Protein Science (1993), *2*, 1249–1254. Cambridge University Press. Printed in the USA.

Design of synthetic gene libraries encoding random sequence proteins with desired ensemble characteristics

THOMAS H. LaBEAN[1] AND STUART A. KAUFFMAN[1,2]

[1] Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104
[2] Santa Fe Institute, Santa Fe, New Mexico 87501

(RECEIVED January 12, 1993; REVISED MANUSCRIPT RECEIVED April 19, 1993)

Abstract

Libraries of random sequence polypeptides are useful as sources of unevolved proteins, novel ligands, and potential lead compounds for the development of vaccines and therapeutics. The expression of small random peptides has been achieved previously using DNA synthesized with equimolar mixtures of nucleotides. For many potential uses of random polypeptide libraries, concerns such as avoiding termination codons and matching target amino acid compositions make more complex designs necessary. In this study, three mixtures of nucleotides, corresponding to the three positions in the codon, were designed such that semirandom DNA synthesized by repeated cycles of the three mixtures created an open reading frame encoding random sequence polypeptides with desired ensemble characteristics. Two methods were used to design the nucleotide mixtures: the manual use of a spreadsheet and a refining grid search algorithm. Using design targets of less than or equal to 1% stop codons and an amino acid composition based on the average ratios observed in natural, globular proteins, the search methods yielded similar nucleotide ratios. Semirandom DNA, synthesized with a designed, three-residue repeat pattern, can encode libraries of very high diversity and represents an important tool for the construction of random polypeptide libraries.

Keywords: amino acid composition; DNA synthesis; gene library; nucleotide mixture; random sequence polypeptides

The utility of genetically encoded random sequence polypeptide (RSP) libraries as sources of interesting new molecules has previously been demonstrated (reviewed by Kauffman [1992] and Scott [1992]). The task of constructing gene libraries that encode RSPs is more complex than simply producing stochastic DNA. Transcription and translation of fully random DNA yields only rather short peptides, because 3 of the 64 codons (4.7%) signal termination of the chain. The assignment of a specific reading frame allows biasing the libraries in favor of particular traits and permits a more "informed" search for molecules possessing some desired property. The engineering problem then becomes the simultaneous design of nucleotide mixtures for each of the three positions in the randomized codons. Designing RSPs to contain particular amino acid compositions is a difficult problem because the nucleotide compositions that translate into a given amino acid composition are not obvious, and the enumeration of every possible set of nucleotide mixtures is not feasible.

Adjustable characteristics of RSP ensembles include sequence diversity, mean length, and amino acid composition. Generally, high diversity should be considered an asset. As diversity and sequence length increase, the fraction of possible sequences that can be effectively sampled diminishes, while the proportion and distribution of useful molecules is usually unknown. Examining representation and distribution questions remains a primary goal of work in this area. Limiting library diversity by imposing constraints on the coding sequences will have unknown effects on the proportion of useful sequences present. For some purposes, longer polypeptides may hold greater promise, but the avoidance of termination codons significantly affects the encoded ensemble amino acid composition. The relative importance of these and other design goals will be discussed below.

In previous studies, RSP libraries were constructed by synthesizing oligonucleotides with three-residue repeat patterns, either "NNK" or "NNS" (where N is all four

Reprint requests to Thomas H. LaBean at his present address: Department of Biochemistry, Duke University Medical Center, 211 Nanaline Duke Building, Box 3711, Durham, North Carolina 27710.

bases, K is T and G, and S is C and G, in all cases equimolar) (see, for example, Scott & Smith [1990]). When cloned in the proper reading frame, these schemes code for an ensemble of peptides containing all 20 amino acids. However, they encode over 3% stop codons. Mandecki (1990) reported a library of randomized genes constructed from DNA fragments containing segments of both "NNY" and "RNN" repeats (where Y represents the pyrimidines and R the purines, both equimolar), which were cloned into an expression system. This design completely eliminated stop codons; however, the polypeptide ensemble encoded fails to contain 2 of the 20 amino acids, 112 of the 400 pairs, and larger fractions of longer subsequences. Diversity is thereby limited. In the present study, we sought to increase RSP length, while maintaining diversity, by biasing the ratios of nucleotides at all three positions in the randomized codons.

Searching nucleotide composition space

A randomized codon corresponds to a point in nucleotide composition space, defined herein as the space of all possible sets of three nucleotide mixtures, $X_1X_2X_3$. Each point in nucleotide space specifies a list of probabilities for the codons and, therefore, values for amino acid and stop codon frequencies. Amino acid design targets will vary depending on the proposed use of the RSP library. The difference between target values and the encoded amino acid ratios corresponds to a "cost" that we seek to minimize. A cost score (C) can be given by:

$$C = \sum_{i=1}^{21} (t_i - e_i)^2, \tag{1}$$

where t_i and e_i are target and encoded values for the 20 amino acids and stop codons. This cost function, a sum of square differences, is defined for every point in nucleotide space and generates a surface that can be explored by various search methods. The deepest "valley" in the cost surface contains the nucleotide compositions that most closely match the design targets.

Design criteria will typically contain conflicting constraints, that is, two or more targets that cannot be met simultaneously. In this study the goals were to minimize stop codons and match amino acid frequencies observed in 207 natural proteins (Klapper, 1977). We partially offset the problem of conflicting constraints in the amino acid ratios by also examining some mean properties of the encoded amino acids. These "secondary descriptors," defined in Figure 1, are derived from the amino acid composition and include average net charge and fraction of exterior, interior, and ambivalent residues. A secondary descriptor (SD) cost function (sum of square differences between target and calculated SD values) codifies compensatory deviations in amino acid frequencies. That is, it allows a deficiency in one amino acid to be offset by an excess in a chemically similar amino acid. For example, glutamate can be replaced by aspartate, and leucine by valine without increasing the SD cost value.

Given design targets and cost functions, the problem becomes that of finding the point in nucleotide space that results in the lowest cost value. The space must be examined at relatively high resolution because small changes in nucleotide composition can have significant effects on some RSP ensemble characteristics. Depending on design criteria, the cost surface might be multipeaked; therefore, simple gradient searches may become arrested on local optima. Complete enumeration of the space is not feasible at 1% or 2% resolution, because the number of possible nucleotide compositions (N) increases as

$$N = \sum_{i=1}^{n+1} \frac{i(i+1)}{2}, \tag{2}$$

where n is the number of divisions or possible values for each dimension of the space. At 1% resolution (100 divisions) there are 176,851 compositions, therefore 5.5×10^{15} possible three-base combinations. Even if one could examine 1 million configurations per second, 175 years would be required to enumerate the entire space. Previously, the space of nucleotide compositions was exhaustively searched at 10% resolution for semirandom, site-directed mutagenesis (Arkin & Youvan, 1992). To examine higher-resolution answers, alternative search procedures were applied.

Nucleotide mixtures via spreadsheet

The use of a spreadsheet allows rational design of nucleotide compositions without definition of a single cost function, as required for complete automation of the searches. Design targets were prioritized as follows: code for no more than 1% termination codons and all 20 amino acids, balance internal versus external side chains, maintain a net charge near zero (the mean for natural globular proteins), and match the individual amino acids as nearly as possible to the target values, including consideration of compensatory deviations. The spreadsheet method is a hands-on approach that allows examination of generalized design targets (AA and SD costs) and individual amino acid frequencies simultaneously, at each step in the optimization.

A portion of the spreadsheet is reproduced in Figure 1, and a working copy of the Excel file (SUPLEMNT directory, file LaBean.xcl) is included on the Diskette Appendix along with more detailed instructions (SUPLEMNT directory, file LaBean.doc). The spreadsheet, written in Excel (Microsoft Corp., Redmond, Washington) on a

Nucleotide Mixtures (in Mole %)[a]

	Position 1	Position 2	Position 3
T	13	24	37
C	20	22	37
A	35	30	0
G	32	24	26

NCHRG (0)	0.11	e
EXT (34%)	33.42	
AMB (42%)	42.33	
INT (24%)	24.25	

AA Cost	83.0	f
SD Cost	0.1	

	Target AA Comp.[b]	Current AA Comp.[c]	% Difference[d]
Ala	8.9	7.04	-20.9
Cys	2.8	2.31	-17.5
Asp	5.5	7.10	29.2
Glu	6.2	2.50	-59.7
Phe	3.5	2.31	-34.0
Gly	7.8	7.68	-1.5
His	2.0	4.44	122.0
Ile	4.6	6.22	35.1
Lys	7.0	2.73	-61.0
Leu	7.5	5.61	-25.2
Met	1.7	2.18	28.5
Asn	4.4	7.77	76.6
Pro	4.6	4.40	-4.4
Gln	3.9	1.56	-60.0
Arg	4.7	6.98	48.6
Ser	7.1	9.08	27.8
Thr	6.0	7.70	28.3
Val	6.9	7.68	11.3
Trp	1.1	0.81	-26.2
Tyr	3.5	2.89	-17.5
stop	0	1.01	

Fig. 1. Sample spreadsheet showing the designed, input nucleotide mixtures and the output amino acid composition. [a] The input nucleotide compositions for the three positions of the codon. [b] The target amino acid composition given in percent. These values represent the average observed in 207 natural proteins (Klapper, 1977). [c] The output amino acid composition given the current nucleotide inputs. [d] The percent difference between the current and target amino acid compositions is: 100% × (current − target)/target. [e] Secondary descriptors of mean properties of the encoded proteins as defined by Zubay (1983) and calculated from the current amino acid composition. NCHRG is net charge per 100 residues: Lys + Arg − Asp − Glu (the target value is 0). EXT is the sum of the exterior (hydrophilic) amino acids: Asp, Glu, His, Lys, Asn, Gln, and Arg (target value is 34%). AMB represents ambivalent amino acids: Ala, Cys, Gly, Pro, Ser, Thr, Trp, and Tyr (target, 42%). INT is interior (hydrophobic) amino acids: Phe, Ile, Leu, Met, and Val (target, 24%). [f] Amino acid and secondary descriptor cost values. Sums of square differences between target and encoded values as defined in the text.

Macintosh computer (Apple Computer, Inc., Cupertino, California), contains a representation of the genetic code such that when the input nucleotide ratios are adjusted, the probability of encoding each triplet and the total for each amino acid are calculated. This list of probabilities is equivalent to the amino acid composition of the protein ensemble encoded by DNA specified by the input nucleotide ratios. Examination of the distribution of the amino acids within the genetic code typically suggests options for the direction of changes in the nucleotide mixtures. Amino acids with similar physical and chemical properties tend to be represented by neighboring triplets in the code (see, for example, Sjostrom & Wold [1985]). A three-dimensional representation of the genetic code is especially useful for visualizing these neighbor relations (Fig. 2).

After entering a change in the nucleotide mixtures, the calculated values for the individual targets and the cost functions were examined. The alteration was reversed or a new change was incorporated to offset some detrimental effects of the first. The process was iterated until no further overall improvements could be found. The final nucleotide composition derived from the spreadsheet (given in the outlined cells of Fig. 1) was a representative member of a family of answers that gave ensemble polypeptide characteristics surrounding the target values. The answers "surrounded" the target in the sense that if the answer was improved according to one criterion, it suffered a loss based on other measures.

Nucleotide mixtures by refining grid search

The optimization of input nucleotides was also investigated by exhaustive search within regions of the space of possible randomized codons at successively higher resolutions. A program written in C on a Sun 4 computer (Sun Microsystems, Inc.) was used to scan through all possible ratios of T, C, A, and G at a given resolution and within given ranges covering a promising region of nucle-

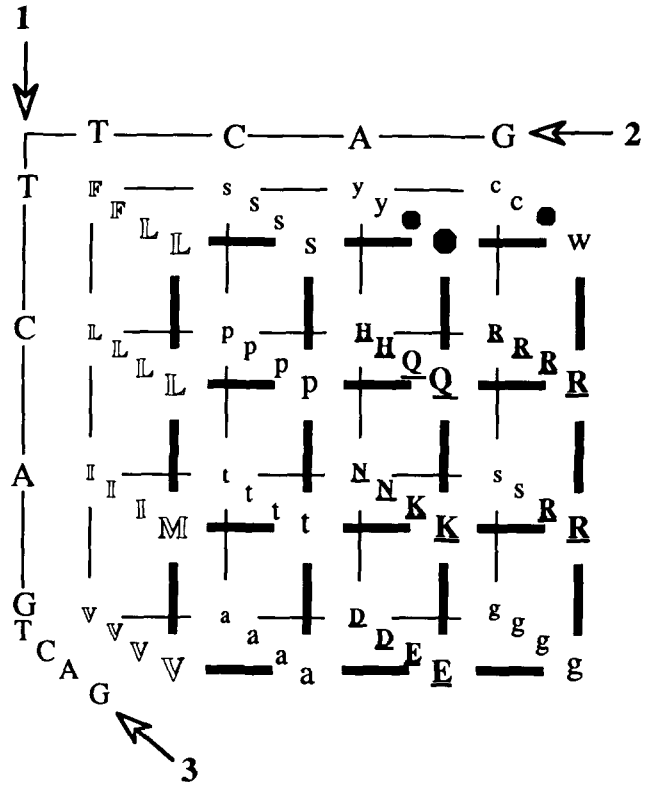

Fig. 2. A three-dimensional representation of the genetic code. Axes 1, 2, and 3 correspond to the first, second, and third positions in the codon, respectively. The single-letter code for the amino acids is used. Interior (hydrophobic) amino acids are given in outlined capital letters (e.g., 𝕃); exterior (hydrophilic) amino acids are in underlined capitals; ambivalent amino acids are in lower case; and termination codons are given as solid octagons. The clustering of similar amino acids as neighbors in the code is obvious from this representation, the most striking example being the left-hand face of the cube, described by NTN (all four nucleotides in positions 1 and 3, but only T in position 2), which contains all the interior amino acids. Similarly, the TNN plane (top face of cube) contains all three termination codons as well as the aromatic amino acids. NCN contains only ambivalent amino acids, and NAN contains mostly exterior amino acids (but also two termination codons. It is also evident in this representation that T and C are equivalent in the third position, whereas A and G in position 3 exhibit differences in 2 of the 16 possible cases. This view of the genetic code is invaluable for use during the design of nucleotide mixtures with the Excel spreadsheet.

otide space. The program calculated the resultant amino acid composition and secondary descriptors mentioned above. The decision to score the input ratio as promising or not was based on empirically determined threshold values for three separate cost functions: the AA and SD costs plus a cut-off for termination codon frequency. The threshold values were decreased for successive searches such that approximately 500–1,000 nucleotide sets were saved as promising in each run. In the final run, the threshold values were ≤ 70 for the amino acid cost score (Equation 1, above) and ≤ 10 for the secondary descriptor cost, with a termination probability of $\leq 1\%$ per codon. The first run searched the entire space at a resolution of 10%, and the final run looked in a restricted area of space at 1% resolution.

During each stage of the search, the best answers were saved and used to define promising regions for the next, higher-resolution run. For example, in the initial, low-resolution run the best configurations (top ~2,000 answers) contained either 0, 10, or 20% T in the first position. Likewise, for each nucleotide in each position, a range of acceptable values was tabulated. For the subsequent search, the ranges were expanded by one step in each direction and then the step size was decreased by half. Interestingly, the midpoints of the ranges from the low-resolution search were nearly equal to the final answer in the high-resolution search, and for all searches the ranges of acceptable values were continuous rather than broken. This implies that, for this set of criteria, the fitness function in nucleotide composition space is globally smooth.

The set of nucleotide compositions that gave the best score in the search program was: first position 8% T, 21% C, 32% A, 39% G; second position 24% T, 25% C, 28% A, 23% G; and third position 60% T, 0% C, 0% A, 40% G. These values are very near those arrived at using the spreadsheet method. Note that in the third position, T and C are interchangeable.

Discussion of results and comparison with previous designs

Neither of the optimization methods used here is guaranteed to find the global optimum answer, but the fact that two very different methods arrived at essentially the same solution supports the validity of the solution. The design targets and cost functions used herein seem to result in a single peak in nucleotide composition space. This topology simplified the use of the spreadsheet. For other targets that may result in a multipeaked landscape, the spreadsheet may prove more difficult to use, but the refining grid search would likely remain useful. We are continuing to investigate optimizations in nucleotide space using other design targets and automated search techniques including gradient hill climbing and other directed walks, a genetic algorithm, and numerical methods for minimization of continuous cost functions (Tozier & LaBean, in prep.).

The method presented here represents significant improvements over previous RSP library designs. The design method is capable of producing longer polypeptides with desired balances of amino acids and greater sequence diversity, including representation of dipeptide, tripeptide, and higher-order subsequences in high diversity. Table 1 compares the makeup of polypeptide ensembles encoded by DNA containing various three-base repeat patterns. The $X_1X_2X_3$ DNA (from the refining grid search) encodes an ensemble of polypeptides that more

Table 1. *Resultant amino acid compositions from three-base-repeat DNA*

Amino acid	Target[a]	NNN[b]	NNK or NNS	NNY + RNN[c]	$X_1X_2X_3$[d]
Ala	8.9	6.2	6.2	9.4	9.8
Cys	2.8	3.1	3.1	3.1	1.1
Asp	5.5	3.1	3.1	6.3	6.6
Glu	6.2	3.1	3.1	3.1	4.4
Phe	3.5	3.1	3.1	3.1	1.2
Gly	7.8	6.2	6.2	9.4	9.0
His	2.0	3.1	3.1	3.1	3.5
Ile	4.6	4.7	3.1	7.8	4.6
Lys	7.0	3.1	3.1	3.1	3.6
Leu	7.5	9.4	9.4	3.1	5.8
Met	1.7	1.6	3.1	1.6	3.1
Asn	4.4	3.1	3.1	6.2	5.4
Pro	4.6	6.2	6.2	3.1	5.2
Gln	3.9	3.1	3.1	0.0	2.4
Arg	4.7	9.4	9.4	6.2	7.8
Ser	7.1	9.4	9.4	9.4	6.4
Thr	6.0	6.2	6.2	9.4	8.0
Val	6.9	6.2	6.2	9.4	9.4
Trp	1.1	1.6	3.1	0.0	0.7
Tyr	3.5	3.1	3.1	3.1	1.3
STOP	0.0	4.7	3.1	0.0	0.9
Net charge	0.0	6.2	6.2	−0.1	0.4
Exterior	33.7	29.5	29.0	28.0	33.8
Ambivalent	42.1	44.3	45.2	46.9	42.0
Interior	24.2	26.2	26.8	25.1	24.2
AA cost score		99.4	95.2	106.8	63.0
SD cost score		65.4	72.8	56.2	0.2

[a] Target refers to the composition from Klapper (1977).
[b] Abbreviations: N = T, C, A, and G; K = T + G; S = C + G; Y = T + C; R = A + G (these are equimolar in each case).
[c] NNY + RNN from Mandecki (1990). Genes containing both "NNY" and "RNN" segments. See description in the text.
[d] The repeat pattern designed during the refining grid search. The definitions for net charge, exterior, interior, and ambivalent are given in Figure 1. The cost scores are explained in the text.

closely resemble natural proteins in amino acid composition and balance. The AA and SD cost scores are given in the last rows of the table. Libraries of stochastic proteins with low cost scores are appropriate arenas in which to search for novel, useful molecules.

Screening an ensemble of random sequence polypeptides is equivalent to examining an arbitrary sample of all possible such sequences, thus random protein libraries are useful as samples for examination of overarching properties of proteins. Our purpose in designing RSP libraries was the construction of a large random sample of protein sequence space with which to investigate folding of unevolved sequences. Solutions of nucleotide precursors, as designed here, were premixed and used during the elongation step in oligonucleotide synthesis. The resulting DNA was cloned into an expression system such that the three-base repeat aligned with the established reading frame. Several such libraries have been constructed, expressed, and purified in this laboratory. They have been examined to determine the extent to which random polypeptides with amino acid compositions similar to those of evolved proteins are capable of folding (LaBean et al., 1992; LaBean, Kauffman, & Butt, in prep.). RSP libraries can be designed for various uses and have other, advantageous biases built in. The general design strategy outlined here is applicable to a wide range of output libraries.

Acknowledgments

We thank Dr. Tauseef R. Butt for encouragement and support, David Penkower for writing the grid search program, and M. McLean Bolton, William A. Tozier, and T.R.B. for critical discussion of this manuscript. This work was supported in part by NIH grant 5-R01-GM-40186-03.

References

Arkin, A.P. & Youvan, D.C. (1992). Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis. *Biotechnology 10*, 297–300.

Kauffman, S.A. (1992). Applied molecular evolution. *J. Theor. Biol. 157*, 1–7.
Klapper, M.H. (1977). The independent distribution of amino acid near neighbor pairs into polypeptides. *Biochem. Biophys. Res. Commun. 78*, 1018–1024.
LaBean, T.H., Kauffman, S.A., & Butt, T.R. (1992). Design, expression, and characterization of random sequence polypeptides as fusions with ubiquitin. *FASEB J. 6*(1), A471.
Mandecki, W. (1990). A method for construction of long randomized open reading frames and polypeptides. *Protein Eng. 3*, 221–226.
Scott, J.K. (1992). Discovering peptide ligands using epitope libraries. *Trends Biochem. Sci. 17*, 241–245.
Scott, J.K. & Smith, G.P. (1990). Searching for peptide ligands with an epitope library. *Science 249*, 386–390.
Sjostrom, M. & Wold, S. (1985). A multi-variate study of the relationship between the genetic code and the physical-chemical properties of amino acids. *J. Mol. Evol. 22*, 272–277.
Zubay, G.L. (1983). *Biochemistry*. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts.