# Inverse translation

## Sequences

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

## Genetic code

Let $G(a) = \{x_1 y_1 z_1, x_2 y_2 z_2 \ldots\}$ 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 $\alpha$ 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 $\beta$ be a penalty for duplicated codons.

## RNA structure

Let $F(\mathbf{X})$ denote the free energy of folding (in kCal/mol) for sequence $\mathbf{X}$, at room temperature, using version 2.2 of the ViennaRNA RNAfold package. (Note that $F < 0$ if the sequence has fold structure.)
Let $\gamma$ be a penalty for RNA structure.

## Runtime

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

## Scoring function

A valid output is scored using the scoring function

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

with defaults $\alpha = \beta = \gamma = 1$ and $\delta = 10$.