Supplementary Information:

The Synthesis Success Calculator: Predicting the Rapid Synthesis of DNA Fragments with Machine Learning

Sean M. Halper[1], Ayaan Hossain[2], Howard Salis[1,2,3,4]

[1]Department of Chemical Engineering, [2]Bioinformatics and Genomics, [3]Department of Biological Engineering, [4]Department of Biomedical Engineering, Pennsylvania State University, University Park, PA 16802. Correspondence should be addressed to H.M.S. (salis@psu.edu).

Supplementary Table 1: Sequence determinants used for initial model training and feature reduction. Rules were derived from synthesis guidelines from multiple commercial service providers as well as unique metrics developed in this work.

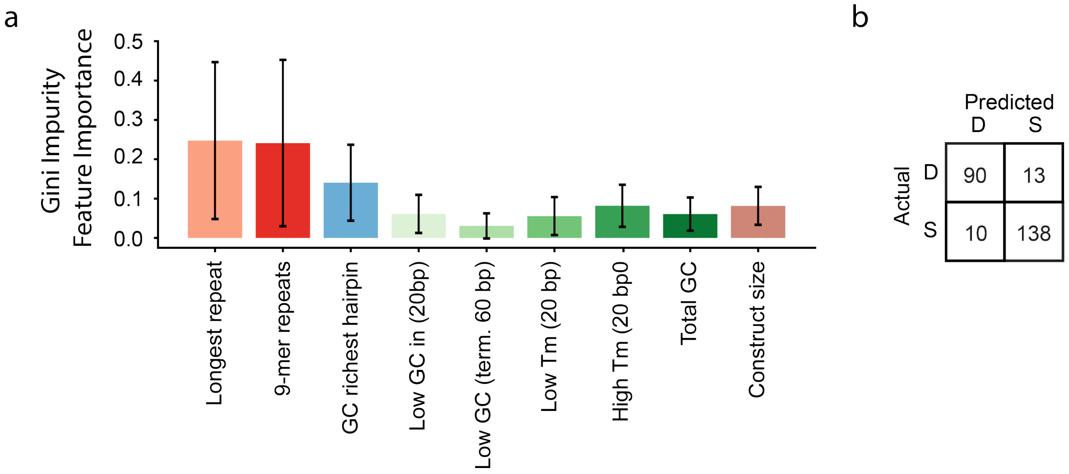
Feature Type	Feature	Description
Repeats	Most frequent repeat count	Number of copies of most abundant repeat in construct
	Scaled 9-mer repeat metric	Scaled count of repeated 9-mers divided by length of construct
	High repeat density (70 bp)	Count of 70 bp windows where 90% of nucleotides participate in a repeat pair
	High repeat density (500 bp)	Count of 500 bp windows where 60% of nucleotides participate in a repeat pair
	Repeats in proportion to sequence length	Count of repeats where 40% of total sequence is a specific repeat
	Total repeat density	Flag if 69% of nucleotides in the total sequence are participating in any repeat pair
	Longest Repeat	Length of longest maximal repeat
	Repeats of length <10 bp	Count of maximal repeats less than or equal to 10 bp
	Repeats of length 10-15 bp	Count of maximal repeats between 11 and 15 bp
	Repeats of length 15-20 bp	Count of maximal repeats between 16 and 20 bp
	Repeats of length 20-25 bp	Count of maximal repeats between 21 and 25 bp
	Repeats of length 25-40 bp	Count of maximal repeats between 26 and 40 bp
	Repeats of length >40 bp	Count of maximal repeats greater than 40 bp
	Tandem repeats >5 bp	Count of repeats 5 or greater separated by 5 or fewer bp
	Terminal repeats	Count of repeats longer than 10 bp in the 5' or 3' 60 bp
Hairpins	Hairpins of length >20 bp	Count of hairpins with a stem length of 20 bp or greater
	Longest hairpin stem	Length of longest stem of predicted hairpins in construct
	Palindromes	Count of direct palindromes in construct
	GC richest hairpin	Highest GC content of hairpins with high GCs
	Strong hairpins	Count of hairpins with GC content above 80%
	Terminal hairpins	Count of hairpins found in the 5' or 3' 60 bp
	Large secondary structures	Secondary structures that sequester 17 contiguous bp of DNA within 100 bp
GC Content	High GC (100 bp)	Count of 100 bp windows with GC >70%
	Low GC (100 bp)	Count of 100 bp windows with GC <30%
	High GC (20 bp)	Count of 20 bp windows with GC >80%
	Low GC (20 bp)	Count of 20 bp windows with GC <20%
	High GC (terminal 60 bp)	Count of 20 bp windows within 5' or 3' 60 bp wher e GC >70%
	Low GC (terminal 60 bp)	Count of 20 bp windows within 5' or 3' 60 bp wher e GC <30%
	High Tm (20bp)	Count of 20 bp windows with Tm ≥ 70˚C
	Low Tm (20bp)	Count of 20 bp windows with Tm ≤ 40˚C
	GC changes (100 bp)	Count of 100 bp windows where the GC content of an pair of 20bp subwindows changes by ≥50%
	Tm changes (100 bp)	Count of 100 bp windows where the Tm of any pair of 20bp subwindows changes by ≥ 30
	Total GC content	Total GC content of the construct
Misc	Construct size	Length in bp of the construct of interest
	Polynucleotide runs	Count of poly N runs
	Motif runs	Count of poly NN or NNN runs
	G quadruplexes	Count of tandem repeated poly Gs that might result in a g quadruplex
	i-motifs	Count of tandem repeated poly Cs that might result in an i -motif

Supplementary Table 2: Design rules used to generate DNA fragment sequences that can not be readily synthesized (negative controls).

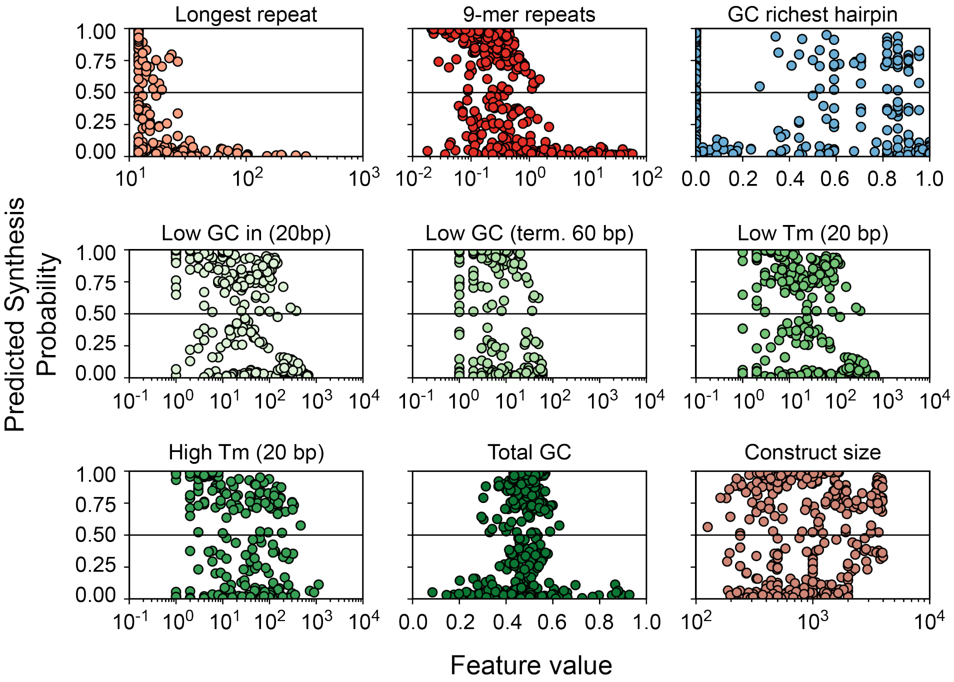
Feature type	Rule	Definition	values
Repeats	Single repeats	N copies of a repeat of length L	5≤N≤10, 14≤L≤25
	Multiple repeats	M unique repeat sets (N copies of a repeat of length L)	3≤M≤5, 5≤N≤10, 14≤L≤25
	Long repeats	N copies of a repeat of length L	3≤N≤5, 30≤L≤80
	Clustered repeats 1	N copies of a repeat of length L within D bp of each other	2≤N≤4, 20≤L≤40, D=160
	Clustered repeats 2	M unique clustered repeats (N copies of a repeat of length L within D bp of each other)	3≤M≤5, 3≤N≤6, 14≤L≤25, D=125
	Tandem repeats	M instances of N repeats of length L adjacent to each other in the D 5' or 3' terminal nucleotides	1≤M≤3, 3≤N≤5, 15≤L≤8, D=40
Hairpins	Long hairpins	N hairpins with stem length S and loop length L	2≤N≤5, 20≤S≤30, 4≤L≤10
	Complex hairpins	N hairpins with stem length S and loop length L	6≤N≤10, 10≤S≤20, 4≤L≤70
	Palindromes	N hairpins with stem length S and loop length L	3≤N≤6, 10≤S≤20, L=0
	Strong hairpins	N hairpins with stem length S and loop length L, GC content of stems G	3≤N≤6, 10≤S≤20, 4≤L≤10, G>0.75
	Terminal hairpins	N hairpins with stem length S and loop length L in the D 5' or 3' terminal nucleotides	2≤N≤4, 12≤S≤20, 4≤L≤10, D=60
	Strong long hairpins	N hairpins with stem length S and loop length L, GC content of stems G	2≤N≤5, 20≤S≤30, 4≤L≤10, G>0.75
	Strong complex hairpins	N hairpins with stem length S and loop length L, GC content of stems G	6≤N≤10, 10≤S≤20, 4≤L≤70, G>0.6
GC content	Total GC high	N regions of length L have a GC content of G	20≤N≤50, 50≤L≤100, G>0.7
	Total GC low	N regions of length L have a GC content of G	20≤N≤50, 50≤L≤100, G<0.3
	100 bp GC high	N regions of length L have a GC content of G	2≤N≤5, 90≤L≤200, G>0.7
	100 bp GC low	N regions of length L have a GC content of G	2≤N≤5, 90≤L≤200, G<0.3
	20 bp GC high	N regions of length L have a GC content of G	4≤N≤10, 20≤L≤40, G>0.8
	20 bp GC low	N regions of length L have a GC content of G	4≤N≤10, 20≤L≤40, G<0.2
	Terminal GC high	N regions of length L have a GC content of G in the D 5' or 3' terminal nucleotides	1≤N≤3, 30≤L≤40, G>0.8, D=45
	Terminal GC low	N regions of length L have a GC content of G in the D 5' or 3' terminal nucleotides	1≤N≤3, 20≤L≤40, G<0.2, D=45
	Terminal GC split	N regions of length L have a GC content of G in the D 5' terminal nucleotides and a GC content of C in the D 3' terminal nucleotides	N=2, 20≤L≤40, G < 0.2, C >.8, D=45
	dGC	N regions of length L in the construct have subregions of length D with GC contents of G and C, respectively	2≤N≤6, 60≤L≤120, 30≤D≤60, G<0.2, C>.8
Other	G quadruplexes	N instances of D adjacent "GGGNNN" motifs	2≤N≤6, 4≤D≤10
	i motifs	N instances of D adjacent "CCCNNN" motifs	2≤N≤6, 4≤D≤10
	Mononucleotide runs	N mononucleotide regions of length L	2≤N≤6, 13≤L≤25
	Dinucleotide runs	N regions composed of D adjacent "NN" motifs	2≤N≤6, 7≤D≤215
	Trinucleotide runs	N regions composed of T adjacent "NNN" motifs	2≤N≤6, 5≤T≤10

Supplementary Table 3: Hyperparameters used for each random forest used by the Synthesis Success Calculator.

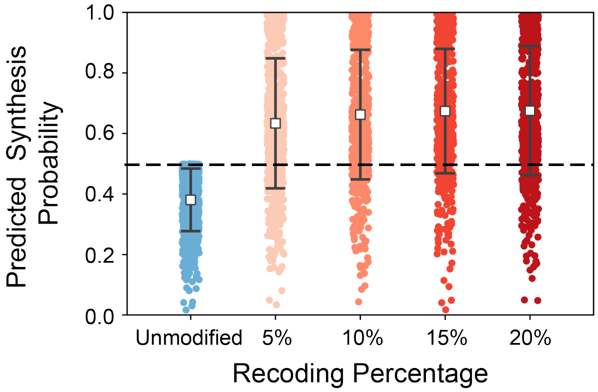
Round	Max_features	Class_weight	N_estimators	Min_samples_split	Min_samples_leaf
Feature reduction	'auto'	None	200	4	4
Round 1: RandomSearchCV 100	'auto'	None	600	8	4
Round 2: GridSearchCV_50	'auto'	None	1200	4	3
Round 3: GridSearchCV 100	'auto'	None	1512	5	2



Supplementary Figure 1: Classifier features and performance, prior to oversampling. (a) Feature importances based on 575 training datapoints. Error bars represent standard deviation of importances for trees in the forest (n=1512) (b) Predicted and actual synthesis outcomes across 251 DNA fragment sequences in the unseen test set. S: Synthesis success. D: Synthesis failure.



Supplementary Figure 2: The predicted probabilities of synthesis success versus feature values across the 595 DNA fragment sequences in the training set and across the reduced feature set.



Supplementary Figure 3: Improvements to predicted synthesis outcomes for 101 *E. coli* proteins after targeted recoding of their protein coding sequences by 5, 10, 15, or 20% amounts. Squares and error bars represent the mean and standard deviation of the predicted synthesis probabilities across 101 proteins (n = 101).