Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Optimal codon randomization via mathematical programming

Yuval Nov*, Danny Segev

Department of Statistics, University of Haifa, Israel

HIGHLIGHTS

- To induce a uniform distribution over the 20 amino acids, 4 conventional oligos are required.
- When spiked oligos are allowed, the number drops to 3.
- Integer Programming can be used to compute the required number for any target distribution.

ARTICLE INFO

Article history:
Received 12 February 2013
Accepted 28 May 2013
Available online 17 June 2013

Keywords:
Codon randomization
Degenerate oligonucleotides
Integer programming
Protein engineering
Spiked oligonucleotides

ABSTRACT

Codon randomization via degenerate oligonucleotides is a widely used approach for generating protein libraries. We use integer programming methodology to model and solve the problem of computing the minimal mixture of oligonucleotides required to induce an arbitrary target probability over the 20 standard amino acids. We consider both randomization via conventional degenerate oligonucleotides, which incorporate at each position of the randomized codon certain nucleotides in equal probabilities, and randomization via spiked oligonucleotides, which admit arbitrary nucleotide distribution at each of the codon's positions. Existing methods for computing such mixtures rely on various heuristics.

1. Introduction

In protein engineering, oligonucleotides carrying biased codons at certain predetermined positions are routinely used to generate protein libraries. Most often, all 20 standard amino acids are encoded, typically via NNK or NNS codons (using the standard IUPAC codes for degenerate codons: N=A/C/G/T, K=G/T, S=C/G; Cornish-Bowden, 1985). This type of codon bias induces a more uniform distribution across the 20 standard amino acids, compared to the distribution induced by the non-biased NNN codon, and lowers the probability of a premature stop codon; both features improve the library's coverage of protein sequence space (Patrick and Firth, 2005). In other cases, only some of the 20 amino acids are allowed to be encoded, according to various physico-chemical consideration; such restricted amino-acid alphabets significantly reduce the complexity of the resulting libraries, and thus make them easier to explore in the laboratory (Hokanson et al., 2011; Tanaka et al., 2010; Reetz et al., 2008; Gilbreth et al., 2008; Fellouse et al., 2004).

Two techniques provide additional flexibility in shaping codon bias beyond conventional codon degeneracy. The first is the use of so-called "spiked" or "doped" oligonucleotides, whereby during DNA synthesis, non-equimolar proportions of the four bases are used at some – or all – of the codon's three nucleotide positions (Neylon, 2004). In the second technique, several degenerate oligonucleotides are mixed together at certain predetermined molar ratios (see below).

The mathematical study of codon bias and its effect on the resulting amino-acid diversity dates back at least two decades. Arkin and Youvan (1992) used an exhaustive search over discretized doping space to optimize the equiprobability of a target amino-acid subset, based on a "group probability" score. Tomandl et al. (1997) combined search via a genetic algorithm with local optimization, to identify doping mixtures that reverse translate a target amino-acid subset. Jensen et al. (1998) used a simulated annealing search guided by three scoring functions, to optimize the design of spiked oligonucleotides. Other related works include those of Siderovski and Mak (1993), Ophir and Gershoni (1995), Wolf and Kim (1999), Mena and Daugherty (2005), and Firth and Patrick (2008).

Recently, Tang et al. (2012) showed that a mixture of four conventional degenerate oligonucleotides – NDT, VMA, ATG, and TGG, at 12:6:1:1 molar ratio – can generate a perfectly uniform distribution across all 20 amino acids, while reducing to zero the

* Corresponding author. Tel.: +972 4 8240203.
E-mail addresses: yuval@stat.haifa.ac.il (Y. Nov), segevd@stat.haifa.ac.il (D. Segev).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.05.034

probability of a premature stop codon. To determine the appropriate mixtures for other target amino-acid subsets, they provided a software tool termed "DC-analyzer," which uses an exhaustive search approach combined with a greedy speed-up procedure, to search the space of relevant mixtures. Kille et al. (2013) showed that the number of degenerate oligonucleotides can be reduced to three – NDT, VHG, and TGG, at 12:9:1 molar ratio – at the expense of a slight deviation from perfect uniformity.

Notwithstanding the sizable literature on the topic, the computational methods used in the aforementioned works are either heuristic, or rely on a discretization of a continuous search space. Thus, to the best of our knowledge, no fully rigorous mathematical approach has been presented yet to determine the minimal number of degenerate oligonucleotides (either conventional or spiked) to be mixed, so as to induce an arbitrary given distribution across the 20 amino acids. The goal of this work is to bridge this gap, and to offer a rigorous mathematical methodology to do so, as well as to address a number of related problems.

Our main vehicle for carrying out this plan is mathematical programming, and in particular, integer programming (Wolsey and Nemhauser, 1988; Schrijver, 1998; Bertsimas and Weismantel, 2005). While many researchers have used mathematical programming in their study of protein engineering (e.g., Saraf et al., 2005, 2006; Bellows and Floudas, 2010; Parker et al., 2011; Chen et al., 2013), we are not aware of any previous study that used this mathematical tool to model and solve the specific above mentioned problems related to codon randomization.

2. Methods

2.1. Notation and terminology

Let $\mathcal{A}=\{\text{Ala, Arg}, \ldots, \text{Val, Stp}\}$ be the set of the 20 standard amino acids *and* the stop codon, with p_a being the target probability of amino acid $a \in \mathcal{A}$, and let $\mathcal{B}=\{A, C, G, T\}$ be the set of the four bases comprising the DNA alphabet

We use the term *conventional degenerate oligonucleotide* to denote an oligonucleotide such as NNS or NNK, i.e., one whose degeneracy arises from allowing only a subset of the four bases at some (or all) of the codon's three positions, and where the allowed bases appear in *equal probabilities*. Technically, an oligonucleotide such as ACG is not degenerate, but we still refer to it as a conventional degenerate oligonucleotide. We denote by $\mathcal{D}$ the set of all conventional degenerate oligonucleotides; since there are 15 non-empty subsets of $\mathcal{B}$, we have $|\mathcal{D}|=15^3=3375$. In contrast to conventional degenerate oligonucleotides, oligonucleotides in which bases are allowed to appear in non-equal probabilities (e.g., 50% A, 25% C, and 25% G at some position) will be referred to as *spiked oligonucleotides*.

For a degenerate oligonucleotide $d \in \mathcal{D}$ and an amino acid $a \in \mathcal{A}$, we let $\pi_{d,a}$ denote the fraction of the codons belonging to d that encode a. For example, the degenerate oligonucleotide AGV (where V=A/C/G) includes the non-degenerate codons AGA, AGG (both encoding Arg), and AGC (encoding Ser), and hence $\pi_{\text{AGV,Arg}}=2/3$, $\pi_{\text{AGV,Ser}}=1/3$, and $\pi_{\text{AGV},a}=0$ for any other amino acid a. Finally, for $a \in \mathcal{A}$, we let $C(a)$ be the set of codons encoding a. For example, $C(\text{Asp})=\{\text{GAT, GAC}\}$.

2.2. Minimizing the number of conventional degenerate oligonucleotides

In this subsection we formulate a mixed integer linear program to answer the following question: When mixing several conventional degenerate oligonucleotides, what is the minimal number of conventional degenerate oligonucleotides required to induce a given set of target probabilities $\{p_a,\ a \in \mathcal{A}\}$ over the 20 amino acids and the stop signal, and what are the corresponding fractions of the degenerate oligonucleotides involved in the mixture?

Let λ_d be the decision variable denoting the fraction of degenerate oligonucleotide $d \in \mathcal{D}$ in the mixture, and let x_d be a binary decision variable indicating whether degenerate oligonucleotide d is included in the mixture ($x_d=1$) or not ($x_d=0$). The mathematical program required to answer the above question is the following:

$$\min \sum_{d \in \mathcal{D}} x_d \tag{P1}$$

$$\begin{aligned} \text{s.t.}\quad &(1)\quad \sum_{d \in \mathcal{D}} \pi_{d,a}\lambda_d = p_a && \forall a \in \mathcal{A} \\ &(2)\quad \lambda_d \le x_d && \forall d \in \mathcal{D} \\ &(3)\quad \sum_{d \in \mathcal{D}} \lambda_d = 1 && \\ &(4)\quad \lambda_d \in [0,1] && \forall d \in \mathcal{D} \\ &(5)\quad x_d \in \{0,1\} && \forall d \in \mathcal{D} \end{aligned}$$

From the very definition of the decision variables x_d, the sum $\sum_{d \in \mathcal{D}} x_d$ is the number of degenerate oligonucleotides involved in the mixture, and therefore this sum is the objective function to be minimized. Constraint (1) guarantees that the resulting fraction of each amino acid a will equal the target fraction p_a. Constraint (2) guarantees that if the fraction λ_d of degenerate oligonucleotide d in the mixture is positive, this will be reflected by adding d to the mixture, i.e., setting $x_d=1$. There is no need to add a constraint to force $x_d=0$ when $\lambda_d=0$, as any solution with $x_d=1$ and $\lambda_d=0$ cannot be optimal: the value of the objective function in such a case can be reduced by one by setting $x_d=0$, without violating any of the constraints. Constraints (3) and (4) guarantee that the mixing proportions λ_d, $d \in \mathcal{D}$, constitute a proper probability distribution, and constraint (5) guarantees that the decision variables x_d are binary.

2.3. Minimizing the number of spiked oligonucleotides

Suppose now that one is allowed to use not only conventional degenerate oligonucleotides, but also arbitrarily spiked ones. In this subsection we formulate a nonlinear mathematical program that will allow us to determine the minimal number of spiked oligonucleotides required to induce a given set of target probabilities $\{p_a, a \in \mathcal{A}\}$ over the 20 amino acids, as well as the corresponding fractions of the spiked oligonucleotides involved in the mixture, and the spiking probabilities in each. Since conventional degenerate oligonucleotides are a special case of spiked oligonucleotides, the objective value (minimal number of oligonucleotides) of problem (P_1) provides an upper bound to the objective value of this problem.

A spiked oligonucleotide may be represented via a triplet $(\beta_1,\ \beta_2,\ \beta_3)$, where each $\beta_i=(\beta_{i,A},\ \beta_{i,C},\ \beta_{i,G},\ \beta_{i,T})$ is a probability distribution over the four DNA bases at position i in the codon to be randomized. Note that while there is a finite number of conventional degenerate oligonucleotides (namely, 3375), there are infinitely many spiked oligonucleotides, as each of the three β_i assumes values in the continuum of the 3-dimensional simplex. However, it is easy to establish an upper bound M on the number of spiked oligonucleotides involved in the optimal mixture: $M=20$ is a trivial such bound, as clearly, 20 non-degenerate oligonucleotides are enough to encode all 20 amino acids, let alone a subset thereof. We exploit this bound in the formulation below.

Let $\beta_{i,b}^{d}$ be the decision variable indicating the probability of base b in position i of spiked oligonucleotide d. The other decision variables, λ_d and x_d, are similar in role to their counterparts in problem (P_1). The mathematical program is the following:

$$\min \quad \sum_{d=1}^{M} x_d \tag{P2}$$

$$\begin{aligned}
\text{s.t.}\quad &(1)\quad \sum_{d=1}^{M} \lambda_d \sum_{b_1 b_2 b_3 \in \mathcal{C}(a)} \beta^d_{1,b_1}\beta^d_{2,b_2}\beta^d_{3,b_3} = p_a \quad \forall a\in\mathcal{A}\\
&(2)\quad \lambda_d \le x_d \qquad \forall 1\le d\le M\\
&(3)\quad \sum_{b\in\mathcal{B}} \beta^d_{i,b} = 1 \qquad \forall 1\le d\le M,\ \ 1\le i\le 3\\
&(4)\quad \beta^d_{i,b}\in[0,1] \qquad \forall 1\le d\le M,\ \ 1\le i\le 3,\ \ b\in\mathcal{B}\\
&(5)\quad \sum_{d=1}^{M} \lambda_d = 1\\
&(6)\quad \lambda_d\in[0,1] \qquad \forall 1\le d\le M\\
&(7)\quad x_d\in\{0,1\} \qquad \forall 1\le d\le M.
\end{aligned}$$

The objective function is identical to the one in problem (P_1), as the objective has not changed: minimizing the number of oligonucleotides in the mixture. Constraint (1) guarantees that the resulting fraction of each amino acid a will equal the target fraction p_a. To see this, note that by the independence of randomization across the codon's three positions, the product $\beta^d_{1,b_1}\beta^d_{2,b_2}\beta^d_{3,b_3}$ is the probability that the codon $b_1b_2b_3$ will be formed from spiked oligonucleotide d; summing these probabilities over all codons encoding amino acid a, gives the probability of getting amino acid a from spiked oligonucleotide d. Constraint (2) is identical to constraint (2) in problem (P_1), and guarantees that $x_d = 1$ whenever $\lambda_d > 0$. Constraints (3) and (4) guarantee that for each position i in each spiked oligonucleotide d, the 4-vector $(\beta^d_{i,A},\ \beta^d_{i,C},\ \beta^d_{i,G},\ \beta^d_{i,T})$ is a proper distribution. Constraints (5) and (6) guarantee that the λ_d constitute a proper probability distribution, and finally, constraint (7) guarantees that the decision variables x_d are binary.

Note that the resulting problem is neither linear nor convex: each of the summands in the left-hand side of constraint (1) is a product of four decision variables (one λ and three β's).

2.4. Allowing small deviations

Consider again the problem described in Section 2.2, but this time suppose that instead of having to match exactly the target probabilities, a small additive deviation of magnitude at most ϵ_a is allowed for each amino acid a. Biological experiments are notoriously noisy, so slight deviations from the theoretical desired probabilities can be tolerated if the number of oligonucleotides in the mixture is reduced.

The only required modification is to replace constraint (1) in problem (P1) with

$$\Big|\sum_{d\in\mathcal{D}} \pi_{d,a}\lambda_d - p_a| \le \epsilon_a, \quad \forall a\in\mathcal{A}. \tag{1}$$

However, this constraint is not linear. A well known trick to linearize such a constraint is based on the observation that $|x|\le c$ if and only if both $x\le c$ and $-x\le c$. Thus, the non-linear constraint in Eq. (1) can be replaced by two linear constraints, as in the following program:

$$\min \quad \sum_{d\in\mathcal{D}} x_d \tag{P3}$$

$$\begin{aligned}
\text{s.t.}\quad &(1a)\quad \sum_{d\in\mathcal{D}} \pi_{d,a}\lambda_d - p_a \le \epsilon_a \qquad \forall a\in\mathcal{A}\\
&(1b)\quad p_a - \sum_{d\in\mathcal{D}} \pi_{d,a}\lambda_d \le \epsilon_a \qquad \forall a\in\mathcal{A}\\
&(2)\quad \lambda_d \le x_d \qquad \forall d\in\mathcal{D}\\
&(3)\quad \sum_{d\in\mathcal{D}} \lambda_d = 1\\
&(4)\quad \lambda_d\in[0,1] \qquad \forall d\in\mathcal{D}\\
&(5)\quad x_d\in\{0,1\} \qquad \forall d\in\mathcal{D}.
\end{aligned}$$

2.5. Minimizing deviations

A dual problem, in a sense, to the one discussed in the previous subsection, is the following: what is the minimal deviation from the target probabilities that can be attained when using a given number K of conventional degenerate oligonucleotides?

To better define this problem, one needs a measure for the magnitude of the *joint* deviation from the target probabilities p_a, across all amino acids a. We consider three such measures, which are all norms of the deviation vector $\Delta = (\Delta_a,\ a\in\mathcal{A})$, where $\Delta_a = |\sum_{d\in\mathcal{D}} \pi_{d,a}\lambda_d - p_a|$. Each such norm may be viewed as a penalty function over the deviations, which is to be minimized.

For brevity of the presentation, we provide the full program only for the L_2 (Euclidean) norm, defined by $\|\Delta\|_2 = (\sum_{a\in\mathcal{A}} \Delta_a^2)^{1/2}$. This norm is closely related to the fitness function used by Tomandl et al. (1997) and to the s score function of Jensen et al. (1998). When using this norm, the following quadratic integer program is to be solved:

$$\min \quad \sum_{a\in\mathcal{A}} \Delta_a^2 \tag{P4}$$

$$\begin{aligned}
\text{s.t.}\quad &(1a)\quad \sum_{d\in\mathcal{D}} \pi_{d,a}\lambda_d - p_a \le \Delta_a \qquad \forall a\in\mathcal{A}\\
&(1b)\quad p_a - \sum_{d\in\mathcal{D}} \pi_{d,a}\lambda_d \le \Delta_a \qquad \forall a\in\mathcal{A}\\
&(2)\quad \lambda_d \le x_d \qquad \forall d\in\mathcal{D}\\
&(3)\quad \sum_{d\in\mathcal{D}} \lambda_d = 1\\
&(4)\quad \sum_{d\in\mathcal{D}} x_d \le K\\
&(5)\quad \lambda_d\in[0,1] \qquad \forall d\in\mathcal{D}\\
&(6)\quad x_d\in\{0,1\} \qquad \forall d\in\mathcal{D}.
\end{aligned}$$

To minimize an L_1 norm of the deviations, defined via $\|\Delta\|_1 = \sum_{a\in\mathcal{A}} |\Delta_a|$, all that needs to be changed is to replace the objective function to $\sum_{a\in\mathcal{A}} \Delta_a$. The L_1 norm is equivalent to the SAE (sum of absolute errors) criterion used by Tomandl et al. (1997).

A modification of a different nature is required to accommodate an L_∞ ("sup") norm, defined in this case via $\|\Delta\|_\infty = \max_{a\in\mathcal{A}} |\Delta_a|$. The change is based on the observation that for a real-valued vector $(x_1, \ldots, x_n)$, we have $\|x\|_\infty \le c$ if and only if $|x_i| \le c$ for each $i = 1, \ldots, n$. To adapt the program (P_4) to an L_∞ norm, we therefore introduce an additional decision variable, $\Delta_{\max}$, add the constraint $\Delta_a \le \Delta_{\max} \forall a\in\mathcal{A}$, and change the objective function to that of minimizing $\Delta_{\max}$.

All norms can be generalized to be weighted, in a similar fashion to the scoring functions used by Tomandl et al. (1997) and Jensen et al. (1998). To do so, a set of positive parameters w_a, $a\in\mathcal{A}$, needs to be added to the model, and the objective function in the L_2 case, for example, needs to be changed to $\min \sum_{a\in\mathcal{A}} w_a \Delta_a^2$.

2.6. Computational notes

In general, optimization problems involving integrality constraints are NP-hard, and thus intractable. However, the specific instances presented above are small enough (in terms of the number of decision variables and the number of constraints) so that all but one of them can be solved by commercial solvers in a fraction of a second. The only exception is problem (P_4) with the L_2 norm, which required several hours of computation time on a standard PC.

All problems were modeled using AMPL (Fourer et al., 2003) and solved by the CPLEX solver (http://www-01.ibm.com/software/integration/optimization/cplex-optimizer), except for problem (P_2) which was solved by the KNITRO solver (http://www.ziena.com/knitro.html).

3. Results

Table 1 shows the optimal mixtures of conventional oligonucleotides required to induce a target probability distribution over select subsets of amino acids (problem (P_1)). In the first four cases, the target distribution is uniform, whereas the fifth case is a non-uniform distribution used by Gilbreth et al. (2008). It can be seen that the minimal number of conventional degenerate oligonucleotides required to induce a uniform distribution over all 20 amino acids is four, confirming that the mixture computed by Tang et al. (2012) is optimal. The optimal solution is not unique, hence the difference between Tang et al.'s solution and the solution in Table 1.

In some cases, the number of oligonucleotides to be mixed could be reduced further if spiked oligonucleotides are allowed (problem (P_2)). Table 2 shows the spiking probabilities and the mixture proportions for two such cases.

Table 1
Optimal mixtures of conventional degenerate oligonucleotides required to induce a uniform distribution across subsets of amino acids (first four cases) and a non-uniform target distribution (last case).

Subset	Degeneracy	Proportions
All amino acids	HAT VMR WKK GDY	3/20 6/20 8/20 3/20
Aliphatic (Ala, Ile, Leu, Pro, Val)	SYN ATA	4/5 1/5
Aromatic (Phe, Tyr, Trp, His)	YAY TGG TTT	2/4 1/4 1/4
Polar (Arg, Lys, Asp, Glu, Asn, Gln)	RAM CRA	2/3 1/3
Gilbreth et al. (2008) (40% Tyr, 20% Ser, 10% Gly, 5% Arg, Leu, His, Asp, Asn, Ala)	VRY GST TMC TYG TAY	3/10 1/10 2/10 1/10 3/10

Table 2
Optimal mixtures of spiked oligonucleotides required to induce two target distributions.

Subset	Proportion	Position	Spiking probabilities			
			A	C	G	T
All amino acids	1/20	1	1/3	1/3	1/3	
		2	2/3	1/3		
		3	2/4	1/4		1/4
	10/20	1	2/5		1/5	2/5
		2			1/2	1/2
		3		1/4	2/4	1/4
	1/20	1				1
		2	1			
		3				1
Gilbreth et al. (2008) (See Table 1)	0.55	1	0.0909	0.0909	0.0909	0.7273
		2	1			
		3		1		
	0.3426	1	0.6	0.1079	0.2921	
		2			1	
		3	0.0271	0.9728		
	0.0575	1		1		
		2			0.1301	0.8699
		3		1		
	0.05	1			1	
		2		1		
		3	1			

Fig. 1 shows the minimal number of conventional degenerate oligonucleotides required to approximate a uniform distribution over the 20 amino acids, up to a deviation of $\epsilon_a = \epsilon$ for each amino acid, as a function of ϵ. Each point in the graph, therefore, corresponds to a solution of problem (P_3), for a different value of ϵ.

Table 3 shows the minimal norm of the deviations from a uniform distribution across the 20 amino acids, when using $K=1$, 2, 3 conventional degenerate oligonucleotides, for the L_1 norm function. Tables 4 and 5 are similar for the L_2 and L_∞ norm functions, respective. For $K \geq 4$, all three norms are zero, as four conventional degenerate oligonucleotides are enough to induce a perfectly uniform distribution. Also shown are the optimal randomization schemes, and the resulting induced probabilities. All values in the table were computed from solutions of problem

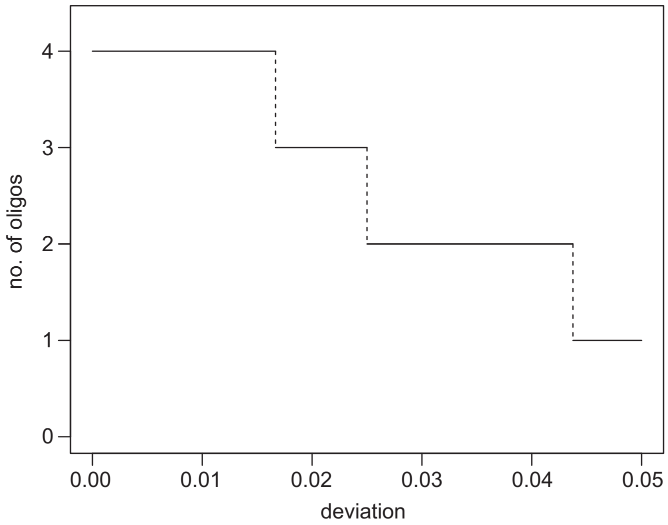

Fig. 1. Minimal number of conventional degenerate oligonucleotides in the mixture, as a function of ϵ, the common allowed deviation from a uniform distribution across the 20 amino acids.

Table 3
Optimal mixtures of $K=1$, 2, and 3 conventional degenerate oligonucleotides, required to minimize the L_1 deviation from a uniform distribution across the 20 amino acids.

K	1	2	3
oligos	NNB 1	NDY 3/5 VHG 2/5	NNY 16/20 VAA 3/20 TGG 1/20
L_1 norm	0.4083	0.1778	0.1000
Ala	0.0625	0.0444	0.05
Cys	0.0417	0.0500	0.05
Asp	0.0417	0.0500	0.05
Glu	0.0208	0.0444	0.05
Phe	0.0417	0.0500	0.05
Gly	0.0625	0.0500	0.05
His	0.0417	0.0500	0.05
Ile	0.0417	0.0500	0.05
Lys	0.0208	0.0444	0.05
Leu	0.0833	0.0944	0.05
Met	0.0208	0.0444	0.00
Asn	0.0417	0.0500	0.05
Pro	0.0625	0.0444	0.05
Gln	0.0208	0.0444	0.05
Arg	0.0833	0.0500	0.05
Ser	0.1042	0.0500	0.10
Thr	0.0625	0.0444	0.05
Val	0.0625	0.0944	0.05
Trp	0.0208	0.0000	0.05
Tyr	0.0417	0.0500	0.05
Stp	0.0208	0.0000	0.00

Table 4
Similar to Table 3, for L_2 norm.

K	1	2	3
Oligos	NNB 1	NDB 0.68365 VMD 0.31635	DDT 0.42953 VMD 0.36171 WKG 0.20876
L_2 norm	0.1054	0.0680	0.0468
Ala	0.0625	0.0527	0.0603
Cys	0.0417	0.0380	0.0477
Asp	0.0417	0.0556	0.0678
Glu	0.0208	0.0541	0.0402
Phe	0.0417	0.0380	0.0477
Gly	0.0625	0.0570	0.0477
His	0.0417	0.0556	0.0201
Ile	0.0417	0.0380	0.0477
Lys	0.0208	0.0541	0.0402
Leu	0.0833	0.0760	0.0522
Met	0.0208	0.0190	0.0522
Asn	0.0417	0.0556	0.0678
Pro	0.0625	0.0527	0.0603
Gln	0.0208	0.0541	0.0402
Arg	0.0833	0.0760	0.0522
Ser	0.1042	0.0380	0.0477
Thr	0.0625	0.0527	0.0603
Val	0.0625	0.0570	0.0477
Trp	0.0208	0.0190	0.0522
Tyr	0.0417	0.0380	0.0477
Stp	0.0208	0.0190	0.0000

Table 5
Similar to Table 3, for L_∞ norm.

K	1	2	3
Oligos	NNK 1	NDK 3/5 VMV 2/5	DKS 16/20 VMS 3/20 WDB 1/20
L_∞ norm	0.0438	0.0250	0.0167
Ala	0.0625	0.0667	0.0667
Cys	0.0312	0.0250	0.0583
Asp	0.0312	0.0472	0.0333
Glu	0.0312	0.0694	0.0333
Phe	0.0312	0.0250	0.0583
Gly	0.0625	0.0500	0.0500
His	0.0312	0.0472	0.0333
Ile	0.0312	0.0250	0.0583
Lys	0.0312	0.0694	0.0500
Leu	0.0938	0.0750	0.0417
Met	0.0312	0.0250	0.0417
Asn	0.0312	0.0472	0.0667
Pro	0.0625	0.0667	0.0667
Gln	0.0312	0.0694	0.0333
Arg	0.0938	0.0750	0.0417
Ser	0.0938	0.0250	0.0583
Thr	0.0625	0.0667	0.0667
Val	0.0625	0.0500	0.0500
Trp	0.0312	0.0250	0.0417
Tyr	0.0312	0.0250	0.0333
Stp	0.0312	0.0250	0.0167

(P_4) and its two variants for the L_1 and L_∞ norms. Note that the three L_∞ norm values in Table 5 are exactly the jump points of the graph in Fig. 1, as expected.

4. Discussion

The goal of this work is to model and analyze in a fully rigorous way certain problems associated with codon randomization, which were previously treated only via heuristic approaches. Our result regarding the optimal mixture of conventional degenerate oligonucleotides (when the target distribution is uniform across all 20 amino acids) proves that four is the minimal number of degenerate oligonucleotides required, so that the mixture reported by Tang et al. (2012) cannot be improved further in this sense.

When allowing spiked oligonucleotides, the number of required oligonucleotides is reduced further to three. Tomandl et al. (1997) used their GALO algorithm to discover a 3-oligonucleotide solution for this problem, but did not provide its details (spiking probabilities and mixture proportions, as shown in Table 2).

We used the spiking model to study only the problem of minimizing the number of spiked oligonucleotides required to induce exactly a target probability distribution (problem (P_2)). With modifications similar to those described in Sections 2.4 and 2.5, this model can also be applied toward approximating a target probability distribution (problem (P_3)), or minimizing deviations under a given number of spiked oligonucleotides (problem (P_4) and its L_1 and L_∞ variants). We note, however, that while the problem instances considered in the context of conventional degenerate oligonucleotides are tractable enough so that the solvers used compute the globally optimal solutions, the spiking model is inherently more difficult to optimize (recall that problem (P_2) is neither linear nor convex), and hence the resulting solutions are guaranteed to be only locally optimal.

Tang et al. (2012) explicitly avoided in their algorithm degenerate codons that code for either of eight rare codons of *E. Coli* (CGA, CGG, AGA, AGG for Arginine; CTA for Leucine; ATA for Ile; GGA for Gly, and CCC for Proline). To incorporate such a constraint into our conventional degenerate oligonucleotide framework, all variables λ_d and x_d corresponding to degenerate oligonucleotides d that encode rare codons need to be removed from the program, which is equivalent to eliminating these codons in advance from the set $\mathcal{D}$. Stop codons can also be avoided in a similar manner, rather than through setting $p_{\text{stp}}=0$, as done above. The resulting programs are substantially smaller, making them even easier to solve. To avoid rare codons in the spiked oligonucleotides model, one should add the constraint

$$\beta^d_{1,b_1}\beta^d_{2,b_2}\beta^d_{3,b_3}=0, \quad \forall 1\le d\le M, \ \ b_1b_2b_3\in F,$$

where F is the set of the codons to be avoided.

References

Arkin, A., Youvan, D., 1992. Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis. Nat. Biotechnol. 10 (3), 297–300.

Bellows, M., Floudas, C., 2010. Computational methods for de novo protein design and its applications to the human immunodeficiency virus 1, purine nucleoside phosphorylase ubiquitin specific protease 7, and histone demethylases. Curr. Drug Targets 11 (3), 264.

Bertsimas, D., Weismantel, R., 2005. Optimization Over Integers. Dynamic Ideas, Belmont, MA.

Chen, T., Palacios, H., Keating, A., 2013. Structure based re-design of the binding specificity of anti-apoptotic Bcl-x$_L$. J. Mol. Biol. 425, 171–185.

Cornish-Bowden, A., 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 13 (9), 3021–3030.

Fellouse, F., Wiesmann, C., Sidhu, S., 2004. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. Proc. Natl. Acad. Sci. USA 101 (34), 12467–12472.

Firth, A.E., Patrick, W.M., 2008. GLUE-IT and PEDEL-AA: new programmes for analyzing protein diversity in randomized libraries. Nucleic Acids Res. 36 (Suppl. 2), W281–W285.

Fourer, R., Gay, D.M., Kernighan, B.W., 2003. AMPL: A Modeling Language for Mathematical Programming. Duxbury Press/Brooks/Cole Publishing Company.

Gilbreth, R., Esaki, K., Koide, A., Sidhu, S., Koide, S., 2008. A dominant conformational role for amino acid diversity in minimalist protein–protein interfaces. J. Mol. Biol. 381 (2), 407–418.

Hokanson, C., Cappuccilli, G., Odineca, T., Bozic, M., Behnke, C., Mendez, M., Coleman, W., Crea, R., 2011. Engineering highly thermostable xylanase variants using an enhanced combinatorial library method. Protein Eng. Des. Sel. 24 (8), 597–605.

Jensen, L., Andersen, K., Svendsen, A., Kretzschmar, T., 1998. Scoring functions for computational algorithms applicable to the design of spiked oligonucleotides. Nucleic Acids Res. 26 (3), 697–702.

Kille, S., Acevedo-Rocha, C., Parra, L., Zhang, Z., Opperman, D., Reetz, M., Acevedo, J., 2013. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. ACS Synth. Biol..

Mena, M., Daugherty, P., 2005. Automated design of degenerate codon libraries. Protein Eng. Des. Sel. 18 (12), 559–561.

Neylon, C., 2004. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. Nucleic Acids Res. 32 (4), 1448–1459.

Ophir, R., Gershoni, J., 1995. Biased random mutagenesis of peptides: determination of mutation frequency by computer simulation. Protein Eng. 8 (2), 143–146.

Parker, A., Griswold, K., Bailey-Kellogg, C., 2011. Optimization of combinatorial mutagenesis. J. Comput. Biol. 18 (11), 1743–1756.

Patrick, W.M., Firth, A.E., 2005. Strategies and computational tools for improving randomized protein libraries. Biomol. Eng. 22 (4), 105–112.

Reetz, M., Kahakeaw, D., Lohmer, R., 2008. Addressing the numbers problem in directed evolution. ChemBioChem 9 (11), 1797–1804.

Saraf, M., Gupta, A., Maranas, C., 2005. Design of combinatorial protein libraries of optimal size. Proteins: Struct. Funct. Bioinformatics 60 (4), 769–777.

Saraf, M., Moore, G., Goodey, N., Cao, V., Benkovic, S., Maranas, C., 2006. Ipro: an iterative computational protein library redesign and optimization procedure. Biophys. J. 90 (11), 4167–4180.

Schrijver, A., 1998. Theory of Linear and Integer Programming. John Wiley and Sons.

Siderovski, D., Mak, T., 1993. Ramha: a PC-based Monte-Carlo simulation of random saturation mutagenesis. Comput. Biol. Med. 23 (6), 463–474.

Tanaka, J., Doi, N., Takashima, H., Yanagawa, H., 2010. Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. Protein Sci. 19 (4), 786–795.

Tang, L., Gao, H., Zhu, X., Wang, X., Zhou, M., Jiang, R., 2012. Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. BioTechniques 52, 149–158.

Tomandl, D., Schober, A., Schwienhorst, A., 1997. Optimizing doped libraries by using genetic algorithms. J. Comput.-Aided Mol. Des. 11 (1), 29–38.

Wolf, E., Kim, P., 1999. Combinatorial codons: a computer program to approximate amino acid probabilities with biased nucleotide usage. Protein Sci. 8 (3), 680–688.

Wolsey, L.A., Nemhauser, G.L., 1988. Integer and Combinatorial Optimization. John Wiley and Sons.