Combinatorial Codons

What input does one need to create a set of random sequences with a desired amino acid composition?

Litterature

Arkin and Youvan chaim that using 25% of each base in synthesis is inefficient [1].
Wolf and Kim made an online tool, which run via a matlab analogue [2] (NB. E. Wolf is not P Wolfe of the Frank-Wolfe alogorith).
Craig et al prove that the task is impossible to get a near perfect match and use a gene algorimth for the search and use 6 separate codon mixes ("tubes") to meet their requirements [3].

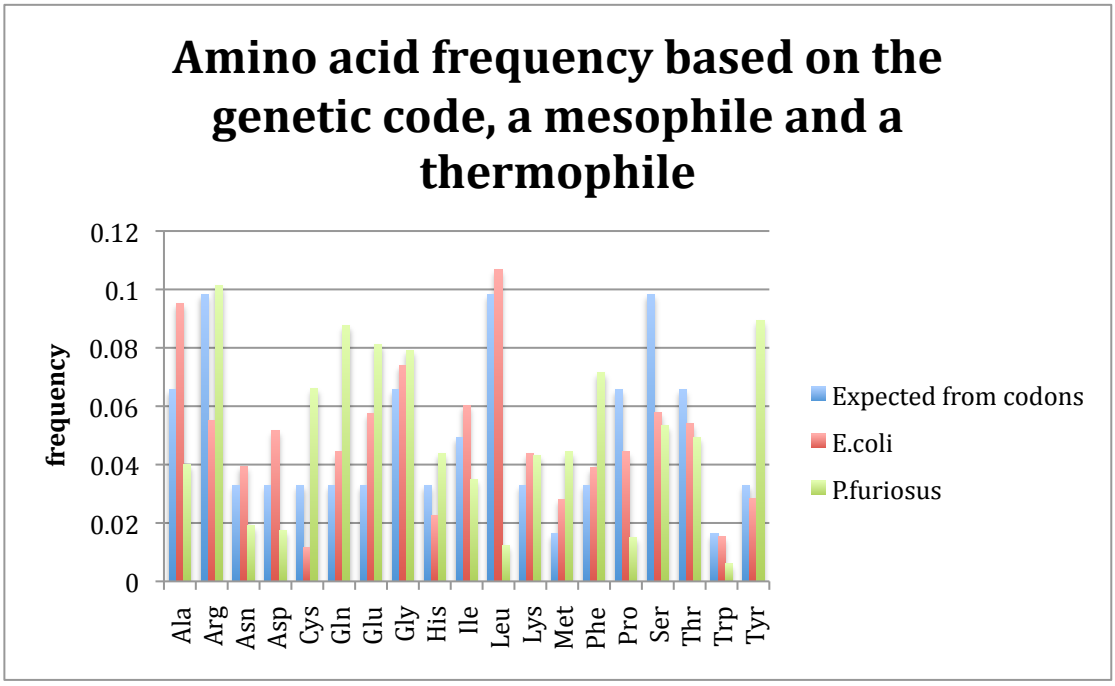
Code

This is an optimization problem, where:

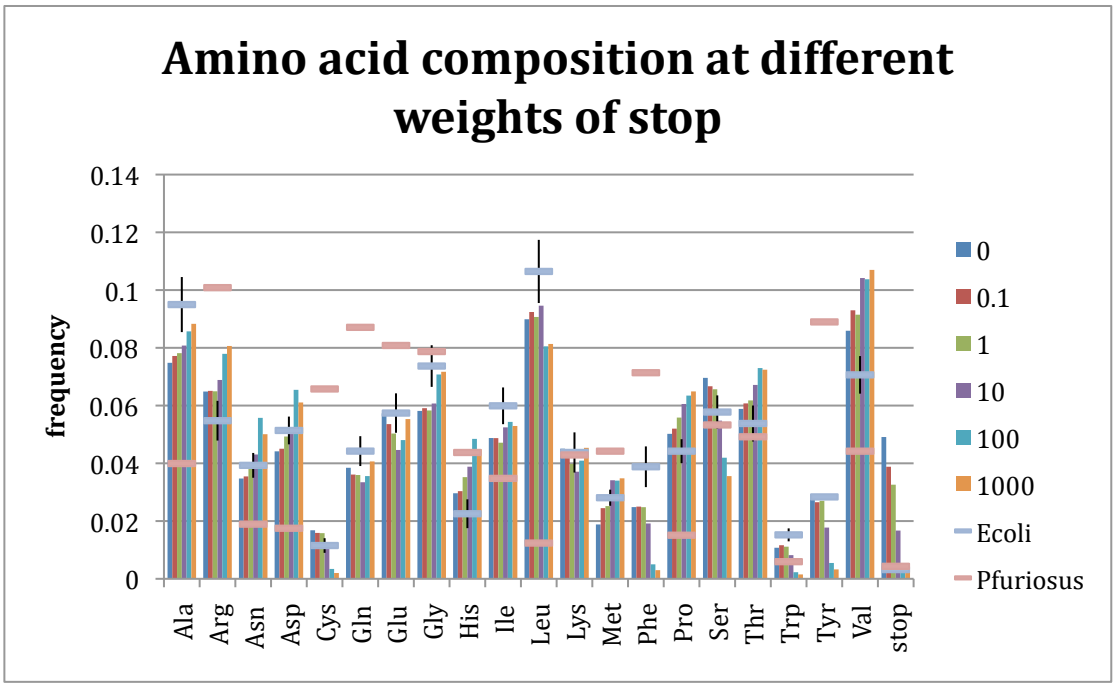
- The matrix to be optimized is the frequency of each base for position 1, 2 & 3 with the constraint that the frequencies are positive and the sum of each base at a position is 1. For simplicity it can be expressed as a 12-dimentional vector.
- The best result is determined by the objective function, for a list of them see [4] or [3]. The most common is the least squares approach, where the residues of the fit are minimized. Wolf uses a cosine basin approach.
- Search method.
    - Systematic. for the XYZ situation it takes 30-60 minutes when searching all combinations of 0.1 increments.
    - Heuristic. Several algorithms exist.
        - Wolf uses an algorithm called CFSQP, a very fancy version of Newton method from what I understand.
        - Craig uses a gene algorithm.
        - For simplicity I used a hill-climbing algorithm.
- A translation function to translate the base vector into the amino acid vector.

Stop codons

The genetic code has been described by Crick as a frozen accident.
The number of codons and the amino acid frequency unfortunately do not correlate too well (fig 1) otherwise this problem would be non existant



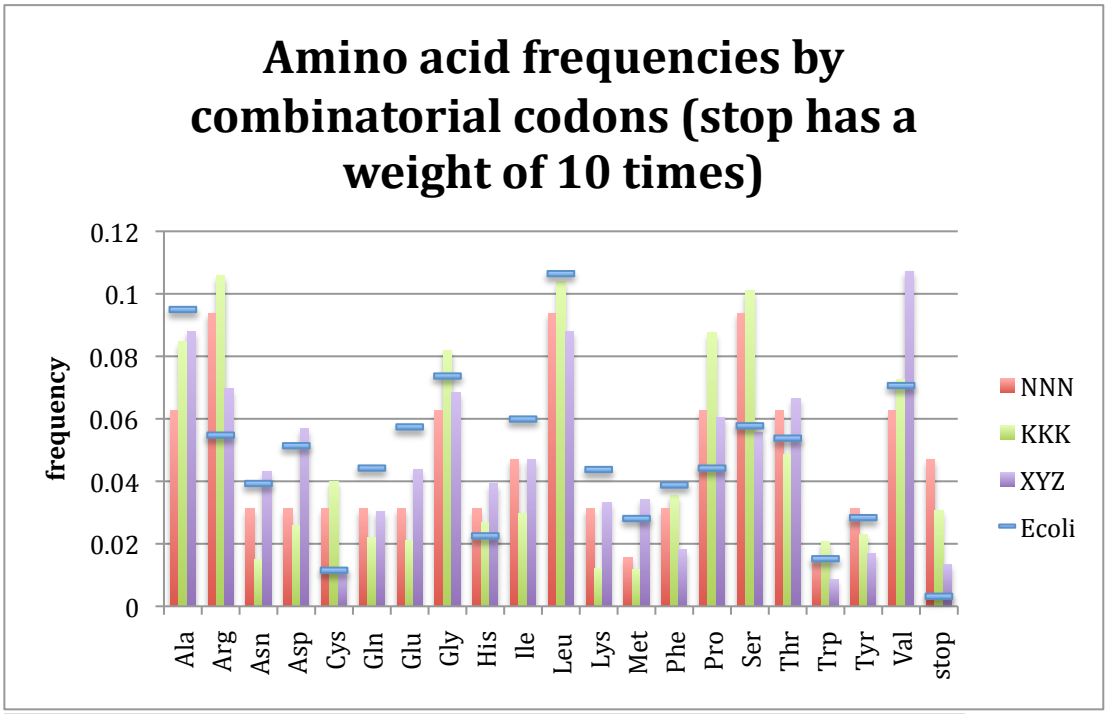
The largest disparity are the 3 stop codons where the average E.coli protein length is 300 amino acids and not 20.
Do stop codons mess stuff up?
They do but it is not such a good fit ignoring them:

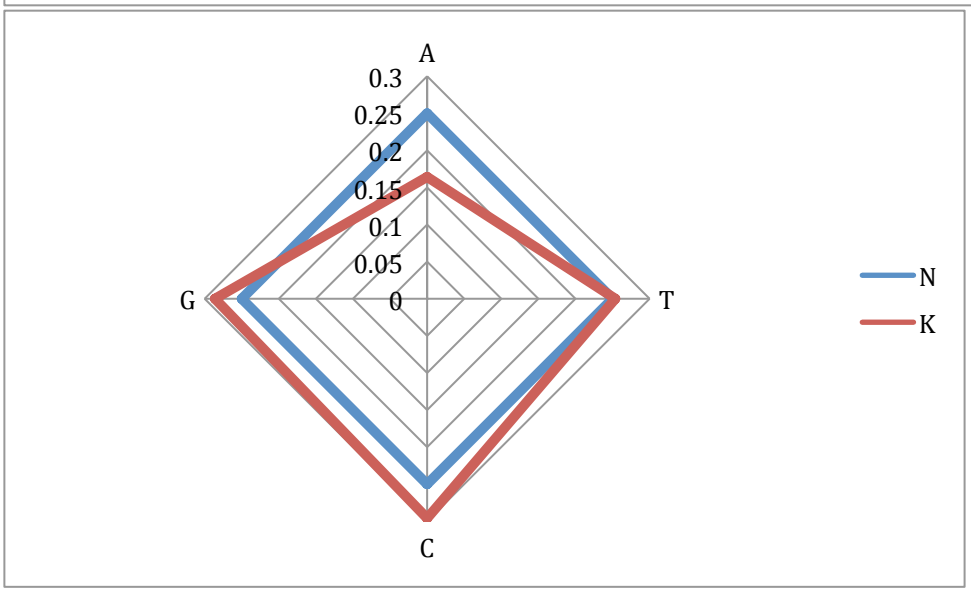


P.furious is a thermophile and it is the graph just to show that there is a large diversity in composition. To achieve the correct amount of stop codons, Cys, Phe, Trp and Tyr are lost.

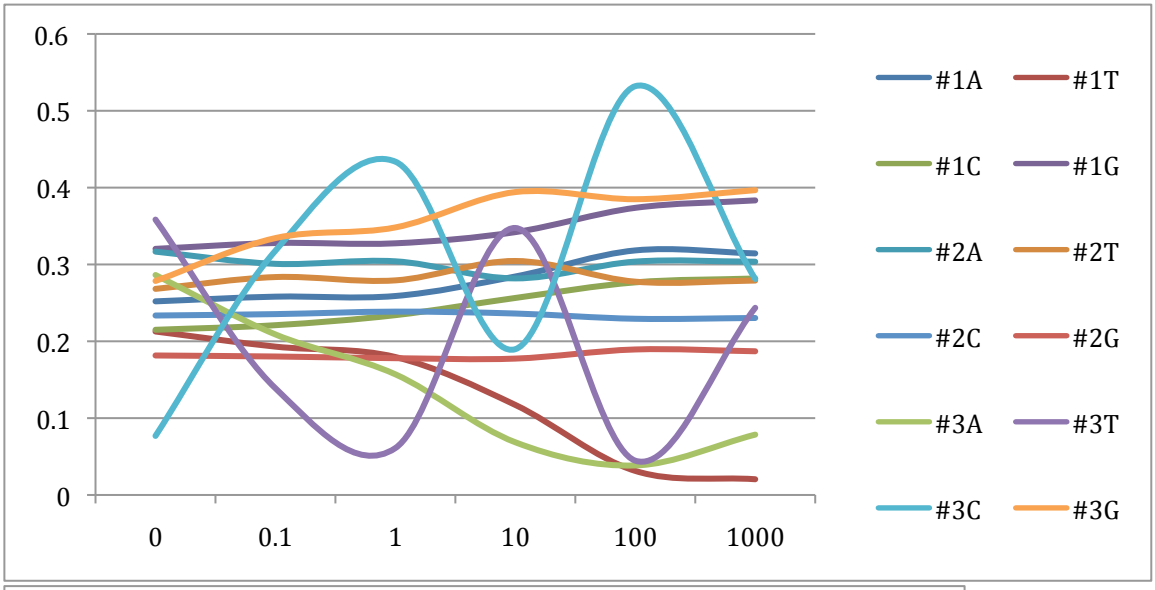
More data beating the bush…

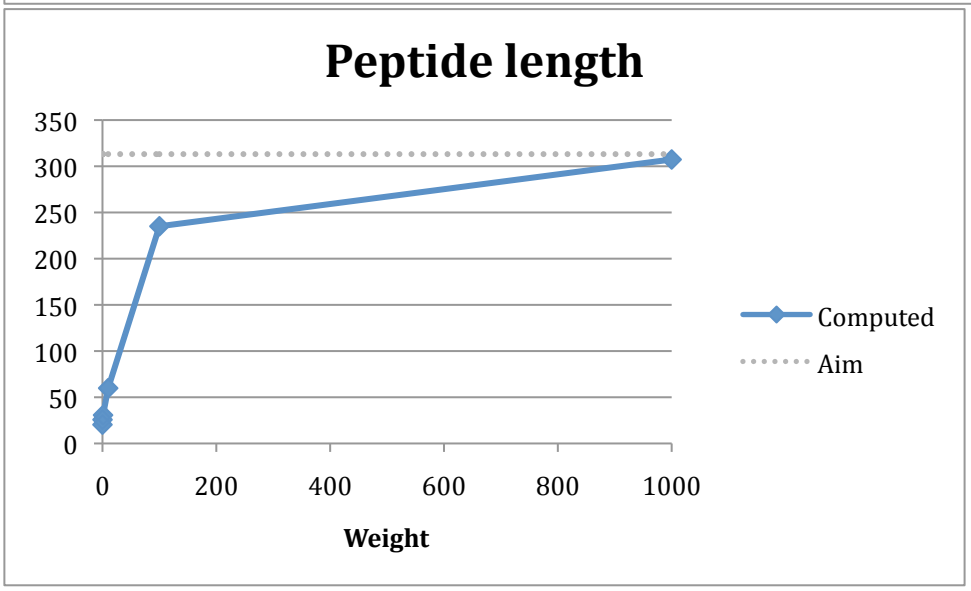
In the KKK model, there is absurdly high amount of proline, as the proportion of A decreases rapidly followed by T (stop = (A:0.5 T:0.3 G:0.2 C:0). Base frequency will be represented as radar plots for ease of visualization (fig 3).



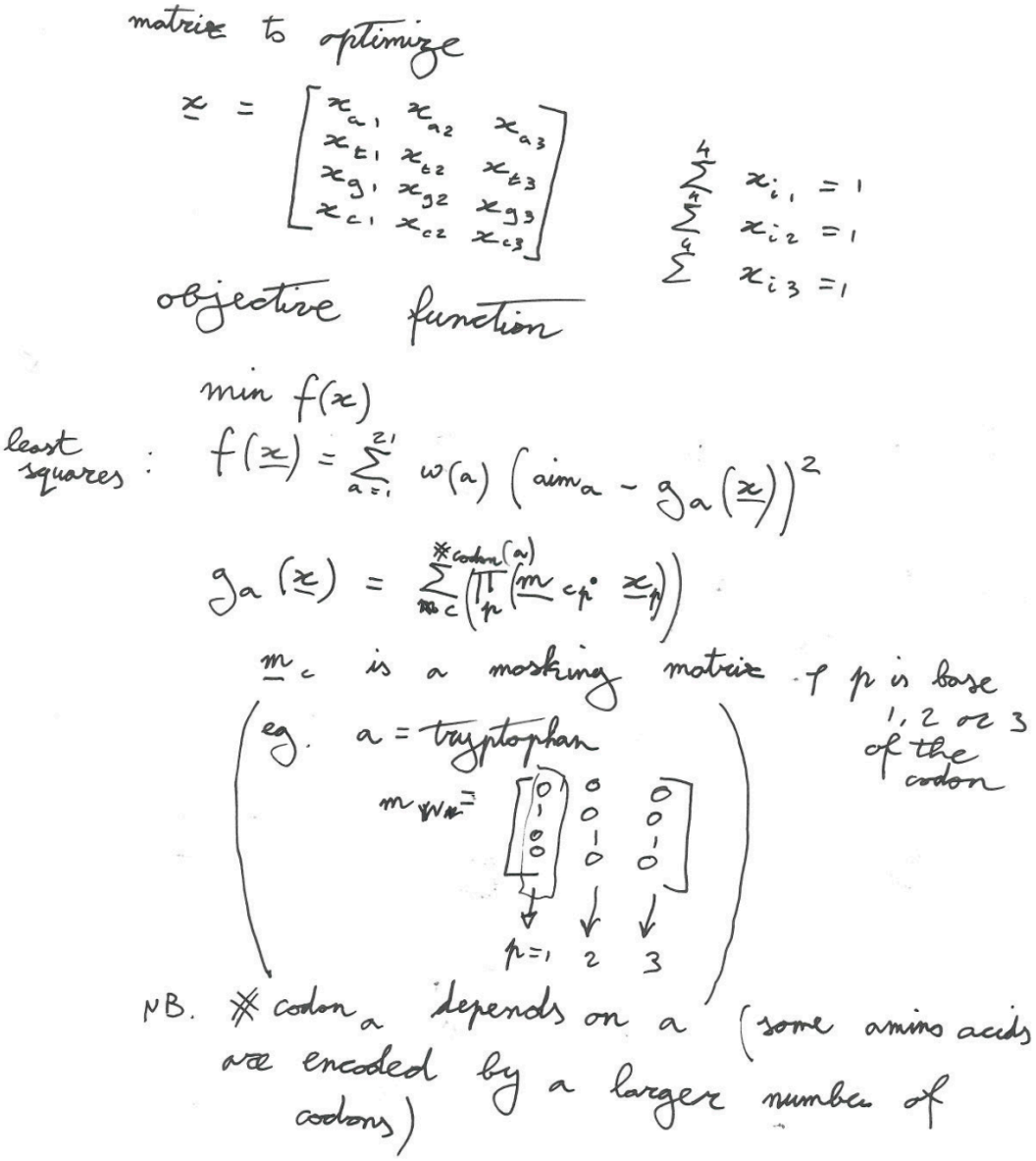


The graph below shows the shift in bases according to several weigths on stop (0, 0.1, 1, 10, 100, 1000). At the third position it does not matter if there is a T or C but A is not allowed there as is T in the first codon.


0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
1
10
100
1000
#1A
#1T
#1C
#1G
#2A
#2T
#2C
#2G
#3A
#3T
#3C
#3G


Peptide length
350
300
250
200
150
100
50
0
0
200
400
600
800
1000
Weight
Computed
Aim

Notes:



1. Arkin, A.P. and D.C. Youvan, *Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis.* Biotechnology (N Y), 1992. **10**(3): p. 297-300.
2. Wolf, E. and P.S. Kim, *Combinatorial codons: a computer program to approximate amino acid probabilities with biased nucleotide usage.* Protein Sci, 1999. **8**(3): p. 680-8.
3. Craig, R.A., et al., *Optimizing nucleotide sequence ensembles for combinatorial protein libraries using a genetic algorithm.* Nucleic Acids Research, 2010. **38**(2): p. e10.
4. Wang, W. and J.G. Saven, *Designing gene libraries from protein profiles for combinatorial protein experiments.* Nucleic Acids Research, 2002. **30**(21): p. e120.