Design, Synthesis, and Amplification of DNA Pools for In Vitro Selection

This unit describes the design, synthesis, and amplification of a random sequence DNA pool. Functional nucleic acid–binding or catalytic species can be selected from these random sequence pools. In designing the DNA pool, careful consideration should be given both to the degree of randomization and the length of the random sequence region (see Strategic Planning). Following pool design, chemical synthesis on a commercial DNA synthesizer will yield a single-stranded DNA pool. The newly synthesized oligonucleotide pool can then be purified (see Basic Protocol 1). Prior to amplification, the initial complexity of the pool should be determined (see Support Protocol 1), the skewing of the pool should be determined (see Support Protocol 2), and amplification reaction conditions should be optimized (Support Protocol 3). If the nascent synthetic oligonucleotide is judged to be suitable for large-scale amplification, it can be enzymatically converted into a double-stranded DNA library (see Basic Protocol 2). Multiple copies of a single-stranded DNA pool can be derived from each double-stranded DNA library, or the library

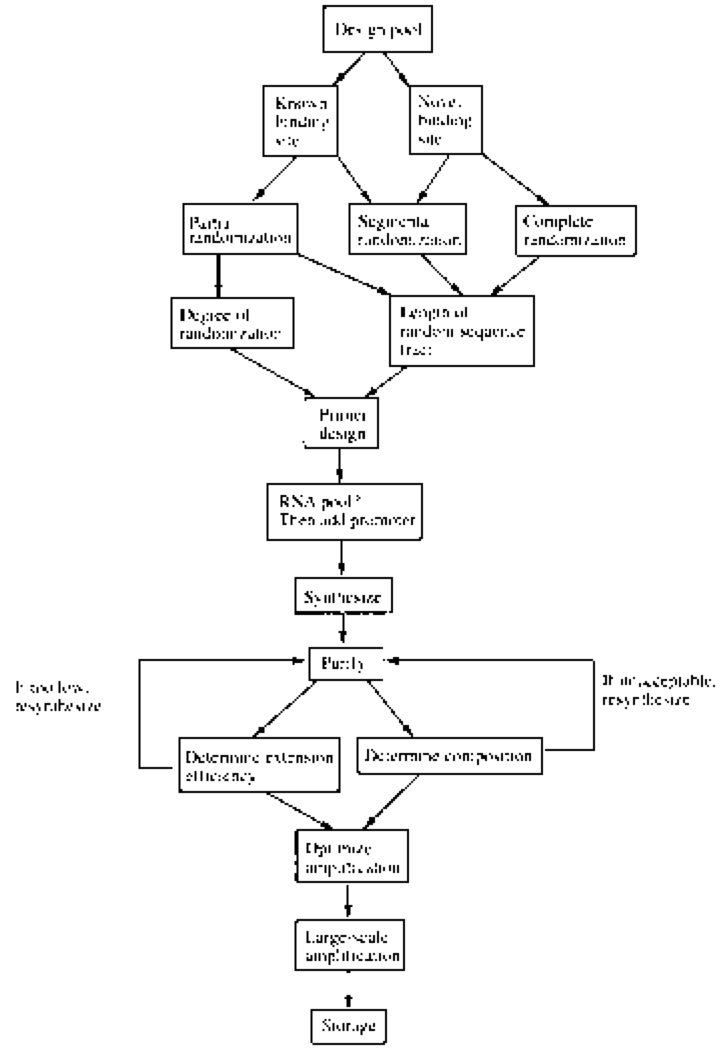

Figure 9.2.1 Flow chart outlining pool design, synthesis, and large-scale amplification.

Contributed by Jack Pollard, Sabine D. Bell, and Andrew D. Ellington

Current Protocols in Nucleic Acid Chemistry (2000) 9.2.1-9.2.23

can be transcribed to yield a RNA pool or a modified RNA pool (see *UNIT 9.3*). Figure 9.2.1 outlines the procedure.

STRATEGIC PLANNING

Designing the Initial DNA Pool

The nucleic acid pools used for in vitro selection experiments typically contain a randomized central core flanked by constant sequences that are required for enzymatic manipulations, such as PCR amplification, in vitro transcription, or restriction digestion (see also Fig. 9.2.2).

Since a pool is relatively expensive to synthesize, both in terms of time and cost, some effort should be devoted to pool design. There are many subtle parameters to consider that can greatly influence the outcome of a selection experiment, including the degree of randomization, pool length, and pool modularity (see Table 9.2.1 for references to selection experiments that have previously been successfully executed with different types and sizes of pools).

Type of selection and degree of randomization

Most researchers who carry out in vitro selection experiments wish to either better define or optimize a known binding site (binding-site selection), or to identify a novel binding site (aptamer selection). Each of these tasks in turn requires the synthesis of different types of pools. The sequences and structures that contribute to known binding sites are frequently best defined by selections that start from partially randomized pools. One example of binding-site definition that started from a partially randomized pool was a selection that defined critical residues of the Rev-responsive element (RRE) of HIV-1 Rev (Bartel et al., 1991). This experiment is also described in more detail below. Biased pools can also be used for the optimization of a previously isolated motif. For example, aptamers that could bind to the Rex protein of HTLV-1 were selected from a partially randomized pool based on the wild-type Rex-binding element (XBE) but in the end bound Rex 9-fold better than the XBE (Baskerville et al., 1995).

In contrast, completely random sequence pools explore a much wider swath of sequence space and are more useful for the isolation of novel binding species (aptamers) or catalytic species (Breaker, 1997; Jaeger, 1997). There are many examples of the selection of novel binding sites from completely random sequence pools (reviewed in Gold et al., 1995; Osborne and Ellington, 1997). Even when a natural binding site is known in advance, a completely different binding site may be selected from a random sequence pool; for

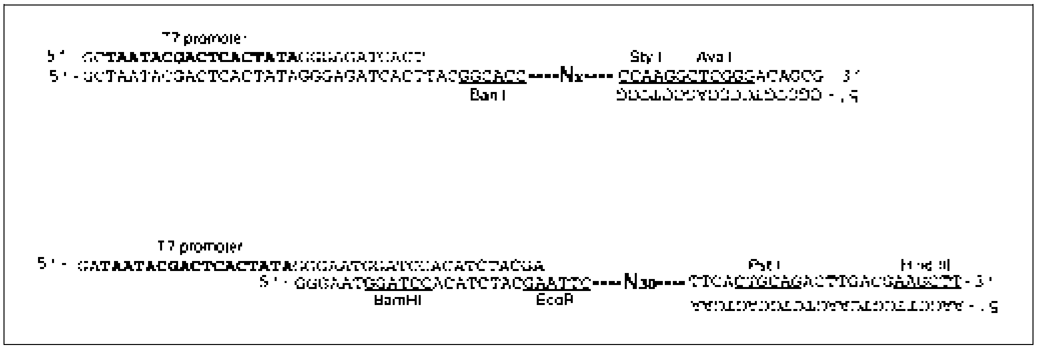
T7 promoter
5' GCTAATACGACTCACTATAGGGAGATCACT
Sty I
Ava I
5'-GCTAATACGACTCACTATAGGGAGATCACTTACGGCACC----Nx----CCAAGGCTCGGGACAGCG 3'
Ban I
T7 promoter
5'- GATAATACGACTCACTATAGGGAATGGATCCACATCTACGA
Pst I
Hind III
5'-GGGAATGGATCCACATCTACGAATTC----N30----TTCACTGCAGACTTGACGAAGCTT-3'
BamHI
EcoR

Figure 9.2.2 Two examples of pools used in in vitro selection. Primers are shown above and below the sequence of the pool. The T7 promoter is delineated in bold. Restriction sites are underlined, with their enzymes listed.

Table 9.2.1 Selection Experiments with Different Types and Sizes of Pools

Target	DNA/RNA	Length of random region	Reference
Bacteriophage T4 DNA polymerase	RNA	8	Tuerk and Gold (1990)
HIV-1 Rev	RNA	66, doped (65% wild type, 30% non-wild type, 5% deleted)	Bartel et al. (1991)
Ribozyme	RNA	120	Bartel and Szostak (1993)
HIV-1 Rev	RNA	30	Tuerk and MacDougal-Waugh (1993)
HIV-1 Rev	RNA	4 and 6, segmental; 6-9 and 6-9, segmental	Giver et al. (1993)
PKCβ	RNA	120	Conrad et al. (1994)
HTLV-1 Rex	RNA	43, doped (70% wild type, 30% non-wild type)	Baskerville et al. (1995)

example, Tuerk and MacDougal-Waugh (1993) isolated unique binders to Rev that bound better than the wild-type RBE sequence in vitro. Completely random sequence pools can also be used to extract aptamers that bind to proteins not normally thought to bind to nucleic acids; an example of this is the selection of an RNA aptamer that bound and inhibited the β isoform of protein kinase C (Conrad et al., 1994). Completely random sequence pools can also be used for the selection of novel nucleic acid catalysts. For example, starting from a pool with a 220-position random region, Bartel and Szostak (1993) isolated a novel ribozyme capable of RNA ligation. Generally, selections for catalysis require pools with a random region greater than 90 residues, while binding selections use pools with a random region of less than 70 residues.

Intermediate between partially random and completely random sequence pools are segmentally random sequence pools. In a segmentally random pool, short tracts of sequence are completely randomized. Segmental randomization thus allows all possible sequences within a short region or set of residues to be examined. Thus, if a natural binding site is known, but a portion of that binding site is suspected to be particularly important for function, then a segmentally random pool can be used to identify all possible, functional sequences within the wild-type sequence context. For example, Tuerk and Gold (1990) selected aptamers that bound T4 DNA polymerase from a pool that contained 8 random sequence positions flanked by wild-type residues. Similarly, many binding sites are known to be presented within a particular structural context, such as a stem-loop or stem-bulge structure. In these cases, a portion of the structure can be completely randomized, and all possible functional stem-loops or stem-bulges can be identified. For example, the Rev-binding element was known to form a stem-internal loop-stem structure. Giver et al. (1993) segmentally randomized only the internal loop portion of the structure and selected Rev-binding species. Many of the anti-Rev aptamers had sequences that were significantly different than the wild-type, yet were still presented in the context of a stem-internal loop-stem structure.

Partially random (doped) pool design (binding site selection)

The most important issue in the synthesis of a doped pool is the level of randomization (the probability of sequence substitution/position). As a general rule, the substitution frequency of a doped pool should roughly correspond to the number of positions thought to be required for function. For example, if 10 residues within a nucleic acid binding site are thought to be functional, then the rate of substitution might be set to yield single mutants at least half the time. If the substitution frequency is set too low, there may be too few varying residues or combinations of residues to yield information about functional sequences or structures. In contrast, if the substitution frequency is set too high, the sequence space nearest the wild-type motif will only be sparsely sampled, and many of the highly mutated molecules may be nonfunctional because their sequences will have diverged too far from the wild-type.

For example, an in vitro genetic analysis has been used to uncover the critical structural interactions between the HIV-1 Rev protein and its primary RNA binding site, the Rev-binding element (Bartel et al., 1991). The RBE had previously been mapped by deletion analysis to a short segment of HIV-1. Bartel and his co-workers assumed that the minimal RBE was smaller even than the region identified by deletion analysis, and thus decided to heavily dope a portion of a 66-nucleotide sequence at a frequency of 35% substitution/position. The initial RRE library contained $\sim 10^{13}$ molecules that had an average of 23 substitutions/template (0.35 probability substitution/position × 66 positions = ~23 substitutions); less than 1 in 10^{12} molecules were completely wild-type. Following selection, a 20-nucleotide core-binding site within the 66-nucleotide pool was readily defined by sequence conservations and covarying residues. A lower substitution rate might not have precisely defined the relatively small binding site, while an even higher substitution rate might have created a mutational load that would have limited the selection of functional molecules or even have allowed the selection of novel, non-wild-type anti-Rev aptamers (Giver et al., 1993; Tuerk and MacDougal-Waugh, 1993). Conversely, if the binding site were larger than originally hypothesized, the relatively high rate of substitution might have meant that few functional molecules could have survived the selection unscathed.

The number and type of sequence substitutions, as opposed to the probable target size for mutation, can also be used to plan the synthesis of a doped sequence pool, as described by the following equations. Typically, a 1-µmol synthesis of a 100-residue template yields a pool of $\sim 10^{15}$ amplifiable molecules. Regardless of the degree of partial randomization or the precise doping strategy employed, the number of different mutational combinations is given by:

$$3^n\{L!/[n!(L-n)!]\}$$

where n is the number of sequence substitutions/template in a template of length L. For example, in the case of the 66-nucleotide RRE pool discussed earlier, there were $\sim 2.17 \times 10^9$ possible 5-residue substitutions and $\sim 1.25 \times 10^{16}$ possible 10-residue substitutions.

To calculate what fraction of a given set of substitutions are actually contained in a doped pool, the binomial probability distribution can be used:

$$P(n,L,f) = \{L!/[n!(L-n)!]\}(f^n)(1-f)^{(L-n)}$$

where P is the fraction of the template population when f is the probability of substitution/position. If primarily single-base substitutions are desired, then f should be maximized for $n = 1$; if multiple mutations (e.g., double or triple substitutions) are desired, then f should be correspondingly higher. If the doping strategy is optimized for n substitutions, then this number of substitutions will occur most frequently, "$n - 1$" and "$n + 1$" substitutions will occur less frequently but in roughly equal numbers, and so forth. Higher levels of sequence substitution skew the mutant frequency distribution, allowing the sampling of some regions of sequence space at the exclusion of others (Fig. 9.2.3).

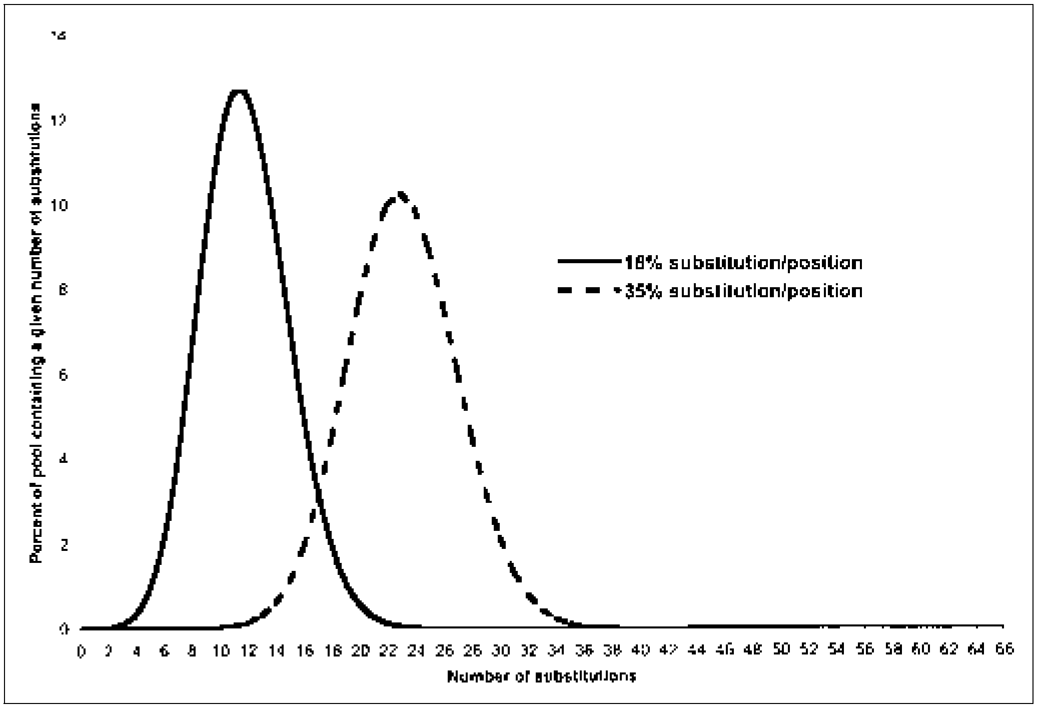

Figure 9.2.3 Comparison of substitution distributions for a 66-nucleotide pool doped to either 18% or 35%.

Therefore, in the RRE example already cited, a pool of 1 × 10^{13} molecules doped at a frequency of 35% would contain few 5-residue substitutions [1 × 10^{13} × $P(5,66,0.35)$ = ~1.82 × 10^{6} 5-residue substitutions out of ~2.17 × 10^{9} possible 5-residue substitutions]. In contrast, if the pool were doped at a frequency of 18%, all 5-residue substitutions would almost certainly be included [1 × 10^{13} × $P(5,66,0.18)$ = ~9.3 ×10^{10} 5-residue substitutions]. Note that in a pool of only 1 × 10^{13} total molecules, neither doping scheme would yield all possible 10-residue substitutions.

Completely random pool design (aptamer selection)

Completely random sequence pools are used to initiate selection experiments when no functional nucleic acid sequence or structural motif is known in advance. There is really only one parameter to consider when designing a completely random pool: the length of the random region. While we will consider this parameter in detail below, we must first dismiss a frequent bogey of selection neophytes, the issue of complexity and representation.

Random sequence space is a vast landscape of possibilities of which only a vanishingly small fraction can be sampled by either nature or man. Assuming a 4-monomer repertoire from which pools can be constructed, there are ~1.6 × 10^{60} unique individual sequences in a sequence space bounded by a 100-residue template (4^{100} = ~1.6 × 10^{60}), a quantity of nucleic acid greater than an Avogadro's number of Earth masses. While this grotesquely large value is clearly beyond the realm of experimental possibility, modern methods of chemical nucleic acid synthesis do allow the sampling of nearly as much sequence information as may be contained in the Earth's biosphere. As a back-of-the-envelope calculation, consider that there are on the order of ~1× 10^{9} species in the biosphere, each with ~1 × 10^{5} genes. If each of these genes in turn is composed of ~1 × 10^{3} residues, then there are ~1 × 10^{17} residues worth of information in a biosphere. In contrast, a typical 1-µmol synthesis of a 100-residue random sequence pool would contain 1 × 10^{15}

molecules × ~1 × 10^2 residues/molecule = ~1 × 10^{17} unique residues or roughly 1 biosphere's worth of information. Obviously, the connection and ordering of sequence information in organisms is important as well.

Typically, a random sequence pool contains ~1 × 10^{15} molecules, and thus can potentially sample on the order of all possible 25-mers (4^{15} = ~1.1 × 10^{15}). In fact, since different 25-mers can be found in different "reading frames," a slightly larger sequence space will likely be sampled. Because of this physical restriction, it is sometimes thought that random sequence pools should be no more than 25 residues in length—any longer, and only a fractional sampling would be possible, and many potential sequences would be lost. While this is true, it should be realized that longer pools do not lose any of the numerical complexity of smaller pools (except in those instances where long syntheses are extremely inefficient) and in fact gain access to some fraction of longer sequence and structural motifs as well. For example, tRNA molecules are roughly 76 nucleotides in length. It might prove more difficult to select tRNA mimics from a random sequence population containing 30 randomized residues than from a pool spanning 80 randomized residues. However, any short functional tRNA mimics present in the shorter population should also be present in equal or greater number in the longer population. In most instances, the relative completeness of the pool is not a consideration in the success of a selection. Indeed, it has been shown that functional nucleic acids are not extremely rare (for recent reviews see Gold et al., 1995; Fitzwater and Polisky, 1996) and can be isolated both from "complete" pools that span 20 random sequence positions and from very "incomplete" pools that span 90 random sequence positions.

Having dismissed considerations of complexity and representation, the one guiding principle that emerges from this analysis is that longer pools are more generally useful for selection experiments than shorter pools. However, this principle must be applied with appropriate caveats. First, aptamers derived from shorter pools are easier to analyze. Sequence and structural motifs embedded within a 30-nucleotide random sequence region are much more readily apparent than sequence and structural motifs embedded within a 90-nucleotide random sequence region, especially if the motifs are not colinear. Second, longer pools are more difficult to synthesize than shorter pools. Finally, longer pools are more likely to yield amplification or other selection artifacts than shorter pools. For example, pools that contain random regions greater than 90 nucleotides in length can form self-aggregates that precipitate from solution upon prolonged incubation, and thus require immobilization on a solid support prior to selection (Bartel and Szostak, 1993; Lorsch and Szostak, 1994). Because of these considerations, pools used for the in vitro selection of aptamers typically contain from 20 to 80 random sequence positions.

Longer pools are not only desirable but are likely required in selections for complex functions, such as catalysis. Pools used for the selection of ribozymes typically contain from 50 to 220 random sequence positions (for recent reviews see Gold et al., 1995; Fitzwater and Polisky, 1996). The optimal length of the random region is an active area of research (Sabeti et al., 1997) where many of the fundamental parameters remain to be defined. Practically, though, longer pools must be synthesized as oligonucleotides of 150 residues or fewer in length because of the constraints of DNA synthetic chemistry. For this reason, pools longer than 150 bases are typically generated in a modular fashion by ligating together individual, synthetic oligonucleotides (Bartel and Szostak, 1993). Segments of shorter DNAs can be stitched together by the inclusion of unique restriction sites (Bartel and Szostak, 1993). Asymmetric restriction sites, such as *Ava*I (C|YCGRG), *Ban*I (G|GYRCC), and *Sty*I (C|CWWGG), are very useful for this task since they minimize intra-pool dimerization via self-ligation. Also, these enzymes are cost-effective for digesting large amounts of DNA. Alternatively, an overlapping region can be included at the 3′ end of each synthetic oligonucleotide and mutually primed synthesis (e.g., *CPMB UNIT 8.2*) of a longer template can be carried out. After assembling pool modules, the complexity (yield) of the new, aggregate pool will need to be freshly assessed. The upper bound of the complexity of an assembled pool (e.g., 10^{11} 100-mer modules × 10^{11} 100-mer modules) will likely be much larger than its actual complexity (e.g., 100 micrograms of ligated 200-mer, 9.12×10^{14} molecules).

Segmentally random pool design (binding site and aptamer selection)

In general, the rules governing the design of segmentally random pools are idiosyncratic, depending on experimental purpose. If the desire is to better define a known binding site, then relatively short sequence tracts (i.e., from four to ten residues) should be completely randomized. The randomization of longer sequence tracts may lead to the selection of novel binding sites rather than variants of a known binding site. The residues can either be colinear (as is the case for many DNA binding sites) or dispersed (as is the case for many RNA binding sites). If the desire is to identify a binding site within the context of a known structural element, then from four to twenty residues can be completely randomized. In this instance, the fewer the number of residues that are randomized, the more likely it will be that the selected sequences will resemble a wild-type binding site or retain an engineered structure. The greater the number of residues that are randomized, the more likely it will be that a novel aptamer sequence or structure will be discovered.

Primer design

Generally, the constant sequences at the 5′ and 3′ ends of a pool function as primer-binding sites and can be almost any sequence or length. Primers of 20 nucleotides in length are convenient because their melting temperatures are convenient for the PCR and they can easily be synthesized in high yields. In designing constant sequences and complementary primers, obvious artifacts associated with the PCR, such as secondary-structure formation or self-association that could lead to the production of primer dimers, should be avoided. Computer programs such as the Genetic Computer Group's PRIME or the Whitehead's PRIMER3 assist in designing constant regions. Other primer design programs include Amplify (Bill Engels, Dept. of Genetics, University of Wisconsin, Madison) and Oligo (National Biosciences). As a rule of thumb, one should try to avoid using the same triplet sequence more than once in either constant region.

Beyond these basal considerations, there are two schools of thought regarding the sequence of the priming site itself. On the one hand, designing primers to possess a 3′ clamp of 5′-WSS-3′ (IUB codes: W = A or T, S = C or G), such as ACC, ensures good extension by polymerases. On the other hand, the inclusion of A/T-rich regions at the 3′ termini of primers reduces the frequency of mispriming and allows virtually "infinite" multiplication of DNA amplicons (Crameri and Stemmer, 1993). The inclusion of restriction sites within primer regions can facilitate cloning of selected nucleic acids, although palindromes adjacent to the 3′ ends can also facilitate the genesis of primer-dimers.

Finally, primers for partially randomized pools should be designed so that they do not conflict with the folding or accessibility of a known DNA or RNA binding site. It is suggested that the secondary structure of the wild-type binding site with any appended primer-binding sites be determined using an algorithm such as Mulfold (Jaeger et al., 1989). If the native or wild-type structure of the binding site is not among the most common folds, then the primers should be redesigned.

If an RNA pool is to be constructed, runoff RNA transcripts for in vitro selection are frequently made with T7 RNA polymerase. There are several known promoters for T7

RNA polymerase (Milligan et al., 1987), but the following minimal sequence gives good yields:

```
 −17                    −1
5′-TAA-TAC-GAC-TCA-CTA-TA-3′
```

Addition of a G and C residue at the −18 and −19 positions of the minimal promoter helps to close the DNA duplex and stabilize the 5′ end of the promoter region, thereby increasing transcriptional yields. Transcription initiation is optimal when there are stretches of purines in the +1 and +2 positions, with GG being the best initiator (Milligan et al., 1987). Transcriptional yields also increase if uridine does not appear in the transcript before position 6. Typical pool designs incorporating all the elements described are shown in Figure 9.2.2.

Chemically Synthesizing the Pool

While pools of genomic DNA sequences have been used for selection (Singer et al., 1997), partially or completely random sequence pools must be chemically synthesized. Modern DNA synthesizers utilize phosphoramidite chemistry (*UNIT 3.3*) or H-phosphonate chemistry (*UNIT 3.4*) and can routinely produce usable amounts of DNA up to 150 nucleotides in length. Longer oligonucleotides can also be synthesized, but side reactions such as branching and depurination accumulate throughout the synthesis and the amount of final, usable product recovered can be vanishingly small. Since stepwise coupling efficiencies for a long oligonucleotide are on average ≥98%, the typical yield of a 100-base synthesis that starts with a 1-µmol column is 13.5%, or 13.5 nmol, or 1×10^{16} different molecules, of which ~10% to 30% can be enzymatically elongated or amplified. Several strategies can be used to enhance the synthetic yield of oligonucleotides that are longer than 100 bases (see *APPENDIX 3C*). Further, if a pool longer than ~150 nucleotides is desired, smaller pools can be modularly synthesized and coupled by ligation or mutually-primed synthesis (see discussion of completely random pool design, above). During synthesis it is wise to prevent the cross-contamination of primers with their corresponding pool. It has recently been discovered (A. Friedman, pers. comm.) that when pools and primers are synthesized on identical ports of a DNA synthesizer, there is some mixing of the molecules. The contamination is sufficient to yield a positive signal following extensive (30 to 50 PCR cycles) amplification of a no-template negative control. The unprogrammed interleaving of pools and primers can lead to extreme skewing of amplified materials, such that only a few species from the original pool may comprise a significant fraction of a subsequent amplification reaction. Therefore, pools and their cognate primers should be synthesized on different synthesizer ports and/or the machine should be extensively flushed with acetonitrile between syntheses.

Most synthesizers can be programmed for in-line, degenerate mixing of bases. While this method is useful when only a few positions must be randomized, because of the extremely fast reaction of the activated phosphoramidite with the newly deprotected 5′ hydroxyl, random sequences will be skewed towards the phosphoramidite that first enters the column. Therefore, for longer pools or pools that should contain a statistically random distribution of nucleotides, it is better to manually mix the phosphoramidites off-line and use this mixture for the synthesis of degenerate sequence positions. A more stochastic distribution can be obtained by including larger amounts of A and C phosphoramidites in the mix to compensate for the faster coupling times of G and T phosphoramidites (Zon et al., 1985). Suggested ratios include a 3:3:2:2.4 molar ratio of A:C:G:T phosphoramidites (D.P Bartel, pers. comm.), and a 1.5:1.25:1.15:1 molar ratio of A:C:G:T (see User’s Manual for PE Biosystems Models 392 and 394 DNA/RNA Synthesis).

Table 9.2.2 Representative Calculations Based on the Masses and Efficiencies for Couplings that Utilize the Canonical Tetrazole Activation Chemisty and Phosphoramidites Bearing Standard Protecting Groups

Phosphoramidite	Molecular mass (g/mol)	Mass correction	Coupling efficiency correction	Overall correction
5′-CE-dA	858	0.87	0.67	0.58
5′-CE-dC	834	0.89	0.67	0.60
5′-CE-dG	840	0.89	1.00	0.89
5′-CE-dT	745	1.00	0.83	0.83

Doped pools are perhaps the most difficult to synthesize (Hermes et al., 1989; Bartel et al., 1991). Doping can be accomplished by using phosphoramidite mixtures that have been adjusted to ensure the proper level of partial randomization of a given nucleotide. For example, if a doped pool is to be synthesized in which non-wild-type residues are included at a rate of 10%/position, then for the adenosine bottle a molar ratio of 33.43:1.50:1.00:1.21 of A:C:G:T phosphoramidites should be used. These ratios were derived by first adjusting for the relative molecular mass and coupling differentials of the individual phosphoramidites and then mixing the phosphoramidite solutions on a percent volume basis to yield the desired extent of doping. This process is described in greater detail below.

To normalize the coupling of different phosphoramidites, relative correction factors that take into account different coupling efficiencies and molecular masses must be calculated. Multiplying together these correction factors gives an overall correction factor to provide equal molar coupling of each phosphoramidite. Table 9.2.2 displays representative calculations based on the masses and efficiencies for couplings that utilize the canonical tetrazole activation chemistry (*UNIT 3.3*) and phosphoramidites bearing standard protecting groups [cyanoethyl for the phosphates along either isobutyryl (N-2 of guanine) or benzoyl (N-6 of adenine and N-4 of cytidine) groups; see *UNIT 2.1*]. Other chemistries and protections may require the substitution of other correction factors.

Most modern synthesizers require that ~1 g of phosphoramidite be dissolved in ~20 mL of acetonitrile to be used in the coupling reaction. Applying this constraint along with the combined mass-coupling (overall) correction factor gives the volumes shown in Table 9.2.3 to dissolve 1 g of each phosphoramidite. Therefore, if equal volumes of each of these solutions are mixed, equal molar coupling should occur since the molar concentrations have been adjusted to account for both the mass and coupling differentials.

As in the example above, if a doped pool is to be synthesized in which non-wild-type residues are included at a rate of 10%/position, then the amidites should be mixed as in Table 9.2.4.

Table 9.2.3 Volumes of Acetonitrile Needed to Dissolve 1 g of Phosphoramidite

Phosphoramidite	Dissolved in X mL of acetonitrile
5′-CE-dA	11.6
5′-CE-dC	12.0
5′-CE-dG	17.8
5′-CE-dT	16.6

Table 9.2.4 Amidite Mixtures for Synthesis of Doped Pool in Which Non-Wild-Type Residues are Included at a Rate of 10% /Position

Phosphoramidite	Mutagenesis (%)	Total Volume (mL)	Volume each amidite to mix (mL)			
			A	C	G	T
A	10	10	9.00	0.33	0.33	0.33
C	10	10	0.33	9.00	0.33	0.33
G	10	10	0.33	0.33	9.00	0.33
T	10	10	0.33	0.33	0.33	9.00
A	20	10	8.00	0.67	0.67	0.67
C	20	10	0.67	8.00	0.67	0.67
G	20	10	0.67	0.67	8.00	0.67
T	20	10	0.67	0.67	0.67	8.00

In addition to varying nucleotide composition, it is also possible to vary the length of random sequence that is synthesized. Deletions can be stochastically incorporated during a synthesis by replacing the capping step with an acetonitrile wash (Bartel et al., 1991). It is more difficult to stochastically incorporate insertions, but the lengths of segmental random sequences in a pool can be mixed. For example, in Giver et al. (1993), four columns were used to generate a pool with two random regions of 6 to 9 positions separated by a constant domain. The first column was synthesized with 6 random positions, the second with 7 random positions, etc. Following the addition of the intervening constant sequence, the synthesis was stopped, the four columns were opened, and the resins from the four columns were mixed. The mixed resins were then equally redivided into four new columns and the synthesis was resumed. The first column incorporated 6 positions, the second column 7 positions, etc. Thus, the first column contained oligonucleotides in which the first random segment was 6, 7, 8, or 9 residues long and a second random segment that was uniformly 6 residues long. The second column contained oligonucleotides in which the first random segment was 6, 7, 8, or 9 residues long and a second random segment was uniformly 7 residues long, and so forth. Following the completion of all four syntheses, the reactions were combined to generate the final random sequence pool.

BASIC PROTOCOL 1

PURIFICATION OF A RANDOM SEQUENCE POOL

A newly synthesized oligonucleotide pool should be purified on a denaturing polyacrylamide gel (see e.g., *CPMB UNIT 2.12*) prior to amplification. Oligonucleotides can also be purified using an HPLC or commercially available spin columns, but HPLC purification is not recommended for ssDNA pools, due to concerns about cross-contamination. Since oligonucleotides of equivalent length but different sequence migrate at slightly varying rates (see User's Guide for PE Biosystems Expedite Nucleic Acid Synthesis System), a pool should appear as a broader band than a homogeneous sequence. In fact, because of the presence of capped failure sequences and depurinated, cleaved fragments, it is likely that the oligonucleotide product will appear even more heterogeneous.

As a general note, since sequences exist as single copies prior to amplification, individual species can be easily lost. Therefore, it is important to wash and elute the various filters, tubes, and tips described below one or more times. The eluates can then be pooled for a final precipitation and eventual amplification.

Contamination of primers or other solutions with a synthesized or isolated pool should be avoided by using aerosol barrier tips. Similarly, gel plates used during purification should be washed thoroughly to ensure that they are free of contamination with other pools or primers.

Materials

DNA pool
Ammonium hydroxide
n-butanol
TE buffer, pH 8.0 (*APPENDIX 2A*)
Urea loading buffer, 2× (*APPENDIX 2A*)
5 M NaCl
Ethanol

Fluorescent TLC plate (VWR), wrapped in plastic wrap
UV lamp
Razor blades
Small-bore syringes
13-mL centrifuge tubes capable of withstanding temperature extremes (Sarstedt)
90°C water bath
Rotary shaker

Additional reagents and equipment for denaturing polyacrylamide gel electrophoresis (e.g., *APPENDIX 3B* or *CPMB UNIT 2.12*)

1. After synthesis, deprotection, and cleavage from the solid support, lyophilize the oligonucleotide solution (in concentrated ammonium hydroxide) to dryness or precipitate with a 10-fold volume of *n*-butanol.

    The n-butanol precipitation can occur quite quickly at room temperature for longer oligonucleotides. Shorter (<20 base) oligonucleotides may require longer or colder incubations. To ensure more efficient recoveries of oligonucleotides it is safest to precipitate for ≥1 hr at −70°C.

2. Pour a denaturing polyacrylamide gel (e.g., *APPENDIX 3B* or *CPMB UNIT 2.12*).

    To allow for good separation of near-full-length from non-full-length products, the acrylamide concentration should be chosen so that the full-length oligonucleotide will migrate approximately one-half to three-fourths of the way into the gel by the time the loading dye reaches the bottom.

3. Resuspend the lyophilized or precipitated pellet in ~100 to 200 µL of water or buffer (e.g., TE buffer, pH 8.0) and add an equal volume of 2× loading dye. Heat denature samples at 75°C for 5 min prior to loading. Load ~20% of a 1-µmole synthesis per 2 cm × 2 cm × 1.6 mm well and perform electrophoretic separation.

4. Place gel on a fluorescent TLC plate that has been wrapped in plastic wrap and excise the oligonucleotide product from the gel with the aid of a UV lamp, using razor blades.

    The desired oligonucleotide product is generally the darkest, shadowed band on the gel (excluding UV-absorbing material that runs at the dye front). If stepwise synthetic efficiency has been low, the product will appear as a smear instead of as a clear band. Since many of the N-1, N-2, etc. products can be converted into full-length products by the polymerase chain reaction, a fairly wide band of near full-length products can be cut from the gel. The excision should be carried out relatively quickly, since unnecessarily long UV exposure can damage the oligonucleotide product.

The full-length oligonucleotide product should be the slowest-migrating band. However, if deprotection has been incomplete, lighter bands that migrate considerably above the major fully deprotected band may be observed.

Unpolymerized acrylamide absorbs strongly at 211 nm and may cause shadowing at the edges and wells of the gel. This can obscure the resolution or recovery of bands in the outer lanes.

5. Elute the oligonucleotide from the gel slices as follows.

   a. To aid in the diffusion of the oligonucleotide from the acrylamide matrix, chop gel slabs into fine particles by forcing the gel through a small-bore syringe.

   b. Place the crushed gel slabs in a 13-ml centrifuge tube capable of withstanding temperature extremes.

   c. Add 3 mL of TE buffer, pH 8.0, per 0.5 mL of gel slab (typically corresponding to one to two wells), and place the sample at −80°C for 30 min or until it is frozen solid.

   d. Quickly thaw the tube in a hot water bath and then let it soak at 90°C for 5 min. Elute the DNA overnight at room temperature on a rotary shaker.

   This freeze-rapid thaw approach (Chen and Ruffner, 1996) allows ice crystals to break apart the acrylamide matrix, increasing yield and decreasing elution time. Typically, 80% of a 20-mer oligonucleotide can be recovered after 3 hr of rotary shaking, making this technique comparable to electroelution (see, e.g., CPMB UNIT 2.7).

   Because elution is a diffusion-controlled process, higher elution volumes or serial elutions from the same gel slice can increase the amount of DNA recovered. Longer oligonucleotides diffuse from the gel more slowly than shorter sequences. Samples of especially long synthetic DNAs and RNAs that are particularly resistant to elution with aqueous buffers may be eluted more easily in 6 vol of formamide (>5 hr at room temperature), followed by a brief elution with an aqueous buffer (~1 hr). Isoamyl alcohol extraction (e.g., CPMB UNIT 2.12) can be used to bring the extracts to a convenient volume for subsequent precipitation.

6. Precipitate the eluted oligonucleotide pool by adjusting the salt concentration to 0.3 M using a 5 M NaCl stock solution, then adding 3 vol of ethanol. Keep at –20°C for 3 hr, then microcentrifuge at maximum speed 4°C. Lyophilize to dryness. Resuspend the synthetic pool in TE buffer, pH 8.0 (to protect against nuclease contamination or drastic pH changes).

   If the volume of the eluted oligonucleotide is too large to conveniently precipitate, concentrate the sample by extracting against an equal volume of n-butanol. Remove the upper butanol layer and repeat until the aqueous volume is convenient for precipitation. About 1/5 of the aqueous layer is extracted into the organic butanol layer for every volume of butanol used. If too much butanol is used, thereby completely extracting the aqueous layer into the butanol, add more water and repeat the concentration.

SUPPORT PROTOCOL 1

DETERMINING THE POOL COMPLEXITY

The number of different molecules present in a population can affect the outcome of a selection experiment (see Troubleshooting). If the pool complexity is too low for a given application, the pool will have to be resynthesized.

Pool complexity is, in turn, a function of yield and of the number of molecules in the pool that can be fully extended by a polymerase. The overall yield of the synthesis can be calculated by determining the UV absorption of the pool. However, deletions, incompletely deprotected residues, or backbone lesions that arise during chemical synthesis decrease by 10% to 40% the fraction of molecules in a synthetic pool that can be fully extended by polymerases. For example, the rate of insertions (presumably due to DMT cleavage via tetrazole) has been measured to be as high as 0.4% per position, and the rate of deletions (presumably due to incomplete capping) has been found to be as high as 0.5% per position (A. Keefe and D. Wilson, pers. comm.). The number of usable DNA molecules that are actually present in a nascent pool can be calculated by determining the fraction of the pool that can be extended by *Taq* polymerase.

Materials

Purified ssDNA pool and labeled primers
50 mM Tris·Cl, pH 7.5 (*APPENDIX 2A*)
10 mM $MgCl_2$
5 mM DTT
[γ-^{32}P]ATP (>3000 Ci/mmol)
T4 polynucleotide kinase
1:1 phenol/chloroform (*APPENDIX 2A*)
Chloroform
4.0 M ammonium acetate
Taq DNA polymerase
TE buffer, pH 8.0 (*APPENDIX 2A*)
PCR amplification buffer (*APPENDIX 2A*)
2× formamide loading buffer (*APPENDIX 2A*)
15 × 17–cm denaturing polyacrylamide gel *(APPENDIX 3B)*

Thermal cycler
Phosphor imager plate and phosphor imager

Additional reagents and equipment for quantitation of DNA (e.g., *CPMB APPENDIX 3D),* end-labeling of DNA (e.g., *CPMB UNIT 3.10*), phenol/chloroform and chloroform extraction of DNA (*APPENDIX 2A*), PCR amplification (e.g., *CPMB* Chapter 15), and denaturing polyacrylamide gel electrophoresis (*APPENDIX 3B*)

1. Quantitate DNA by UV absorption assuming that A_{260} of 1.0 indicates ~37 µg/ml of single stranded DNA.

   Also see, e.g., CPMB APPENDIX 3D.

2. Label the 5′ end of the 3′ PCR primer with [γ-^{32}P]ATP by preparing the following reaction mixture.

   For 30-µl reaction (volume of reaction and concentration of DNA and [γ-^{32}P]ATP will vary depending on application):

   50 mM Tris·Cl, pH 7.5
   10 mM $MgCl_2$
   5 mM DTT
   1 to 50 pmol dephosphorylated DNA, 5′ ends
   50 pmol (150 µCi) [γ-^{32}P]ATP
   50 µg/ml BSA
   20 U T4 polynucleotide kinase

   Incubate 60 min at 37°C, then stop reaction by adding 1 µl of 0.5 M EDTA. Phenol/chloroform and chloroform extract the labeled oligonucleotide (see recipe for phenol/chloroform/isoamyl alcohol in *APPENDIX 2A*), and precipitate by adding an equal volume of 4.0 M ammonium acetate and 2 vol ethanol. Microcentrifuge to collect the pellet, remove the supernatant, and redissolve the labeled DNA pellet in 10 µL of TE buffer, pH 8.0.

This procedure ensures that most of the unincorporated label remains in the supernatant.

3. Incubate ~50 pmol of labeled primer with a 2- to 5-fold molar excess of pool in a 50-µL extension reaction, under the same conditions that will be used in the final amplification, in a thermal cycler as follows (see, e.g., *CPMB UNIT 15.1* for PCR).

   a. Denature and anneal the primer and template DNA in PCR amplification buffer (usually 94°C for the denaturation step and ~50°C for the annealing step).

   b. Add *Taq* or other DNA polymerase (scaled to the anticipated enzyme concentration to be used in the large-scale amplification), then ramp the temperature to 72°C for 20 min.

      It may be useful to take time points to determine whether the reaction has gone to completion.

   c. Finally, terminate the reaction by the addition of 2× formamide loading buffer.

4. Heat the extension reaction to 90°C for 3 min and load the reaction on a 15 × 17–cm denaturing polyacrylamide gel with appropriate radiolabeled size markers. Electrophorese until the dye is at or near the bottom of the gel, but do not let the radiolabeled primers run off.

   It is also useful to load a separate well with an aliquot of the primer alone. Choose an acrylamide percentage that allows efficient separation of small primers from larger extended products.

5. Dry and expose the gel to a phosphor imager plate. Using a phosphor imager, quantify the control primer band and the extended product band.

   There may be a smear leading up to the extended band. One should use one's best judgement in determining how much near-full-length material will be included in the quantitation. Calculate the percent extension by dividing counts of labeled, extended product by counts of labeled primer. Percent extension for a gel-purified ssDNA pool can range from 10% to 30%. The complexity of the pool is then the yield (determined in step 1) multiplied by the extension efficiency (percent extension determined above). If the complexity of the pool is insufficient for planned experiments, then the pool must be resynthesized.

SUPPORT PROTOCOL 2

DETERMINING THE POOL BIAS

Following extension, the reaction should be repeated using a cold primer and the nonradioactive double-stranded DNA pool should be amplified in a PCR reaction, cloned (e.g., using a TA cloning kit from Invitrogen), and individual members sequenced to determine the degree of partial or completely randomness. The cloning step could also be carried out following PCR optimization (see Support Protocol 3). From 20 to 30 clones should be sequenced to determine the base composition of the starting pool. The random region should be composed of roughly 25% of each base. A pool with the random region skewed toward one or more bases (>30%) should be resynthesized.

SUPPORT PROTOCOL 3

SMALL-SCALE PCR OPTIMIZATION OF POOL AMPLIFICATION

To enhance yield and further avoid bias, the amplification conditions for a pool should be optimized prior to carrying out a large-scale amplification. Moreover, since amplifying a pool is costly in terms of both time and money, any optimization of the PCR should first take place on a small scale. The more involved large-scale amplification can then be carried out with confidence.

Materials

dNTPs (*APPENDIX 2A*)
Taq DNA polymerase (e.g., Boehringer Mannheim)
PCR amplification buffer containing 1.5 mM Mg^{2+} (*APPENDIX 2A*)
dsDNA mass markers (e.g., Life Technologies)
4% Nu Sieve agarose gell (FMC Bioproducts)

Thermal cycler
Densitometer

Additional reagents and equipment for agarose gel electrophoresis (e.g., *CPMB UNIT 2.5*)

1. Carry out a 0.1 mL PCR reaction using 2 nM of synthetic pool oligonucleotide as template, 2 µM primers, and PCR buffer with 1.5 mM magnesium. Use the manufacturer's suggested quantity of *Taq* (e.g., 2.5 U of Boehringer Mannheim *Taq*) in a reaction containing 200 µM dNTPs. A suggested temperature regime is:

10 to 15 cycles:	2 min	95°C	(denaturation)
	1 min	55°C	(annealing)
	3 min	72°C	(extension).

After 10 to 15 cycles of amplification, check the length and purity of the amplified DNA on a 4% Nu Sieve agarose gel in 1× TBE buffer (e.g., *CPMB UNIT 2.5*).

Annealing temperature may need to be adjusted to as low as 45°C depending on primer composition (e.g., for a small or AU-rich primer).

A 0.1 mL reaction typically yields ~1 µg, but the amount can vary from 0.1 to 10 µg. A fuzzy band may indicate that too many cycles of PCR have been carried out. In this case, set up the reaction again and perform fewer cycles.

2. Dilute the double-stranded PCR DNA product 1:128, and repeat the PCR reaction, removing a 5- to 10-µL aliquot during the last 10 sec of the cycle-7 extension step. Serially dilute the amplified product 1:2, 1:4, ... 1:128. Electrophorese all of the samples on a large agarose gel.

Note that it is quite difficult to accurately pipet solutions at 72°C. It may therefore be desirable to pipet an amount slightly larger than that intended for use in the serial dilution.

3. Calculate the average PCR efficiency by identifying to what extent the cycle-7 PCR reaction is the result of progressive doublings of the original synthetic DNA. Determine which dilution lanes lack detectable DNA.

The largest dilution that lacks detectable DNA is also the dilution that is a minimum estimate of the number of doublings. For example, if the 1/64 dilution is the largest dilution without detectable DNA, this implies that 6 "doublings" of the synthetic DNA yielded at least 64-fold more DNA. This is expressed as follows:

$$(\text{average efficiency})^{\text{no. of theoretical doublings (i.e., PCR cycles)}} = \text{fold increase in DNA}$$

Thus, if 7 cycles of PCR were performed, then the average number of doublings per cycle is ~1.81 [from $(\sim 1.81)^7 = 64$].

4. Modulate PCR conditions to enhance PCR efficiency.

If the pool's average number of doublings per cycle is <1.8, then the PCR conditions chosen may skew the representation of the pool. In that case PCR conditions should be modulated to enhance PCR efficiency. The following parameters or variables are most amenable to modification. It is best to begin the optimization with a single set of reaction conditions, modify individual parameters relative to this one reference reaction, and then combine all advantageous alterations into a single reaction. In addition, one may wish to consult CPMB UNIT 15.1.

Theoretically PCR can proceed until the primers or dNTPs are depleted. Therefore, primer and dNTP concentrations should be well above those used for the amplification of small amounts of DNA. Primer concentrations from 1 µM to as high as 5 µM have been used (although concentrations >5 µM are generally not helpful). It may be useful to scan both above and below 2.5 µM in 0.5-µM increments.

Magnesium concentration affects both primer annealing and the fidelity of Taq (which decreases with increasing magnesium concentration). Starting at the magnesium supplied in the PCR buffer (usually 1.5 mM), scan in 1-mM increments toward 5 mM as a maximal concentration.

DNA denaturation at temperatures above 95°C is usually impractical since this greatly reduces Taq's half-life. While other thermostable polymerases can be more resistant to higher temperatures, they usually have a lower extension efficiency and are more expensive than Taq. Annealing temperatures are dependent upon both primer sequence and length. The primer annealing temperatures should already be known from the primer design process, or may be calculated via an algorithm that can be found at http://paris.chem.yale.edu/extinct.html. This algorithm takes into account nucleotide composition, stacking energies (according to Turner's rules), and empirical data. An annealing temperature ~5°C less than the calculated annealing temperature is a good place to begin optimization. The amplification is more efficient at a lower annealing temperature, but mispriming and secondary structural problems are more pronounced. Higher temperatures improve the specificity, but decrease the overall yield of the reaction. To determine the optimum annealing temperature for a given primer and magnesium concentration, one should scan in both directions around the annealing temperature in 5°C increments. Finally, extension temperatures are modulated by the properties of Taq, which will extend (although inefficiently) at temperatures as low as 65°C. When extending at temperatures above Taq's optimum temperature (70° to 75°C) somewhat more polymerase may be required; scanning of the enzyme quantity should be done in 2.5-U increments. However, too much Taq may be harmful to structured single-stranded nucleic acids (Lyamichev et al., 1993).

5. Confirm the results of the extension reaction described in Support Protocol 1 by the optimization method as follows. After optimizing pool PCR conditions for >1.8 average number of doublings per cycle, determine the pool complexity by performing another 0.1-ml PCR reaction with 2 nM of the original, synthetic pool oligonucleotide under the now optimized reaction conditions. After 7 or more cycles of PCR, perform agarose gel electrophoresis on serial dilutions of the PCR reaction adjacent to serial dilutions of dsDNA mass markers. Calculate the amount of amplified DNA using either a densitometer or by estimating which dilutions are most similar. Calculate the approximate pool complexity as follows:

$$\frac{\text{g of PCR DNA after}N\text{ cycles of PCR}}{\text{g avg no. of doublings per cycle (see step 4)}} = \text{g of starting extendable ssDNA}$$

$$\frac{\text{g of starting extendable ssDNA}}{330\text{ g/mole} \times \text{(no. of bases in full}w\text{length product)}}$$

$$= \text{mol starting extendable ssDNA}$$

$$\text{mol starting extendable ssDNA} \times (6.02 \times 10^{23}) = \text{molecules of starting extendable ssDNA}$$

$$\frac{\text{molecules of starting extendable ssDNA}}{\text{starting molecules}} = \text{fraction of extendable ssDNA}$$

$$\text{fraction of extendable ssDNA} \times \text{no. of synthetic pool molecules} = \text{pool complexity}$$

PCR efficiency should be optimized to balance the average number of doublings per cycle against the total reaction volume. A pool of 1 × 10^{15} molecules (~1.7 × 10^{9} mol) at a starting template concentration of 2 nM will require 0.85 L for amplification. Therefore, it is greatly desirable to amplify the pool at the highest template concentration that still gives a reasonable number of doublings per cycle. The amplification should generate at least 8 copies of pool DNA if the pool complexity is to be archived and preserved (see Basic Protocol 2).

BASIC PROTOCOL 2

LARGE-SCALE PCR AMPLIFICATION OF POOL DNA

Very long and complex pools often require PCR amplification on a multiple-milliliter scale. Large-scale PCR differs from conventional PCR in that it is typically conducted in water baths using 15 mL, 17 × 120–mm, screw-capped (Sarstedt) thermostable tubes to accommodate the larger volumes. Amplification reactions of up to 2.5 L have been carried out in this way. Medium-scale amplifications can sometimes be carried out in thermal cyclers that can accommodate multiple samples (e.g., 96-well PCR plates).

Materials

Purified ssDNA pool and primers
EDTA
1:1 phenol/chloroform (*APPENDIX 2A*)
Chloroform
4 M ammonium acetate
Ethanol
TE buffer, pH 8.0, containing 50 mM of a salt such as potassium chloride

Thermal cycler or three water baths (one must be a circulating water bath)
96-well PCR plate or 13-mL thermostable tubes (Sarstedt)
Thermometer
Styrofoam racks
Spectrophotometer or fluorometer

Additional reagents and equipment for PCR amplification (*CPMB UNIT 15.1*; see Support Protocol 3 for determination of conditions on a small scale) and phenol/chloroform and chloroform extraction of DNA (*APPENDIX 2A*)

Plan the reaction

Since large-scale reactions are quite expensive in terms of nucleotides and enzyme, preparedness and planning for the large-scale amplification cannot be overemphasized. Primers <20 bases in length usually do not need to be gel purified and can instead be purified by precipitation.

1. After identifying the optimal PCR conditions on a small scale (see Support Protocol 3), prepare reagents for the large-scale reaction. Set aside time for the large-scale amplification, which will probably consume an entire day.

   The size of the large-scale reaction will be determined in part by the amount of DNA pool to be amplified and by the number of copies of the library that are desired. For example, assume that 100 (extendable) μg of a pool are to be amplified 16-fold. Since the typical amount of DNA recovered from a 100-μL PCR reaction is 1 μg, then each 100-μL reaction should have 1 μg/16 = 60 ng of DNA. 100 μg total/60 ng/100 μL = 1667 × 100 μL, or a 167-mL reaction.

Choose how the amplification will be carried out

If the volume of the large-scale amplification reaction is to be ≤100 mL

2a. Use a commercially available thermal cycler repetitively. Set the reaction mixture up in advance, and pipet 100-µL aliquots into individual wells of a 96-well PCR plate.

3a. Carry out several small amplification reactions in advance to ensure that the optimized conditions determined in Support Protocol 3 work with the PCR plate format, and that amplification is uniform across the PCR plate.

4a. Perform thermal cycling on the entire reaction using eleven PCR plates.

For larger volumes

Reactions will be divided into aliquots in 13-mL thermostable (Sarstedt) tubes and amplified in a series of water baths. Construct floating racks by cutting off the bottom of the tubes' Styrofoam packing material. Reinforce these racks by wrapping their edges with heavy tape. Place the racks iteratively in three circulating or static water baths held at the denaturation, annealing, and elongation temperatures previously determined (see Support Protocol 3).

2b. Determine how long it will take for the reaction mixture in a tube to come to thermal equilibrium by constructing a temperature probe, placing a thermometer through the top of a Sarstedt tube filled with 10 mL of water. Place the probe in a rack with other, similar tubes.

Typical equilibration times range from 2 to 8 min, depending on the temperature differential. Annealing, and extension times of 5, 6, and 7 min are typical. It should be noted that these ramping temperature profiles are very slow relative to a commercial PCR machine and can yield more amplification artifacts.

3b. To ensure that the reaction conditions actually work as planned, fill the rack with tubes of water, a single amplification reaction, and the temperature probe. Denature the sample for 30 min, and then add *Taq* after the first annealing step. Take aliquots at each cycle to monitor the progress of the reaction.

4b. When reaction conditions have been confirmed, proceed with the remaining amplification reactions. Allow the final extension step to proceed for at least 20 min to ensure that all templates are completely double-stranded.

Do not be alarmed if the solution becomes cloudy; the detergent in the buffer causes the turbidity.

Amplification efficiencies of 3 to 4 doublings in 5 cycles can typically be achieved using this method.

5. Following the amplification, pool the reactions from the individual wells or tubes. Chelate the magnesium in the buffer by adding 1.1 molar equivalents of ETDA, pH 8.0.

The reactions can be left at 4°C overnight.

6. Add an equal volume of 2-butanol and extract to concentrate the reaction to a manageable volume (usually 10- to 20-fold). Mix the layers by vortexing and then separate by centrifuging 5 min at 1200 × *g* at room temperature, then discard the upper, butanol layer. Repeat as necessary.

About one-fifth of the aqueous layer is extracted into the organic butanol layer for each volume of butanol used.

7. After concentrating the DNA, carry out a phenol/chloroform extraction, followed by two successive chloroform extractions (see recipe for phenol/chloroform/isoamyl alcohol in *APPENDIX 2A*).

   At this point, it should be possible to easily precipitate the DNA. Be sure to temporarily save all of the organic layers in case of a mishap. Falcon tubes (50 mL) work well for these extractions, as they are conveniently sized and have a small surface area. Alternatively, a Teflon extraction funnel may be useful since nucleic acids will not stick to its surface.

8. Precipitate the DNA by adding an equal volume of 4 M ammonium acetate (final concentration, 2 M) and 2 vol ethanol in 13-mL Sarstedt tubes if possible.

   If larger tubes are required, prepare a set of Beckman 250-mL high-speed centrifugation jars. Wash the jars with 15 mL of 3% hydrogen peroxide for 30 min and then rinse three times with 100 mL of distilled water to remove any residual DNases that may remain from previous use (typically bacterial cell pelleting).

9. Resuspend the amplified DNA in TE buffer, pH 8.0, containing 50 mM of a salt such as potassium chloride.

   It is unwise to resuspend a double-stranded DNA pool in water, since the random segments may denature, reassort, and become transcriptionally incompetent.

   If it is suspected that the pool has become denatured (for example, if a large single-stranded DNA component is seen on a nondenaturing agarose gel), simply repeat one to two cycles of PCR.

10. Quantitate the PCR DNA.

    This can be done by carrying out gel electrophoresis in parallel with a DNA ladder of known concentration. The concentration can also be determined spectrophotometrically or by monitoring the change in absorbance of an intercalated fluorescent dye, Hoechst 33258 (Sigma) on a fluorometer (e.g., DyNA Quant 200, Amersham Pharmacia Biotech). These latter methods are much more quantitative (although the fluorometer method may not be accurate for sequences <100 nucleotides in length). However, these methods may not distinguish precipitated double-stranded DNA from residual, precipitated nucleotides or single-stranded primers. Determine the overall PCR efficiency and the final number of DNA molecules.

    The amount of DNA obtained from large-scale amplification is often referred to in terms of the number of copies of the original synthetic pool's complexity. For example, if the starting pool had a complexity of 1×10^{15} molecules and 8×10^{15} total DNA molecules were recovered, then, on average, 8 copies of the original starting pool were obtained from the amplification. It should be noted that skewing that may arise during amplification, and sampling errors that occur during the use of the amplified pool, may cause this estimation to be grossly inaccurate; nevertheless, it is empirically useful.

11. Following large-scale amplification, store at least 4 copies of the pool at −80°C.

    Because of the aforementioned sampling errors, archiving at least 4 copies worth of the pool DNA ensures the preservation of most of the pool's complexity. The amount of preserved pool complexity can be calculated using the following equation:

    % of the pool complexity in a given sample = $100 \times \{1-[(x-y)/x]^x\}$

    where x is total number of pool copies, and y is the number of pool copies archived.

    Therefore, in the example given above, if 4 of the 8 copies of the pool generated through amplification are archived, then ~99.6% of the original starting pool's complexity is preserved. Similarly, at least 4 copies of the pool should be used whenever manipulations such as ligation, transcription, or biotinylation, are carried out, so that the original complexity is also manifest in the manipulated or synthesized copies.

COMMENTARY

Critical Parameters

Synthesis

Depending on the size of the pool to be synthesized, the operation of the DNA synthesizer may first need to be optimized. Short pools (<80 total nucleotides in length) can be synthesized using standard protocols (see, e.g., PerSeptive Biosystems, 1997). In order to synthesize longer pools (>80 total nucleotides in length), all reagents should be fresh and special care should be taken to exclude water from the synthesis (see *APPENDIX 3C*). To ensure equimolar base incorporation in the random region of longer pools, the phosphoramidites must be mixed in a skewed ratio (see Strategic Planning). Coupling efficiency should be monitored throughout the synthesis by following the trityl output (see *APPENDIX 3C*).

Amplification

Optimization of PCR conditions according to established protocols is vital to the success of the large-scale amplification. Cycle temperatures and times, as well as the concentrations of polymerase, primers, and dNTPs (see, e.g., *CPMB UNIT 15.1*) should be addressed prior to the large-scale workup. Most importantly, since extremely large quantities of relatively expensive reagents (e.g., *Taq* polymerase) may be required, care should be taken to make sure that all reagents and procedures are in readiness. Different priming sequences often require distinct PCR buffers for optimal extension efficiency; the best buffer for a given pool and primer combination can be easily and systematically identified through the use of a PCR optimization kit (e.g., the PCR Optimizer Kit from Invitrogen).

Troubleshooting

The most common problem with the synthesis of a random sequence pool is the overall synthetic yield. However, researchers should carefully decide how many sequences are really necessary for their selection experiments. In selection experiments from a pool with a relatively limited potential diversity (i.e., a segmentally random pool with only 1 × 10^{11} possible sequences or less), even a low synthetic yield should be sufficient. However, in vitro selection from a pool with a very high potential diversity (i.e., a completely random pool with 1 × 10^{15} possible sequences or more) should use at least 1 × 10^{14} different sequences initially in order to adequately sample the potential sequence space. Pools that contain fewer than 1 × 10^{13} possible sequences should not be used.

The most likely sources of low yields and coupling efficiencies are old (i.e., water-contaminated) synthesis reagents. Thus, instead of attempting to amplify an incomplete pool, the pool should be resynthesized with fresh reagents; the old and new pools can then be combined, if desired. If fresh synthesis reagents do not significantly raise yields, then more serious problems, such as line or valve blockage, may be the cause, and the instrument service representative should be contacted.

The second most common problem is that the base composition of a partially or completely random region is skewed. Unfortunately, skewing cannot be detected until after completion of a large-scale amplification. Fortunately, unless the degree of skewing is extreme, it should not seriously affect the outcome of a selection. Moreover, if the degree of skewing is known in advance of a selection, it can be taken into account when analyzing the results of the selection. For example, Baskerville et al. (1995) selected functional Rex-binding elements from a partially randomized pool. Despite the fact that the initial pool did not contain equimolar representation of non-wild-type bases at partially randomized positions, these authors were able to determine the relative importance of individual residues by comparing the degree of conservation or variance before and after selection. If a researcher decides that extant skewing of base ratios is unacceptable, this can only be fixed by adjustment of the randomized phosphoramidite mixture and resynthesis of the pool.

The third most common problem is that the pool fails to efficiently elongate. With the proviso that the efficiency of extension may be as low as 10% of the available pool, it should not be much lower (i.e., 1% of the available pool). If extension or PCR efficiency is dauntingly low, the PCR conditions should be reexamined and optimized as described, including buffer and enzyme concentrations, temperatures, and extension times. Switching to a different thermostable polymerase, or to a combination of polymerases, will sometimes improve primer extension. If all possible PCR optimization conditions have been addressed, poor extension efficiency could reflect a problem with the synthetic DNA. For example, the pool may not have been completely deprotected or

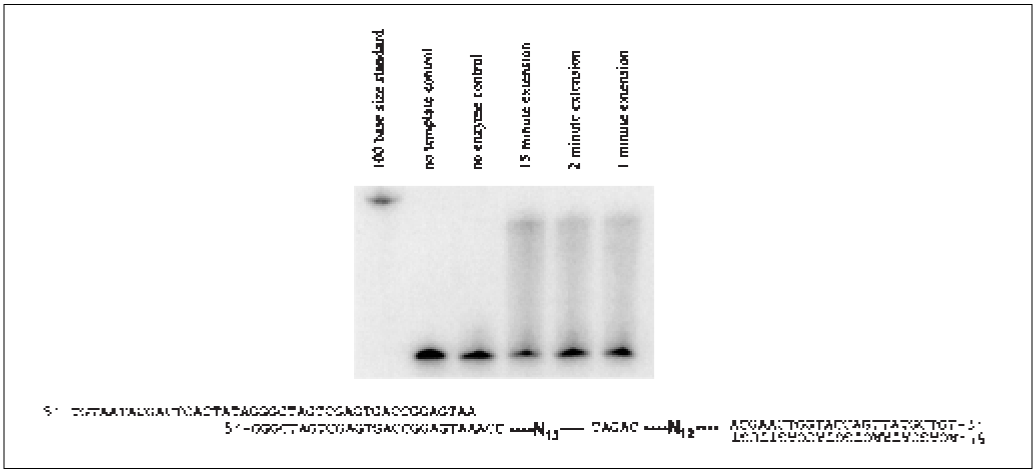

Figure 9.2.4 Typical extension reaction. The pool used (N30P) is shown below the figure of the gel. Lane 1 is a size standard, lanes 2 and 3 show control reactions, and lanes 4, 5, and 6 follow the extension efficiency after different incubation times.

a primer binding site may have become largely depurinated during the course of a long synthesis. Although incomplete deprotection is rarely a problem, small aliquots of the pool can be further treated with ammonia, and extension and amplification can again be assessed. If additional deprotection instead yields oligonucleotide degradation, then it is likely that apurinic sites have accumulated, and the pool will have to be resynthesized.

Anticipated Results

It is apparent from the discussion earlier in this unit that there is no one correct way to design and amplify a random sequence pool. However, by following the protocols described above, results similar to the following should be observed.

If the integrity of the nascent, synthetic pool is good, then the primer extension efficiency (described in Support Protocol 1) should be relatively high. Figure 9.2.4 shows a typical extension reaction for a pool synthesized in the authors' laboratory (N30P, a segmentally random pool). Lane 1 is an RNA size standard; lanes 2 and 3 show a control reaction with "no template" and "no enzyme," respectively. In these lanes, only the radiolabeled 3′ primer (24 nucleotides in length) is visible. Lanes 4, 5, and 6 show the primer extension reaction at various incubation times. Molecules that were incapable of full extension make up the smear leading to the full-length product. By determining the number of counts in the full-length product relative to the radiolabeled primer, the extension efficiency for the N30P pool was calculated to be ~8%. Moreover, it appeared as though the extension reaction had gone to completion within 2 min.

Assuming that the nascent pool is intact and can serve as a template for the primer extension reaction, then it should be possible to amplify the pool via the polymerase chain reaction. Figure 9.2.5 shows the results of an amplification "cycle course" for a different pool (N71, with a 71-nucleotide random sequence core). An 8-mL PCR reaction was placed in a 15-mL Falcon tube and cycled through a series of three water baths. The samples in the figure were drawn at 0, 2, 4, 6, 8, and 10 cycles. This initial PCR reaction was only a trial, and for the final, large-scale amplification of the entire pool, a 150-mL PCR reaction was distributed to 18 Falcon tubes and 7 PCR cycles were carried out. Following amplification, a portion of the N71 pool was cloned into a TA cloning vector (Invitrogen) and ten clones were sequenced. The proportions of different nucleotides in the final pool reflected almost perfect equimolar coupling efficiencies: A, 25.22%; C, 25.37%; G, 25.82%, and T, 23.58%.

Time Considerations

The amount of time required for the protocols described in this section should not be underestimated. Pool design will require at least one day, depending on the degree of background research. It is strongly recommended that pool design be discussed with one or more colleagues prior to synthesis. The synthesis of oligonucleotides <150 bases in length can be easily accomplished in one day, allowing 1 hr to ensure proper instrument setup. Pool purification and optimization of PCR conditions

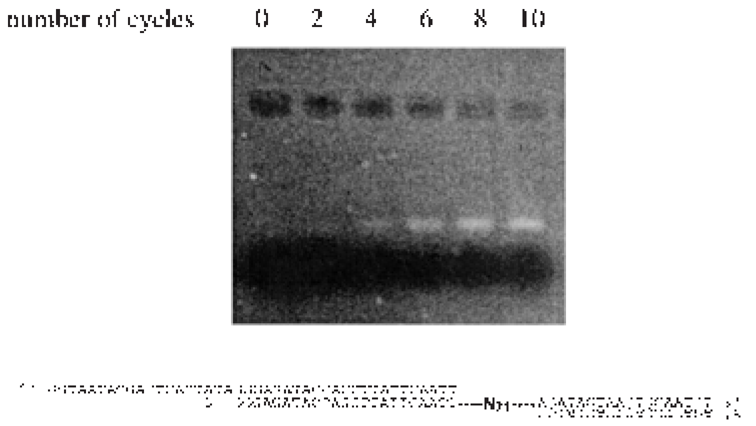

Figure 9.2.5 Cycle course. The gel follows amplification of the N71 pool after 0, 2, 4, 6, 8, and 10 cycles. The pool used in the cycle course is depicted below the figure of the gel.

should take 1 to 2 additional weeks. Finally, the actual large-scale amplification and subsequent isolation of the dsDNA pool will require the researcher's undivided attention for ~2 days.

Literature Cited

Bartel, D.P. and Szostak, J.W. 1993. Isolation of new ribozymes from a large pool of random sequences. *Science* 261:1411-1418.

Bartel, D.P., Zapp, M.L., Green, M.R., and Szostak, J.W. 1991. HIV-1 Rev regulation involves recognition of non-Watson-Crick base pairs in viral RNA. *Cell* 67:529-536.

Baskerville, S., Zapp, M., and Ellington, A.D. 1995. High-resolution mapping of the human T-cell leukemia virus type 1 rex-binding element by in vitro selection. *J. Virol.* 69:7559-7569.

Breaker, R.R. 1997. In vitro selection of catalytic polynucleotides. *Chem. Rev.* 97:371-390.

Chen, Z. and Ruffner, D.E. 1996. Modified crush-and-soak method for recovering oligodeoxynucleotides from polyacrylamide gel. *BioTechniques* 21:820-822.

Conrad, R., Keranen, L.M., Ellington, A.D., and Newton, A.C. 1994. Isozyme-specific inhibition of protein kinase C by RNA aptamers. *J. Biol. Chem.* 269:32051-32054.

Crameri, A. and Stemmer, W.P.C. 1993. 10^{20}-fold aptamer library amplification without gel purification. *Nucl. Acids Res.* 21:4410.

Fitzwater, T. and Polisky, B. 1996. A SELEX primer. *Methods Enzymol.* 267:275-301.

Giver, L., Bartel, D., Zapp, M., Pawul, A., Green, M., and Ellington, A.D. 1993. Selective optimization of the Rev-binding element of HIV-1. *Nucl. Acids Res.* 21:5509-5516.

Gold, L., Polisky, B., Uhlenbeck, O., and Yarus, M. 1995. Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* 64:763-797.

Hermes, J.D, Parekh, S.M., Blacklow, S.C., Koster, H., and Knowles, J.R. 1989. A reliable method for random mutagenesis: The generation of mutant libraries using spiked oligodeoxyribonucleotide primers. *Gene* 84:143-151.

Jaeger, J.A., Turner, D.H., and Zuker, M. 1989. Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol.* 183:281-306.

Jaeger, L. 1997 The new world of ribozymes. *Curr. Opin. Struct. Biol.* 7:324-335.

Lorsch, J.R. and Szostak, J.W. 1994. In vitro evolution of new ribozymes with polynucleotide kinase activity. *Nature* 371:31-36.

Lyamichev, V., Brow, M.A., and Dahlberg, J.E. 1993. Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. Science 260:778-783.

Milligan J.F., Groebe D.R., Witherell G.W., and Uhlenbeck O.C. 1987. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucl. Acids Res.* 15:8783-8798.

Osborne, S.E. and Ellington, A.D. 1997. Nucleic acid selection and the challenge of combinatorial chemistry. *Chem. Rev.* 97:349-370.

PerSeptive Biosystems. 1998. Expedite Nucleic Acid Synthesis System: User's Guide. PerSeptive Biosystems, Framingham, Mass.

Sabeti, P.C., Unrau, P.J., and Bartel, D.P. 1997. Accessing rare activities from random RNA sequences: The importance of the length of molecules in the starting pool. *Chem. Biol.* 4:767-774.

Singer, B.S., Shtatland, T., Brown, D., and Gold, L. 1997. Libraries for genomic SELEX. *Nucl. Acids Res.* 25:781-786.

Tuerk, C. and Gold, L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to baceriophage T4 DNA polymerase. *Science* 249:505-510

Tuerk, C. and MacDougal-Waugh, S. 1993. In vitro evolution of functional nucleic acids: High affinity RNA ligands of HIV-1 proteins. *Gene* 137:33-39.

Zon, G., Gallo, K.A., Samson, C.J., Shao, K., Summers, M.F., and Byrd, R.A. 1985. Analytical studies of "mixed sequence" oligodeoxyribonucleotides synthesized by competitive coupling of either methyl- or β-cyanoethyl-*N*,*N*-diisopropylamino phosphoramidite reagents, including 2′-deoxyinosine. *Nucl. Acids Res.* 13:8181-8196.

Contributed by Jack Pollard
Harvard University
Cambridge, Massachusetts

Sabine D. Bell and Andrew D. Ellington
University of Texas
Austin, Texas