) Tests of the helix dipole model for stabilization of alpha-helices. *Nature* **326,** 563–567.

33. Chou, P. Y. and Fasman, G. D. (1978) Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47,** 251–276.
34. Fasman, G. D. (1989) Prediction of Protein Structure and the Principles of Protein Conformation. Plenum, New York.
35. Creighton, T. E. (1993) Proteins: Structures and Molecular Properties (2nd ed.). Freeman, New York.
36. Pace, C. N. and Scholtz, J. M. (1998) A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75,** 422–427.
37. DeBoer, H. A. and Kastelein, R. A. (1986) Biased codon usage: an exploration of its role in optimization of translation, in *Maximizing Gene Expression* (Rezinikoff, W. and Gold, L., eds.) Butterworth, Stoneham, MA, pp. 225–285.
38. Kane, J. F. (1995) Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* **6,** 494–500.
39. Bradley, L. H., Kleiner, R. E., Wang, A. F., Hecht, M. H., and Wood, D. W. (2005) An intein-based genetic selection enables construction of a high-quality library of binary patterned *de novo* sequences. *Protein Eng. Des. Sel.* **18,** 201–207.
40. Maizel, J. V., Jr. and Lenk R. P. (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Nat. Acad. Sci. USA* **78,** 7665–7669.
41. Tinoco, I., Jr., Uhlenbeck, O. C., and Levine, M. D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature* **230,** 363–367.
42. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415.
43. Hofacker, I .L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.* **31,** 3429–3431.
44. Cornish-Bowden, A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.* **13,** 3021–3030.