Nucleic Acids Research, 1998, Vol. 26, No. 4 879–886

Combinatorial library diversity: probability assessment of library populations

Brian Ward* and Thomas Juehne

Sigma Chemical Company, PO Box 14508, St Louis, MO 63178, USA

Received December 11, 1997; Revised and Accepted December 30, 1997

ABSTRACT

A method is described for measuring the diversity of combinatorial oligonucleotide libraries that entails extrapolating the base composition of a co-synthesized model library (dNC, N = A, C, G, T) to that of a multibase library template. The base composition of dNC was measured by HPLC. The ability of dNC to predict the base composition of a multibase library template was corroborated by measuring the composition of a 12 base combinatorial library. The base composition of the 12 base library was determined by several template dependent incorporation assays: measurement of restriction fragment specific activities from polymerase incorporation/restriction enzyme digests, template directed radionucleotide primer extension and quantitative dideoxynucleotide sequencing. Additionally, a convention for describing oligomeric combinatorial library (OCL) diversity is proposed. The convention uses a quantity termed the diversity quotient (Q_d) to describe library breadth and the mole fraction of the least represented monomeric unit of the OCL to calculate minimum library quantity requirements. Similar methods/conventions could presumably be developed/adopted for non-nucleic acid libraries.

INTRODUCTION

Random sequence oligodeoxyribo and ribonucleic acid libraries have been used to isolate and identify sequences that bind to sequence specific ligands and less obvious targets (recently reviewed in 1–6). Such libraries are composed of random sequence blocks flanked by defined/primer sequences. Defined/primer sequences are for the amplifications used during the selection process. In practice, a DNA library is prepared from a synthetic template using the polymerase chain reaction. RNA libraries are prepared by transcription from such DNA libraries. In either case, the first step in preparing an oligonucleotide library is solid phase synthesis of a library template. The diversity of a library's amplification/transcription products can at best equal the parent synthetic template's diversity. Development of sequence independent amplification/transcription protocols will ensure that an aptamer's diversity is equal to its parent template's diversity. Such protocols would also benefit selections by eliminating target independent sequence biasing.

The random sequence section of the template can be prepared by allowing a DNA synthesizer to mix the building blocks or by filling an auxiliary bottle with a building block mixture (7,8). In planning an *in vitro* selection experiment, one must consider the amount of library initially being screened for molecules that display the property of interest. The initial amount of library is dependent upon combinatorial sequence length, desired pool complexity and practicality. Due to the reaction volumes necessary for inclusion of all possible sequences, it may be impractical to begin a selection experiment with a library pool that contains all possible sequences. In any case, assaying a library's population distribution before the commencement of a selection protocol is desirable. This is sometimes accomplished by sequencing a number of the individual library members (9,10; M.Yarus, personal communication). Here we describe methods for measuring the diversity of a library's synthetic template, henceforth referred interchangeably as the library or template.

The probability of finding a particular sequence in a library sample will depend upon the breadth of library diversity and the number of molecules sampled. Adequate breadth will serve to enhance the probability that failed or missed selections are not due to library deficiencies. It was recently reported that a first attempt at selecting binding sequences for a 19 base triplex forming oligonucleotide (TFO) failed due to deficiencies in the composition of the initially prepared template (9). Selections starting from a subsequently synthesized template successfully identified binding sites for the TFO. At the other end of the library length extreme, a class I ribozyme ligase has been isolated starting from a 220 nucleotide library (11). It was fortuitous that the ligase evolved from the pool since statistically such molecules should be isolated only once in 2000 selections (11). Intuitively, a deficient or biased library could make the odds even worse. Since library deficiencies can result in failed selection experiments and likely reduce the probability of fortuitous isolations, knowledge of a library's diversity prior to selection should enable one to design experiments that minimize library deficiencies.

To facilitate the discovery of novel structures using combinatorial techniques, it is hoped that the burden of library preparation will shift from user to secondary and/or commercial sources. In this event, it will be imperative that the user know the breadth of a supplier's library. It is anticipated that conventions for reporting library diversities and accepted methods for their measurement will be essential. Our immediate interest in combinatorial oligonucleotide libraries is derived from the preparation of duplex libraries for combinatorial type IIS restriction enzyme

*To whom correspondence should be addressed. Tel: +1 800 521 8956; Fax: +1 314 772 6797; Email: bwardhome@rocketmail.com

footprinting experiments (9,12). It occurred to us that having a quantity that describes a library's diversity, a quantity used for calculating minimal sequence representation and less demanding methods for their measurement would be desirable.

It was our hypothesis that the base composition of a template could be conveniently measured by determining the base composition of a co-synthesized one base library (dNC). To demonstrate this point, we found that the composition of dNC was consistent with an identically prepared multibase library (dN_{12} lib). The base composition of the model was measured by HPLC, dN_{12} lib was measured by several template dependent incorporation assays (TDIA). The quantitative TDIA included measurement of restriction fragment specific activities from polymerase incorporation/restriction enzyme digests, template directed radionucleotide primer extension and quantitative dideoxy-nucleotide sequencing. Quantitative dideoxy sequencing revealed that base addition during solid phase synthesis is approximately sequence independent. These data support our contention that easily prepared and characterized one base model libraries should be useful for the characterization of any length combinatorial template. Similar methods could presumably be developed for non-nucleic acid oligomeric libraries.

To report library diversity, we propose that a quantity termed the diversity quotient (Q_d) be used. Q_d is the mole fraction of the least abundant combinatorial monomer (**l**) divided by the most (**m**). This convention yields for completely random libraries a Q_d maximum of 1. Libraries with Q_ds that are <1 have reduced molar diversities (i.e. the number of individual sequences per mole of library). The quantity of library that achieves minimal sequence representation for an initial selection experiment is calculated using the quantity **l**. Since **l** is the mole fraction of the least represented combinatorial monomer (L), the greatest likelihood for all oligomers to be statistically represented is if the homo L containing oligomer is present. Since these two quantities are independent of oligomer structure, they could be used to describe the diversity of any oligomeric library.

MATERIALS AND METHODS

Unless otherwise noted, all materials and reagents were from Sigma. Radionucleotides were from Amersham. Oligonucleotides were prepared on a Biosearch 8600 DNA synthesizer using standard synthesis programs. The mixed base sequences were prepared by filling the auxiliary fifth base bottle ('U') with a solution prepared from a 3:3:2:2 molar ratio of dA:dC:dG:dT CED phosphoramidites (2,10).

dNC standard

The dimer standards were purified using conditions similar to the analytical separations. Dimer purity was assayed by HPLC and electrophoresis. Chromatograms of the individual dimers contained single-homogeneous peaks. Homogeneity was demonstrated by comparison of the up and down slope spectra with that of the apex. In all cases the up slope, down slope and apex spectra were in absolute agreement and had correlation coefficients that were ≥0.999. 5′-^{32}P-end-labeled dimers were resolved among themselves and shown to be single-banded by 20% PAGE. The concentrations of the individual dimers that were used for the preparation of the equimolar (1 mM each) HPLC standard were measured optically assuming the extinction coefficients for dAC, dCC, dGC and dTC are 10.6, 7.3, 8.8 and 8.1 mM^{-1} cm^{-1} respectively (13). The ratios of the dimer standards were also assayed by HPLC analysis of the exhaustively hydrolyzed (overnight digestion using snake venom phosphodiesterase from *crotalus durissus terrificus*) dimer mixture (14,15). The analysis revealed that the dimers were quantitatively hydrolyzed to nucleosides. The ratios of the nucleosides, and hence the dimer ratios, were calculated assuming the 260 nm extinction coefficients are 15.4, 7.4, 11.5 and 8.7/mM/cm for dA, dC, dG and dT respectively (13).

dNC HPLC analysis

Analytical HPLC was performed using Shimadzu hardware (LC-10AT liquid chromatograph, SCL-10A system controller, SPD-M10AV diode array detector) and software (EZChrom v.3) with a 250 × 4.6 mm Vydac 218TP C18 reversed phase column. The mobile phases were 10% aqueous acetonitrile (A), 1 M triethyl ammonium acetate (B) (Glen Research) and water (C). The gradient was from 30 to 70% A (1 ml/min, 20 min), B remained at 10% throughout. Dilute acetonitrile (Buffer A) was used because the separations were not reproducible when a 3–7% acetonitrile gradient was attempted using neat acetonitrile.

Library labeling and purification

Using standard methods (16), 15 µl of a 1:100 dilution of the crude template was 5′-^{32}P-end-labeled and purified on a 12% denaturing polyacrylamide gel. The full length band was excised and eluted by agitating the gel slice in 1 mM EDTA overnight. The oligonucleotide was concentrated by repeated extraction with sec-butanol followed by ethanol precipitation. The pellet was suspended in 15 µl of water (1× library) and stored at –70°C between uses.

[α-^{32}P]dNTP incorporation/restriction enzyme digestion assay

Modified T7 DNA polymerase/BglII. To seven 0.5 ml micro-centrifuge tubes were aliquoted 6 µl of library mix [0.1× library, 8.4 µM P2 and 104 µM dNTPs (N = A, C, G, T)]. To tubes 1–7 were added respectively 1 µl of H_2O, H_2O, [α-^{32}P]dATP, [α-^{32}P]dATP, [α-^{32}P]dCTP, [α-^{32}P]dGTP and [α-^{32}P]dTTP (3000 Ci/mM, 10 Ci/l). To each tube was added 3 µl of diluted polymerase [3.33× reaction buffer (Amersham), 25 mM DTT, 0.32 U/µl modified T7 DNA polymerase (Amersham)]. The reactions were incubated at 37°C for 5 min, temperature ramped to 65°C at 2.5°C/min followed by a 30 min 65°C incubation. The tubes were centrifuged (15 s) and equilibrated at 37°C. To reactions 2 and 4–7 were added 1 µl of *Bgl*II (10 U/µl). The restriction digests were incubated at 37°C for 30 min and quenched by the addition of 5 µl of gel loading buffer (Amersham). After heat denaturation (2 min, 94°C), 5 µl aliquots were loaded onto a 0.4 mm × 40 cm long 12% denaturing polyacrylamide gel. The gel was run at 85 W (constant power) until the xylene cyanol marker had migrated ~2/3 down the gel. The bands were excised from the gel and quantitated by scintillation methods using a Beckman 3801LS scintillation counter. *Taq* DNA polymerase/*Bgl*II, *Nsi*I. To four tubes, each containing 4 µl of [α-^{32}P]dNTP (N = A, C, G and T, respectively, at 3000 Ci/mM, 10 Ci/l) was added 21 µl of PCR mix. The PCR mix was prepared to contain: 2×10^{-5}× dilution of the crude dN_{12} library, 2.3 µM P1 and P2, 1.2× PCR buffer, 120 µM dNTPs and 0.05 U/µl *Taq* DNA polymerase. The amplification protocol was

10 cycles of 95°C for 30 s, 65°C for 30 s, 72°C for 60 s, followed by 10 min incubation at 72°C and 12°C soak. To polish the 3′-termini, 10 U of T4 DNA polymerase was added to each reaction and incubated at 12°C for 15 min (17). The amplimers were spin filter purified (Qiagen) and concentrated by rotary evaporation. After adjusting the volumes to 45 µl, 20 µl of each reaction were aliquoted into two sets of four 0.5 ml microcentrifuge tubes. Aliquots of 2.5 µl each of 10× reaction buffer and *Bgl*II (10 U/µl) were added to each tube of the first set. 10× buffer and *Nsi*I (20 U/µl) were identically added to the second set. The restriction digests were incubated at 37°C for 1 h. The reactions were quenched by the addition of 10 µl of gel loading buffer (Amersham). Denaturation, gel loading, electrophoresis and restriction fragment quantitation were as described for the modified T7 polymerase/*Bgl*II experiment.

Template directed radionucleotide primer extension

An aliquot of 9 µl of extension mix [0.1× library, 43 nM P2, 1.12× reaction buffer (Amersham), 5 mM DTT, 0.2 U/µl modified T7 DNA polymerase (Amersham)] were aliquoted into four tubes each containing 1 µl of 3000 Ci/mM, 10 Ci/l [α-^{32}P]dATP, [α-^{32}P]dCTP, [α-^{32}P]dGTP and [α-^{32}P]dTTP (Amersham) respectively. To maximize the ability of the polymerase to read through secondary structures, the reactions were incubated at 37°C for 5 min, temperature ramped to 65°C at 2.5°C/min followed by a 5 min incubation at 65°C (B. Ward, unpublished). The reactions were quenched and heat denatured as previously described. Aliquots (4 µl) were loaded in three sets onto a 20% denaturing polyacrylamide gel and electrophoresed at 85 W (constant power). To visualize the less intense bands, several sheets of film (XAR-5) were serially exposed to the gel. The first six bands (i.e. P2+1 through P2+6) were excised from the gel and scintillation counted. The triplicate scintillation data were averaged and the base compositions were calculated for the 72 bands as described in the appendix.

Quantitative dideoxynucleotide sequencing

To three sets of sequencing reactions were aliquoted 2 µl of [α-^{33}P]ddNTP (450 mCi/l, 1500 ci/mmol) and 4 µl of reaction mix [1:6.25 reaction buffer (Amersham), 0.1× library, 0.1 µM P3, 0.6 U/µl modified *Taq* DNA polymerase (Amersham)]. To each set were added respectively 4 µl of 7.5, 3.75 and 1.87 µM dNTPs. The reactions were layered with 15 µl of mineral oil and ramped to 95°C. The ramp protocol was [temperature (°C)/time (min)]: 50/3, 55/3, 60/4, 65/6, 70/9, 75/9, 80/9, 85/6, 90/3, 95/3, 4/soak. The reactions were quenched, heat denatured and electrophoresed (12% PAGE) as described above. The relative amount of each termination product was measured by microdensitometry of three separate autoradiograms using a Molecular Dynamics Personal Densitometer SI. Band volumes were measured using identical rectangles that encompassed the entire band. Background was compensated for by subtracting the average volumes of five identically sized background rectangles from the band volumes. Base compositions were calculated as described in the appendix.

RESULTS AND DISCUSSION

In the search for oligomers that have a desired property, the quantity of library initially present to achieve a desired level of representation is of critical significance. Intuitively, libraries that are deficient in key populations will have a lower probability of evolving sequences demonstrating the desired properties. For experiments aimed at characterizing the consensus binding sites for sequence specific ligands, one would ideally want to begin a selection experiment with a library pool that contained all possible sequences. To avoid missed target sequences due to sequence concentration effects, an ideal experiment would also have all of the possible sequences at equal concentrations. In the search for sequences that display other specific properties, it may be that subsets of the combinatorial sequence be thoroughly represented. In some *in vitro* evolutions one may want to skew the library away from maximal diversity (18). In any case, knowing the base composition of the library will allow one to begin a selection experiment with a library pool that contains a desired level of sequence representation.

A library's diversity can minimally be described by its length and Q_d. Since Q_d is defined as $\mathbf{l/m}$, for libraries containing j randomized positions the ratio of the least represented (L_j) to the most abundant (M_j) library member is Q_d^j, the library's breadth. From this it follows that the ratio of any two library members is $\geq Q_d^j$. Figure 1A shows that as a function of combinatorial sequence length, a library's breadth decreases rapidly as Q_d decreases. Even modestly low Q_ds (e.g. Q_d = 0.68) result in vanishingly small amounts (≤1%) of the minor sequences compared with the major at short library lengths (≥11 bases). Though Q_d describes a library's breadth, it alone cannot be used to determine the amount of library necessary for a selection experiment. For this, the mole fraction of the least represented monomeric species suffices. Statistically, a single copy of L_j can be achieved by $\mathbf{l}^{-j}$ molecules. For libraries that contain an equal distribution of all nucleotides at each randomized position (i.e. Q_d = 1), the probability (P) of a library sample containing a unique sequence is $\sim 1 - e^{-k}$, where k, the representation factor, is the number of molecules in the pool (p) divided by the number of possible monomeric species raised to the length of the combinatorial sequence (i.e. $k = p/4^j$) (19,20). For libraries with Q_ds < 1, it follows that the probability of a library pool containing L_j, and hence all possible sequences, is given when $k = p \times \mathbf{l}^j$. To calculate the quantity of library necessary to attain a desired level of representation, one sets P equal to an acceptable probability and solves for p. For example, a pool of $4.6 \times \mathbf{l}^{-j}$ molecules will have a 99% probability (P = 0.99) of having one L_j. In such a pool, all members are represented with a probability that is ≥0.99. From this it is immediately apparent that the amount of library initially needed to achieve minimal sequence representation is critically dependent upon $\mathbf{l}$. For example, it requires 180 µg (4.6×0.25^{-25} molecules, P = 0.99) of a Q_d = 1 (i.e. $\mathbf{l}$ = 0.25), 65 base library (25 combinatorial positions and two 20mer primer sites) to have at least one copy of each possible sequence. If $\mathbf{l}$ were 0.23, that representation is achieved with 1480 µg of library. Figure 1B shows for several values of $\mathbf{l}$ ($\mathbf{l}$ = 0.2–0.25) the amount of library necessary to attain singular representation as a function of combinatorial sequence length. As described in the Figure legend, the library amounts were normalized to a Q_d = 1 library. From this it is seen that modest reductions in $\mathbf{l}$ result in large increases in the amount of library necessary to begin an all inclusive selection experiment. For libraries with combinatorial sequences that are beyond the practical limits of all inclusivity, $k = p \times \mathbf{l}^j$ is the representation factor for calculating the lowest probability of finding a unique sequence. From this it is evident that Q_d and $\mathbf{l}$ are adequate for describing any oligomeric library's breadth

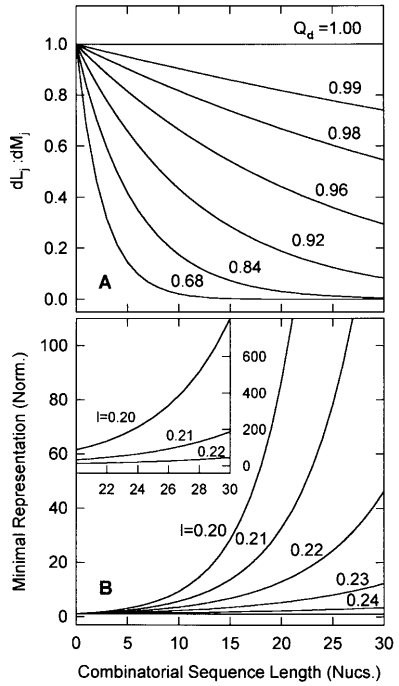

Figure 1. (**A**) The ratio of the least abundant library member (dL_j) to the most (dM_j) as a function of combinatorial sequence length (j) for several values of Q_d. dL_j/dM_j is Q_d^j. (**B**) The amount of library necessary to include one copy of dL_j normalized to a $Q_d = 1$ library as a function of j. Minimal representation is given by $(\mathbf{l}/0.25)^{-j}$. **l** is the mole fraction of the least represented base.

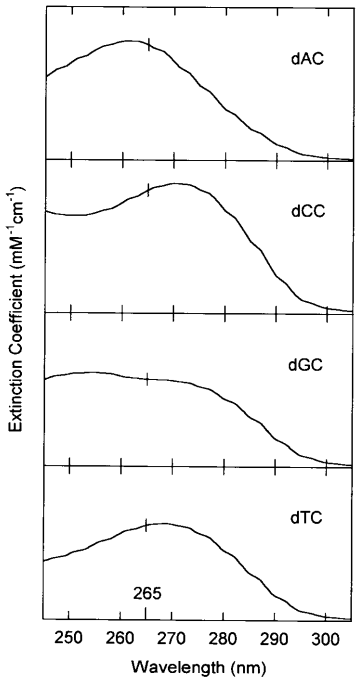

Figure 2. HPLC apex absorbance spectra of dNC dinucleotides.

and for calculating pool requirements. For oligonucleotide libraries, it is sufficient to report the mole fraction of each monomeric unit. However, such convention does not directly address a library's diversity. Q_d intuitively expresses library diversity. This may particularly benefit libraries built from a larger number of monomers.

To measure Q_d and **l** for combinatorial oligonucleotide libraries, it was our hypothesis that the base composition of a co-synthesized one base library would predict the composition of a full length template. We consider co-synthesis as that occurring immediately before, after or concurrent with the full length library using identical reagents and conditions. For most exacting work, and where the stability of the monomers may be in question, it may be of benefit to co-synthesize before and after library template synthesis. If the one base library compositions agree, it should be safe to assume that no biasing occurred during library template synthesis. Since standard phosphoramidites are expected to not degrade significantly during a synthesis, this precaution is likely not routinely necessary. The one base library would be of the type dNX where N is a randomly incorporated nucleotide and X is a solid support defined moiety. The 3′ position of the library could have been occupied by any solid support defined moiety, however, a deoxynucleotide was chosen because these supports are universally found in DNA synthesis laboratories. The one base library chosen for this experiment was dNC because preliminary HPLC experiments revealed that dNC components were more easily separated than were dNA, dNG or dNT. These preliminary experiments indicated that relative base incorporation was independent of the solid support bound base. Figure 2 contains the chromatographic apex absorption spectra of dAC, dCC, dGC and dTC. Because the four dimers do not share an absorbance maximum, signal to noise can be maximized by monitoring the separations at 265 nm. Figure 3 contains chromatograms of purified (std) and co-synthesized (lib) dNC. The standard was prepared by mixing individually synthesized and purified dimers. The crude dNC library (lib Fig. 3) was prepared from a premixed solution of the phosphoramidites (2,8,10). From the standard it is clearly evident that the four components of dNC are baseline resolvable by reversed phase HPLC. The order of elution is the same as reverse phase separation of the individual nucleosides (14). In addition to the four dNCs, lib contains minor peaks attributable to hydrolysis/deprotection products and any unreacted starting material (i.e. dC from dC-CPG). The expected hydrolysis/deprotection products that significantly absorb at 265 nm are dA, 3′dAMP, dC, 5′dCMP, 3′dCMP, dG, 3′dGMP, dT, 3′dTMP and benzamide (Bza). These peaks were identified by spectral and co-elutional analysis. The seemingly excessive amount of dC, 5′dCMP and 3′dCMP (dC in Fig. 3) results from summing all dimer degradations with any unreacted starting material. Because dTC's extinction coefficient was found to be dependent upon eluant composition , the mole fractions of dAC, dCC, dGC and dTC contained in dNC lib were calculated from the composition of the dimer standard (see Appendix). The dimer library was found to be 0.31 dAC, 0.26 dCC, 0.20 dGC and 0.23 dTC (mole fractions).

To test our contention that each combinatorial nucleotide position of a co-synthesized library would have a base composition that was similar or equal to that of a model one base library, the library shown in Figure 4 (dN_{12} lib) was synthesized immediately after dNC. A short library was chosen in an effort to circumvent or minimize any sequence biasing during the quantitative TDIAs that were used to measure the base composition of the library's combinatorial sequence (dN_{12}). The TDIAs have in common that

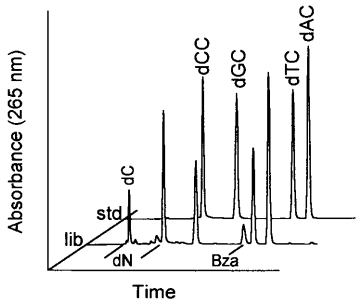

Figure 3. Chromatograms of pure dNC (std) and crude co-synthesized dNC (lib). dN are nucleosides and 3′ nucleotide monophosphates, dC is dC + 5′dCMP + 3′dCMP, Bza is Benzamide. Time is from 0 to 20 min.

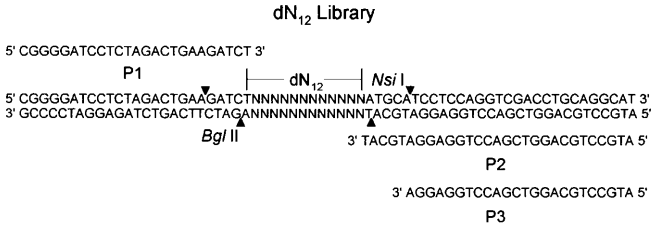

Figure 4. dN_{12} library. dN_{12} is the random (N) section of the library. P1, P2 and P3 are primers used for TDIA experiments. The proximal restriction fragments are those that contain the randomized sequence and P1 or P2. The distal fragments are those that are 3′ of the proximal fragments. The proximal compliment fragments are the remaining labeled fragments and are complimentary to the proximal fragments.

a DNA polymerase was used to catalyze the template directed incorporation of a radionucleotide into a complimentary copy of the synthesized library. To determine base composition, the specific activities of the copies were measured as a function of radionucleotide identity. In the first experiments described below, the mole fractions of the nucleotides contained in dN_{12} were measured from ^{32}P-labeled dN_{12} lib restriction fragments. In a second TDIA, the geometric reduction of homogenous base sequences was measured by primer extension in the presence of a single radionucleotide. Finally, the per position and average base mole fraction of the nucleotides contained in the library were measured by quantitating dideoxynucleotide termination products.

Cleavage of radionucleotide incorporated copies of dN_{12} lib with *Bgl*II or *Nsi*I (Fig. 4) results in radiolabeled restriction fragments with specific activities that are proportional to their nucleotide composition (Table 1). Template directed radionucleotide incorporation was catalyzed by modified T7 and *Taq* DNA polymerases using primers P2 and P1+P2 respectively. For either restriction endonuclease, the cleavage products produce single-stranded fragments that contain the combinatorial sequence and a defined sequence fragment. The defined sequence fragment was engineered to be used as an internal control for the calculation of the combinatorial sequence's base composition. The average occupation of a base in dN_{12} is given by the specific activity ratio of the dN_{12} containing fragment(s) to the defined sequence fragment corrected to the number of incorporated nucleotides (see Appendix). A possible source of error in this experimental design is restriction cleavage within dN_{12}. Errors resulting from this occurrence are negligible. For an idealized 12 base-Q_d = 1 library, only 0.15% (i.e. 6×0.25^{-6}) of dN_{12} would be cleavable by a six base requiring restriction enzyme.

For the modified T7 polymerase experiment, an aliquot of the library template was 5′-^{32}P-end-labeled and gel purified. Purification was performed so that synthetic failure sequences would be unable to over represent the proximal fragment (i.e. *Bgl*II 42mer). The primer P2 (Fig. 4) was extended in four separate reactions each containing an [α-^{32}P]dNTP (N = A, C, G or T) and unlabeled dNTPs. After the extension reaction, the polymerase was heat inactivated and the double-stranded products were cleaved with *Bgl*II and separated by 12% PAGE. The expected bands were excised from the gel and quantitated using scintillation methods. For this experiment, only the *Bgl*II 42 and 24mers contain incorporated radionucleotides. After normalization (i.e. a + c + g + t = 1 where a, c, g and t are the mole fractions of dA, dC, dG and dT), the composition of dN_{12} was: a = 0.273 ± 0.002, c = 0.285 ± 0.002, g = 0.209 ± 0.002 and t = 0.233 ± 0.003. The standard deviations are derived from triplicate scintillation measurements of a single experiment. Except for the [α-^{32}P]dCTP containing reaction, the denaturing gels showed that the extension and restriction reactions had proceeded smoothly. Even after prolonged film exposures, there were no visible polymerase stalling bands in the [α-^{32}P]dATP, -dGTP or -dTTP containing reactions. Reactions containing [α-^{32}P]dCTP did contain some faint stalling bands throughout dN_{12}. The most intense of these bands corresponded to the addition of the first two dCs to P2. The stalled bands were presumably due to the inability of modified T7 DNA polymerase to read through some dG-rich template sequences. From this it appears that the dG content of dN_{12} was slightly underestimated in this experiment.

Table 1. Number of primer extended nucleotides contained in single stranded dN_{12}(lib) *Bgl*II and *Nsi*I restriction fragments

Fragment	dA	dC	dG	dT
*Bgl*II 24mer (d)	5	7	6	6
*Bgl*II 42mer (p)	12t + 1	12g	12c	12a
*Bgl*II 46mer (pc)	12a + 6	12c + 10	12g + 7	12t + 6
*Nsi*I 24mer (d)	4	9	6	5
*Nsi*I 38mer (pc)	12t + 6	12g + 7	12c + 6	12a + 6
*Nsi*I 42mer (p)	12a + 2	12c + 1	12g + 1	12t + 1

dA, dC, dG and dT are the number of the respective nucleotide contained in each fragment. a, c, g and t are the mole fractions of dA, dC, dG and dT at each N in the 12 base combinatorial sequence of the chemically synthesized library strand (upper strand, Fig. 4). d, p and pc are distal, proximal and proximal compliment respectively.

The incorporation of radiolabeled nucleotides during amplification offers the opportunity to measure base composition of both strands. Except for the primer bases, all positions are subject to radiolabeling. Radiolabeled dN_{12} lib was prepared by direct amplification of crude dN_{12} lib using *Taq* DNA polymerase and primers P1+P2 (Fig. 4) in four separate [α-^{32}P]dNTP containing reactions. The expected 3′ overhang was removed after the amplification by a low temperature end polishing step (17). To assure that the specific activity of the distal restriction fragments could not be altered, the polishing step was performed by adding T4 DNA polymerase directly to the amplification reactions. The duplexes were then purified using routine methodologies (see Materials and Methods), separately cleaved with *Bgl*II and *Nsi*I

and separated by 12% PAGE. Unlike the modified T7 DNA polymerase experiment, no bands were observed which would have resulted from template dependent polymerase stalling. As summarized in Table 1, cleavage of the amplimer with *Bgl*II or *Nsi*I yields three labeled restriction fragments (i.e. proximal, proximal compliment and distal). The base composition of dN_{12} was calculated from the *Bgl*II and *Nsi*I defined and combinatorial sequence containing bands as described in the appendix. The composition averages are: a = 0.32 ± 0.08, c = 0.27 ± 0.04, g = 0.22 ± 0.02 and t = 0.20 ± 0.06. Deviations between gel loadings were ≤4%. The majority of the above error is the result of significant variability between the upper and lower strand base incorporations. Though the composition averages are consistent with all of the other measurements (Table 2), it would appear that some selection may have occurred during amplification. This result suggests that it would be of value in the design of combinatorial libraries to flank the random sequence with restriction sites so that equal base complimentarity can be a part of an amplification optimization scheme (20). No attempt was made here to develop a sequence independent amplification protocol.

The base composition of dN_{12} was also measured by template directed primer extension reactions (P2 + gel purified library, modified T7 DNA polymerase) that contained individual [α-^{32}P]dNTPs. The resulting (P2 + i)mers, where i refers to the number of radiolabeled nucleotides added to the primer, were separated by 20% PAGE and quantitated as above. As expected (see Appendix), due to the geometric reduction of homologous base sequence compliments it was possible to only measure the specific activities of the first six base addition products. The positional base mole fractions (n_i) of dN at positions P2+2 through P2+6 for three sets of gel loadings were calculated relative to the first labeled band (i.e. P2 + 1) and the band immediately preceding the band of interest (i.e. P2 + i – 1) according to equations **5** and **6** in the Appendix. The n_is were normalized so that at each i the sum of a_i, c_i, g_i and t_i =1. These n_is were averaged to yield: a = 0.266 ± 0.009, c = 0.266 ± 0.009, g = 0.222 ± 0.012 and t = 0.247 ± 0.006. From the modified T7 DNA polymerase/*Bgl*II experiment, it is expected that here too c may be slightly underestimated. Unfortunately this could not be reconciled by doing the parallel experiment using *Taq* DNA polymerase. In the presence of only one radionucleotide, *Taq* DNA polymerase promiscuously catalyzed terminal base additions that were compliment independent.

The TDIA experiments described so far all assume that no sequence bias occurred during the solid phase synthesis of the template. That is, the rate of adding dA, dC, dG or dT to the growing oligomer was independent of all previous additions. To substantiate this assumption, the gel purified library was sequenced three times using P3, modified *Taq* DNA polymerase and three different [α-^{33}P]ddNTP:dNTP ratios. The sequencing reactions were repeated with different nucleotide ratios to ensure terminator:deoxynucleotide independence. The termination products were quantitated using densitometric methods. The positional n_is were calculated by dividing the combinatorial sequence band volumes by the average of the defined sequence band volumes. The n_is were then normalized for each terminator:dNTP reaction set (see Appendix). Bar graphs of these base mole fractions as a function of base position for the three sets of sequencing reactions were indistinguishable (standard deviations ≤6%). This result demonstrated that the reactions and quantitations were performed under conditions that were independent of the nucleotide ratios. The relative base occupations from the three experiments were then averaged and plotted as the bar graphs shown in Figure 5. From this, it is evident that the proportion of each base is relatively independent of base position, substantiating that base addition is sequence unbiased during solid phase synthesis. The average base mole fractions were: a = 0.300 ± 0.002, c = 0.287 ± 0.002, g = 0.172 ± 0.001 and t = 0.241 ± 0.002.

The TDIA experiments all yielded similar base compositions for dN_{12} (Table 2). These base composition results are in excellent agreement with the base composition of co-synthesized dNC. The TDIA averages and dimer (parenthesis) base mole fractions are: a = 0.29 ± 0.02 (0.31), c = 0.28 ± 0.01 (0.26), g = 0.21 ± 0.02 (0.20) and t = 0.23 ± 0.02 (0.23). This agreement leads to the conclusion that dNC can accurately predict the diversity of a co-synthesized library. The **l**s obtained from the dimer and TDIA average data predict respectively that 41 and 23 pg of dN_{12} lib would have a 99% probability of containing at least one copy of each possible sequence. An identically composed 25 randomized position library (65 bases total) would require 51 versus 14 mg (dimer versus TDIA). These small differences are considered inconsequential in view of the variance between the individual TDIA p's (Table 2).

Table 2. Nucleotide mole fractions, l, Q_d and p (P = 0.99) of dN_{12}

Method	a	c	g	t	l	Q_d	p(pg)
Dimer HPLC	0.31	0.26	0.20	0.23	0.20	0.66	41
mod T7/*Bgl*II	0.273	0.285	0.209	0.233	0.209	0.77	24
Taq/*Bgl*II, *Nsi*I	0.32	0.27	0.22	0.20	0.20	0.62	41
mod T7/[α-^{32}P]dNTP	0.266	0.266	0.222	0.247	0.222	0.83	12
mod *Taq*/[α-^{33}P]ddNTP	0.300	0.287	0.172	0.241	0.172	0.57	250
TDIA	0.29	0.28	0.21	0.23	0.21	0.72	23

a, c, g and t are the mole fractions of dA, dC, dG and dT in dN12. **l** is the least of a, c, g, t. Q_d is **l/m** where **m** is the largest of a, c, g, t. p is the minimum quantity of library necessary to have a 99% probability of containing dL_{12}. Dimer HPLC are the results from measuring the mole fractions of dAC, dCC, dGC and dTC contained in the co-synthesized dNC library. Mod T7/*Bgl*II and *Taq*/*Bgl*II, *Nsi*I are from measuring the relative specific activities of the fragments after restriction digestion of the copied (modified T7 DNA polymerase) and amplified (*Taq* DNA polymerase) dN_{12} (lib) respectively. Mod T7/[α-^{32}P]dNTP is from [α-^{32}P]dNTP primer extension of P2 using modified T7 DNA polymerase. Mod *Taq*/[α-^{33}P]ddNTP is the quantitative dideoxynucleotide sequencing results using modified *Taq* DNA polymerase. TDIA (Template Dependent Incorporation Assay) columns a, c, g, t are the averages of the preceding four radionucleotide incorporation assays. TDIA columns l, Q_d and p(pg) are from the TDIA columns a, c, g, t.

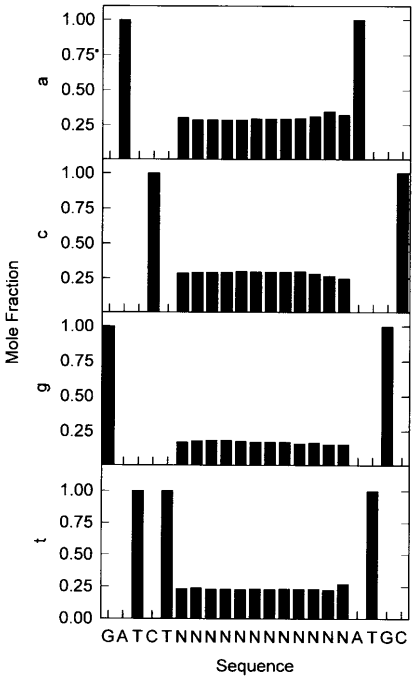

Figure 5. Histograms of nucleotide mole fractions as a function of sequence determined from quantitative dideoxynucleotide sequencing. a, c, g and t are the normalized (i.e. a + c + g + t = 1) mole fractions using the terminators: [α-^{33}P]ddATP, [α-^{33}P]ddCTP, [α-^{33}P]ddGTP and [α-^{33}P]ddTTP respectively. Average standard deviations of a, c, g and t from three experiments are 2.4, 1.9, 1.5 and 6.1% respectively.

Combinatorial libraries have been shown to contain sequences that specifically bind a diverse array of molecular targets. Because of this, one might suspect that enzyme linked methods of base composition measurement could be subject to a level of sequence selection. For such methods to be void of any selection, the polymerization, restriction digestion and cloning/sequencing processes must process all of the sequences present with equal efficiency. The measurement of restriction fragment specific activities indicated that this condition may not have been entirely met. It is therefore desirable to develop methods that measure combinatorial library diversity that minimize the opportunity for selection. Co-synthesis of dNC seems ideally suited for this since it is unlikely that a selection could take place during its preparation or analysis that had selectively removed or enriched any of its components. This is supported by the agreement between dNC's composition and TDIA results for dN_{12} lib. Quantitative dideoxy sequencing of dN_{12} lib, a model for longer libraries, further demonstrated that template synthesis is approximately sequence independent. From this we conclude that it should be possible to use co-synthesized dNC to model the diversity of any synthetic oligonucleotide template. The dimer should also be useful for identifying phosphoramidite ratios that would produce any desired level of diversity.

CONCLUSION

We have described a convenient method for assessing oligonucleotide library diversity. It was our hypothesis that the base composition of a co-synthesized one base library (dNC) would be approximately equal to that of a multibase library. This proposal was tested by comparing the base composition of dNC with that of a 12 base combinatorial library template (dN_{12} lib). The base composition of dNC was determined by HPLC, dN_{12} was measured by several template dependent incorporation assays (TDIA). The TDIA experiments yielded dN_{12} base compositions that were in excellent agreement with the base composition of dNC. From a quantitative dideoxy sequencing experiment we have shown that base addition during solid phase oligonucleotide synthesis is essentially independent of the growing oligomer. From these data we conclude that dNC should adequately model the diversity of any synthetic oligonucleotide library. Assuming that sequence independent amplification and transcription protocols are used, the model dimer too will predict the diversity of selection libraries. Assuming that building block addition is independent of all previous additions, similar methods could be developed for other types of oligomeric libraries.

A convention for reporting combinatorial oligomeric library diversity was proposed. The diversity coefficient (Q_d) and the mole fraction of the least represented monomer (**l**) minimally represent the diversity of such libraries. Q_d is **l**/**m** where **m** is the mole fraction of the most represented monomer. Libraries are most random at the Q_d maximum of 1. Q_ds that are <1 contain fewer members per mole. The probability that a library pool (p) will contain at least one copy of each possible library member is $1 - e^{-k}$, where k is $p \times \mathbf{l}^j$. Q_d and **l** have been presented here in the context of random sequence oligonucleotides. It is not necessary that these quantities be limited to the diversity of randomized nucleic acids. Due to their respective definitions, they can be universally applied to any oligomeric combinatorial library.

ACKNOWLEDGEMENTS

We thank, in alphabetical order, Drs A. Ellington, R. Hernon, R. Lirette, M. Van Dyke and M. Yarus for critical manuscript review and helpful suggestions. We also thank Dr J. Szostak for help with reference 10.

REFERENCES

1 Osborne,S. and Ellington,A.D. (1997) *Chem. Rev.*, **97**, 349–370.
2 Conrad,R.C., Giver,L., Tian,Y. and Ellington,A.D. (1996) In Abelson,J.N. (ed.), *Methods in Enzymology.* Academic Press, San Diego, Vol. 267 pp. 336–367.
3 Gold,L., Polisky,B., Uhlenbeck,O. and Yarus,M. (1995) *Annu. Rev. Biochem.*, **64**, 763–797.
4 Conrad,R.C., Baskerville,S. and Ellington,A.D. (1995) *Mol. Diversity*, **1**, 69–78.
5 Trotta,P.P., Beutel,B.A. and Sherman,M.I. (1995) *Med. Res. Rev.*, **15**, 277–298.
6 Chapman,K.B. and Szostak,J.W. (1994) *Curr. Opin. Struct. Biol.*, **4**, 618–622.
7 Ellington,A. and Green,R. (1989) In Ausubel,F.M., Brent,R., Kingston,R.E., Moore,D.D., Seidman,J.G., Smith,J.A. and Struhl,K. (eds), *Current Protocols in Molecular Biology.* John Wiley and Sons, Inc., New York, Vol. 1 pp. 2.11.1–2.11.8.
8 Fitzwater,T. and Polisky,B. (1996) In Abelson,J.N. (ed.), *Methods in Enzymology.* Academic Press, San Diego, Vol. 267 pp. 275–301.
9 Hardenbol,P. and VanDyke,M.W. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 2811–2816.
10 Bartel,D.P. and Szostak,J.W. (1993) *Science*, **261**, 1411–1418.
11 Ekland,E.H., Szostak,J.W. and Bartel,D.P. (1995) *Science*, **269**, 364–370.
12 Ward,B. (1996) *Nucleic Acids Res.*, **24**, 2435–2440.
13 Borer,P.N. (1975) In Fasman,G.D. (ed.), *Handbook of Biochemistry and Molecular Biology, Nucleic Acids.* CRC Press, Boca Raton, 3rd ed. Vol. 152, pp. 589.
14 Connolly,B.A. (1991) In Eckstein,F. (ed), *Oligonucleotides and Analogues: a Practical Approach.* IRL Press, Oxford, pp. 179–180.
15 Kallansrud,G. and Ward,B. (1996) *Anal. Biochem.* **236**, 134–138.

16 Sambrook,J., Fritch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. 2nd ed. Cold Spring Harbor University Press, Cold Spring Harbor, pp. 10.59–10.61.
17 Sambrook,J., Fritch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. 2nd ed. Cold Spring Harbor University Press, Cold Spring Harbor, F.4–F.5.
18 Huizenga,D.E. and Szostak,J.W. (1995) *Biochemistry*, **34**, 656–665.
19 Tabler,M., Benos,P. and Dörr,M. (1996) *Nucleic Acids Res.*, **24**, 3437–3438.
20 Ciesiolka,J., Illangasekare,M., Majerfeld,I., Nickels,T., Welch,M., Yarus,M. and Zinnen,S. (1996) In Abelson,J.N. (ed.), *Methods in Enzymology*. Academic Press, San Diego, Vol. 267 pp. 315–335.

APPENDIX

dNC HPLC

The mole fraction ratios calculated for the dimer library were based upon the standard's ratios according to Equation **1**.

$$\left(\frac{n'}{n}\right)_{lib} = \left(\frac{n'}{n}\right)_{std} \times \left(\frac{A_{dN'C}}{A_{dNC}}\right)_{lib} \times \left(\frac{A_{dNC}}{A_{dN'C}}\right)_{std} \quad \mathbf{1}$$

n, n′, A_{dNC} and $A_{dN'C}$ are respectively the mole fractions and peak areas for bases N and N′. The subscripts lib and std refer to the library and standard. The mole fraction (n) of base N in dNC_{lib} was calculated using Equation **2**.

$$n = \frac{1}{\left(\frac{n'}{n}\right)_{lib} + \left(\frac{n''}{n}\right)_{lib} + \left(\frac{n'''}{n}\right)_{lib} + 1} \quad \mathbf{2}$$

n, n′, n″ and n‴ are the mole fractions of dNC, dN′C, dN″C and dN‴C.

Restriction fragment quantitation

Incorporation of [α-^{32}P]dNTPs (N = A, C, G or T) into copies of the library followed by restriction digestion produces restriction fragments with specific activities that are proportional to the number of Ns contained in the fragment. The relationship between the defined sequence fragments and the combinatorial sequence containing fragments is:

$$\frac{N_{def}}{B_{def}} = \frac{jn + N_{comb}}{B_{comb}} \quad \mathbf{3}$$

N_{def} and N_{comb} are the number of known dNs contained in the defined and combinatorial sequence containing restriction fragments, j is the length of the combinatorial sequence and n is the mole fraction of dN in the combinatorial sequence. B_{def} and B_{comb} are the specific activities of the defined and combinatorial sequence containing bands. After rearrangement, the mole fraction of base dN in the combinatorial sequence is given by Equation **4**.

$$n = \left(N_{def}\frac{B_{comb}}{B_{def}} - N_{comb}\right)/j \quad \mathbf{4}$$

Homologous base addition quantitation

The positional mole fraction (n_i) of dN for the i^{th} primer addition product (P2 + i) relative to the first labeled primer extension product (P2 + 1) is:

$$n_i = \left(\frac{\sum_{i=i}^{j} B_i/i}{\sum_{i=1}^{j} B_i/i}\right)^{1/(i\,-\,1)} \quad \mathbf{5}$$

B_i is the specific activity of the P2 + i band and j is the length of the combinatorial sequence. In practice j = 6 (see below), though strictly the summation should include any bases incorporated after i = j. The summations result from the total P2 + i being the measured P2 + i and its elongated homologs. Similarly, n_i was calculated relative to its preceding addition product according to Equation **6**.

$$n_i = \frac{\sum_{i=i}^{j} B_i/i}{\sum_{i=i-1}^{j} B_i/i} \quad \mathbf{6}$$

The summation of the i's through j = 6 is the result of only being able to quantitate bands corresponding to P2 + 1 through P2 + 6. This is understandable because the specific activity of P2 + i (B_i) is proportional to the number of labeled nucleotides added to P2 times the mole fraction of dN in the template raised to the number of incorporated nucleotides less the P2 + i homologous addition products (i.e. Equation **7**).

$$B_i \propto in_i^i - \sum_{i=i+1}^{j} n_i{}^i \quad \mathbf{7}$$

For an idealized $Q_d = 1$ library, the relative specific activities of P2 + 1 through P2 + 6 are: 1.00, 0.625, 0.250, 0.0859, 0.0273 and 0.00830.

Dideoxynucleotide termination product quantitation

To minimize errors due to lane to lane variations, the positional dN mole fractions (n_i) were calculated relative to the average of the defined sequence band volumes according to Equation **8** followed by normalization (i.e. $a_i + c_i + g_i + t_i = 1$).

$$n_i = V_i \frac{N}{\sum_{1}^{N} V_N} \quad \mathbf{8}$$

V_i and V_N are the combinatorial and defined sequence autoradiogram band volumes respectively. N is the number of defined sequence band volumes used for the normalization. The mole fraction of dN in dN_{12} (n) was calculated according to Equation **9**.

$$n = \frac{\sum_{i=6}^{17} n_i}{\sum_{i=6}^{17}(a_i + c_i + g_i + t_i)} \quad \mathbf{9}$$

The summations begin at i = 6 because the first nucleotide added to P3 that is complementary to a combinatorial position corresponds to P3 + 6.