Supplementary Material

Pareto optimization of combinatorial mutagenesis libraries

Authors: Deeptak Verma[1], Gevorg Grigoryan[1,2] and Chris Bailey-Kellogg[1,*]

Affiliations:
Departments of [1]Computer Science and [2]Biological Sciences
Dartmouth College, Hanover, New Hampshire, United States of America

[*]To whom correspondence should be addressed: cbk@cs.dartmouth.edu; 6211 Sudikoff Laboratory, Hanover, NH 03755

Fig. S1: Pareto optimal frontiers for GFP libraries. Pareto frontiers at mutational loads of 10 (red), 15 (green), and 20 (blue), using (a) specific point mutations and (b) degenerate oligos. The insets zoom in on the regions where the slope starts ramping up.

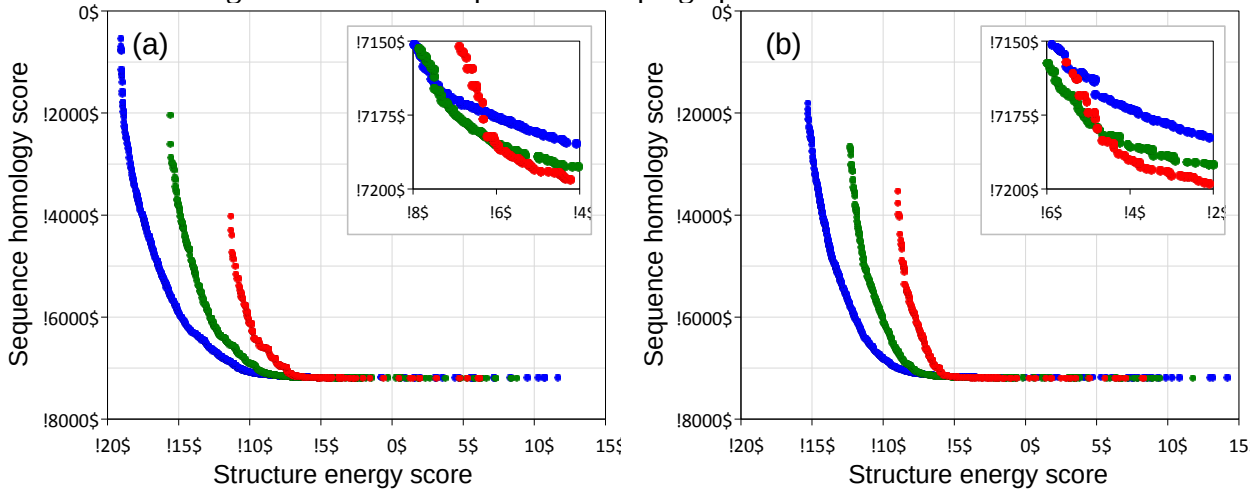
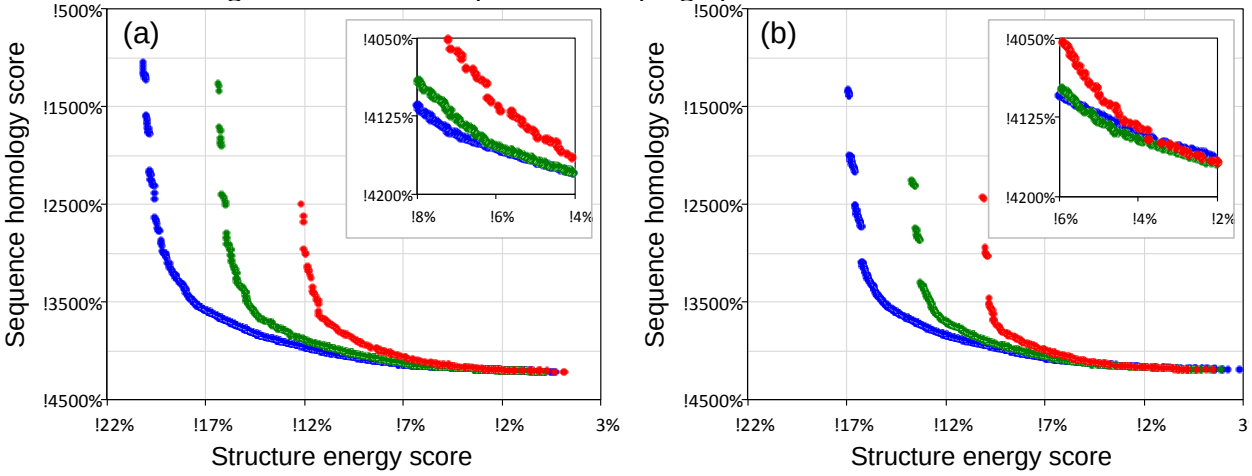


Fig. S2: Pareto optimal frontiers for P450 libraries. Pareto frontiers at mutational loads of 10 (red), 15 (green), and 20 (blue), using (a) specific point mutations and (b) degenerate oligos. The insets zoom in on the regions where the slope starts ramping up.



Data preprocessing: The sequence potential and sequence-based allowed mutations for each of the proteins were derived from a set of homologs (Table S1), filtered to eliminate gappy members of a multiple sequence alignment (at most 25% gaps) and to be sufficiently similar to the wild type (at least 35% identity), and sub-selected to be sufficiently different from each other (at most 95% identity). A background distribution [1] provided threshold for considering an amino acid as a mutational choice only if its MSA frequency exceeded the background frequency for the amino acid type.

The homology-based allowed mutations were augmented with structure-based allowed mutations whose secondary structure Chou-Fasman propensities [2] were similar enough (propensity cutoff of 1.50 or more) to those of the corresponding wild-type residues in the target structure. Mutations that could introduce major structural changes, such as proline and cysteine, were excluded from the list.

Structure potentials were derived from the PDB structures listed in Table S1. A training set of randomly mutated structures was produced with Rosetta [3], and the energies were used for fitting a Cluster Expansion (CE) [4] based potential that could be computed from the amino acid sequence. The model accuracy was verified on a randomly chosen unique test set of structures; correlations are listed in Table S1. The complete process of CE training and modeling for each target requires less than 24 hours on a compute cluster.

Table S1: Preprocessing data for extracting sequence and structure potentials.

Target protein	Organism	UniProt ID	Number of sequence homologs	Filtered sequence representatives	Structure (PDB)	Random structures for CE training	CE Predicted energy correlation
GFP	*Aequorea victoria*	GFP_AEQVI	243 (from Pfam PF01353)	44	1GFL	24000	0.80
P450	*Bacillus Subtilis*	CYPC_BACSU	238 (from BLAST)	160	2ZQJ	36000	0.70
β-lactamase	*Escheridia coli*	BLAT_ECOLX	148 (from BLAST)	42	1BT5	36000	0.75

References:

[1] McCaldon P, Argos P. Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. Proteins. 1988 January;4(2):99–122.

[2] Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974 January 15;13(2):222–45.

[3] Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein Structure Prediction Using Rosetta. Methods in Enzymology. 2004;383:66–93.

[4] Grigoryan G, Reinke AW, Keating AE. Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature. 2009 April 16;458(7240):859–64.