Inverse translation

Sequences

Input amino acid sequence: $\mathbf{A} = A_1 \dots A_N$
Output RNA sequence: $\mathbf{R} = R_1 \dots R_{3N}$
Let $C_n = R_{3n-2}R_{3n-1}R_{3n}$ denote the n'th codon.

Genetic code

Let $G(a) = \{x_1y_1z_1, x_2y_2z_2 \dots\}$ be the set of codons xyz that translate to amino acid a.
For an output $\mathbf{R}$ to be valid for a given input $\mathbf{A}$, we require that $C_n \in G(A_n)$ for all n.

Codon frequencies

Along with the codon usage table, we are given a probability distribution over codons, $q(xyz)$, reflecting the frequency with which each codon xyz is observed in coding sequence from some specified target organism.
Since q is a probability: $0 \leq q(xyz) \leq 1$ and $\sum_{xyz} q(xyz) = 1$.
Let α be a penalty for using rare codons.

Repeated codons

Let $D = \{n : 1 \leq n \leq N, \exists m < n, C_m = C_n\}$ be the set of all input positions where a codon is duplicated.
For $n \in D$, let $d(n) = \max\{m : m < n, C_m = C_n\}$ be the most recent position using the same codon.
Let β be a penalty for duplicated codons.

RNA structure

Let $F(\mathbf{X})$ denote the maximal free energy of folding (in kCal/mol) for sequence $\mathbf{X}$, at room temperature, using version 2.2 of the ViennaRNA RNAfold package.
Let γ be a penalty for RNA structure.

Runtime

Let T be a program's runtime (in seconds), and δ a time penalty.

Scoring function

A valid output is scored using the scoring function

$$S(\mathbf{R}|\mathbf{A}) = \alpha \sum_{n=1}^{N} \log(q(C_n)) - \beta \sum_{n \in D} (n - d(n))^{-1} - \gamma F(\mathbf{R}) - \delta T$$