Decodon Calculator: Degenerate Codon Set Design for Protein Variant Libraries

Dimitris Papamichail[*]
papamicd@tcnj.edu
The College of New Jersey
Ewing, New Jersey, USA

Nicholas Carpino
The College of New Jersey
Ewing, New Jersey, USA

Tomer Aberbach
The College of New Jersey
Ewing, New Jersey, USA

Georgios Papamichail
New York College
Athens, Greece

1 INTRODUCTION

Mutant libraries representing protein variants are increasingly used to optimize protein function. Protein Engineering involves screening mutant libraries for novel proteins that show enhanced expression levels, solubility, stability, or enzymatic activity. To reach such objectives, it is often necessary to modify extant proteins, developing variants with improved properties [3, 4]. However, there exists a massive space of potential variants to consider.

Computational design of combinatorial libraries [1, 2, 6, 7] provides a reasonable approach in the development of improved variants. Library-design strategies seek to experimentally evaluate a diverse but focused region of sequence space in order to improve the likelihood of finding a beneficial variant. Such an approach is based on the premise that prior knowledge can inform generalized predictions of protein properties, but may not be sufficient to specify individual, optimal variants. Libraries are particularly appropriate when the prior knowledge does not admit detailed, robust modeling of the desired properties, but when experimental techniques are available to rapidly assay a pool of variants.

The design of mutant protein libraries typically involves a manual process in which required sites for mutation are selected and ambiguous *degenerate* codons (those containing mixtures of nucleotides) are designed to introduce controlled variation in these positions. This is particularly useful in cases where definitive decisions regarding specific amino acid substitutions are non-obvious [4]. The design of the protein variant library is complemented by use of synthesized degenerate oligonucleotides which enable annealing based recombination. Custom oligonucleotide overlaps enable the targeted introduction of crossovers at only desired positions, in turn enabling the desired level and type of diversity in a combinatorial library.

Table 1: Degenerate Bases and their codings

Degenerate Base	Actual Bases Coded
N	A or C or G or T
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
K	G or T
M	A or C
R	A or G
S	C or G
W	A or T
Y	C or T

2 THE PROBLEM: TARGETED MUTANT PROTEIN LIBRARIES

Traditional mutant protein library design methods involve the incorporation of a single degenerate codon (thereafter referred to as *decodon*) at each position where amino acid substitutions are explored. Decodons contain ambiguous bases (*degenerate* bases), as shown in Table 1.

An online tool called CodonGenie [5] was created to aid the effort of designing decodons that code for any provided set of amino acids. The CodonGenie tool ranks candidate decodons by specificity, attempting to minimize coding of undesirable amino acids and/or STOP codons. Even so, when using a single decodon to code for a set of amino acids, it is often unavoidable to code for additional unwanted amino acids. Using an example from [5], when coding the non-polar residues A, F, G, I, L, M and V, CodonGenie picks decodon DBK ([AGT][CGT][GT]) as its top choice, which, in addition to the desired set, codes also for amino acids C, R, S, T, and W. In total, the decodon DBK codes 26 total DNA variants, 18 DNA variants coding for desired amino acids, and 8 DNA variants for undesirable ones.

[*] Corresponding author

In our work we explored the coding of a set of amino acids by potentially multiple decodons. The usage of annealing based recombination of degenerate oligos containing the decodons can produce libraries on the productive portion of the space by eliminating unwanted mutations, therefore improving the yield of beneficial variants and the overall quality of the library. In turn, this method can significantly reduce labor costs assaying the pool of variants, at the expense of additional oligo synthesis, whose comparative cost is modest and continuously dropping.

The Decodon Calculator Tool

We have designed and implemented an algorithm that, given any set of amino acids, produces the minimum number of decodons necessary to code for exactly this set, i.e. without coding for extraneous amino acids or STOP codons. There are 15 nucleotide codes ("letters"), ranging from the completely unambiguous A, C, G and T representing a single nucleotide, to the completely ambiguous N representing all 4 nucleotides. There are $15^3 = 3,375$ decodons that can be assembled from this 15-letter alphabet of ambiguous codes, compared to the 43 = 64 codons that can be constructed from the standard 4-letter alphabet of unambiguous nucleotides.

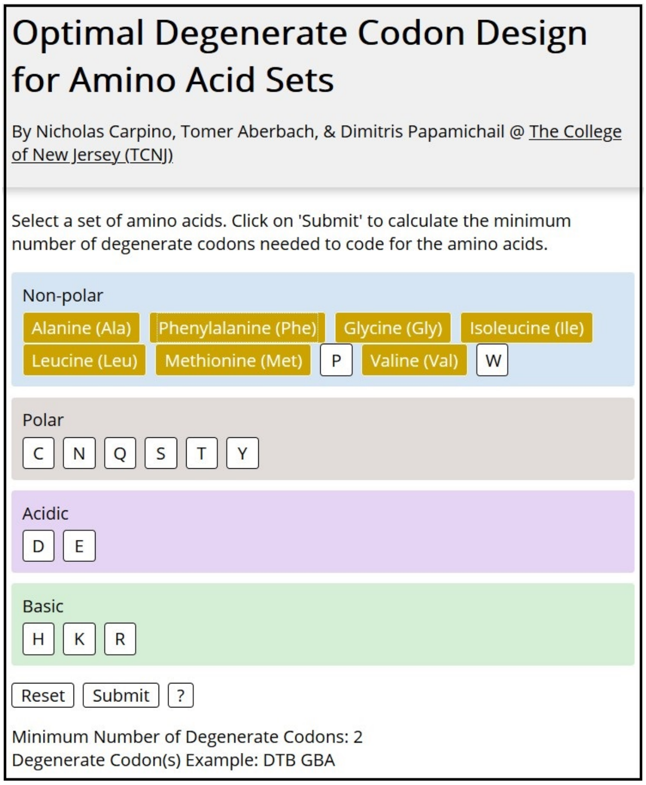
Using our algorithm we calculated minimum cardinality decodon sets for all $2^{20} - 1 = 1,048,575$ possible amino acid subsets. Our results indicate that 6 decodons are always sufficient to code for any amino acid subset, where at most 4 decodons are sufficient to encode more than 90% of all amino acid subsets. Our algorithm also produces an example of a decodon set of minimum cardinality for each amino acid subset.

We also built a web tool called *Decodon Calculator* that allows the calculation of the minimum number of decodons needed to code any amino acid subset. Once a set of amino acids is selected and the Submit button is pressed, results are displayed on the bottom of the screen, as shown in in Figure 1. In this particular example, we can observe that the non-polar residues A, F, G, I, L, M and V can be coded by the two degenerate codons DTB and GBA, which code for 12 desirable DNA variants, in contrast to the 26 variants of the single best decodon generated by CodonGenie, 8 of which are undesirable.

The Decodon Calculator can be accessed at http://algo.tcnj.edu/decodoncalc/.

3 ACKNOWLEDGEMENTS

This work has been supported in part by NSF Grant CCF-1418874. The authors also acknowledge use of the ELSA high performance computing cluster at The College of New Jersey for hosting the web service reported in this paper. This cluster is funded by the NSF grant OAC-1828163.



Optimal Degenerate Codon Design for Amino Acid Sets

By Nicholas Carpino, Tomer Aberbach, & Dimitris Papamichail @ The College of New Jersey (TCNJ)

Select a set of amino acids. Click on 'Submit' to calculate the minimum number of degenerate codons needed to code for the amino acids.

Minimum Number of Degenerate Codons: 2
Degenerate Codon(s) Example: DTB GBA

Figure 1: Calculating the minimum number of decodons necessary to encode the amino acid set { }

REFERENCES

[1] Meyer, M. M., Silberg, J. J., Voigt, C. A., Endelman, J. B., Mayo, S. L., Wang, Z.-G., and Arnold, F. H. Library analysis of SCHEMA-guided protein recombination. *Protein Science* (2003).

[2] Pantazes, R. J., Saraf, M. C., and Maranas, C. D. Optimal protein library design using recombination or point mutations based on sequence-based scoring functions. *Protein Engineering, Design and Selection* (2007).

[3] Parker, A. S., Zheng, W., Griswold, K. E., and Bailey-Kellogg, C. Optimization algorithms for functional deimmunization of therapeutic proteins. *BMC Bioinformatics* (2010).

[4] Reetz, M. T., and Carballeira, J. D. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nature Protocols* (2007).

[5] Swainston, N., Currin, A., Green, L., Breitling, R., Day, P. J., and Kell, D. B. CodonGenie: Optimised ambiguous codon design tools. *PeerJ Computer Science* (2017).

[6] Treynor, T. P., Vizcarra, C. L., Nedelcu, D., and Mayo, S. L. Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proceedings of the National Academy of Sciences of the United States of America* (2007).

[7] Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L., and Arnold, F. H. Protein building blocks preserved by recombination. *Nature Structural Biology* (2002).