PROTEIN DESIGN

An evolution-based model for designing chorismate mutase enzymes

William P. Russ[1]*, Matteo Figliuzzi[2]*, Christian Stocker[3], Pierre Barrat-Charlaix[2,4], Michael Socolich[5], Peter Kast[3], Donald Hilvert[3], Remi Monasson[6], Simona Cocco[6], Martin Weigt[2]†, Rama Ranganathan[5]†

The rational design of enzymes is an important goal for both fundamental and practical reasons. Here, we describe a process to learn the constraints for specifying proteins purely from evolutionary sequence data, design and build libraries of synthetic genes, and test them for activity in vivo using a quantitative complementation assay. For chorismate mutase, a key enzyme in the biosynthesis of aromatic amino acids, we demonstrate the design of natural-like catalytic function with substantial sequence diversity. Further optimization focuses the generative model toward function in a specific genomic context. The data show that sequence-based statistical models suffice to specify proteins and provide access to an enormous space of functional sequences. This result provides a foundation for a general process for evolution-based design of artificial proteins.

Approaches for protein design typically begin with atomic structures and physical models for forces between atoms, but the marked expansion of protein sequence databases and the growth of new computational methods (*1*–*4*) have opened new strategies to this problem. Statistical analyses of homologs comprising a protein family have recently enabled successful prediction of protein structure (*5*–*8*), protein-protein interactions (*9*–*13*), and mutational effects (*14*–*18*) and, for a family of small protein interaction modules, have led to the successful design of artificial amino acid sequences that fold and function in a manner similar to their natural counterparts (*19*–*21*). These findings indicate that sequence-based statistical models may represent a general approach for a purely data-driven strategy for the design of complex proteins that can work in vivo, in specific organismal contexts. A first key step is to demonstrate the sufficiency of these models for specifying functional proteins.

An approach for evolution-inspired protein design is shown in Fig. 1, A to C, and is based on direct coupling analysis (DCA), a method originally conceived to predict contacts between amino acids in protein three-dimensional (3D) structures (*1*). The starting point is a large and diverse multiple sequence alignment (MSA) of a protein family, from which we estimate the observed frequencies (f_i^a) and pairwise co-occurrences (f_{ij}^{ab}) of all amino acids (a, b) at positions (i, j), the first- and second-order statistics (Fig. 1A). From these quantities, we infer a model comprising a set of intrinsic amino acid propensities [fields $h_i(a)$] and a set of pairwise interactions [couplings $J_{ij}(a, b)$] that optimally account for the observed statistics (Fig. 1A). This model defines a probability P for each amino acid sequence $(a, \cdots, a_L)$ of length L:

$$P(a_1, \cdots, a_L) \sim exp[-H(a_1, \cdots, a_L)] \quad (1)$$

with the Hamiltonian $H(a_1, \cdots, a_L) = -\sum_i h_i(a_i) - \sum_{i<j} J_{ij}(a_i, a_j)$ representing a statistical energy (which provides a quantitative log-likelihood score for each sequence; see materials and methods). Lower energies are associated with higher probability. Monte Carlo (MC) sampling from the model allows for generating nonnatural sequence repertoires (Fig. 1B), which can then be screened for desired functional activities (Fig. 1C). If positional conservation and pairwise correlation suffice in general to capture the information content of protein sequences and if the model inference is sufficiently accurate, then the artificial sequences should recapitulate the functional diversity and properties of natural proteins.

To test this process, we chose the AroQ family of chorismate mutases (CMs), a classic model for understanding principles of catalysis and enzyme design (*22*–*24*). These enzymes occur in bacteria, archaea, fungi, and plants and operate at the central branch point of the shikimate pathway leading to the biosynthesis of tyrosine (Tyr) and phenylalanine (Phe) (Fig. 1D). CMs catalyze the conversion of the intermediary metabolite chorismate to prephenate through a Claisen rearrangement, leading to more than a million-fold acceleration of the reaction rate (*25*), and are necessary for growth of bacteria in minimal media. For example, *Escherichia coli* strains lacking a CM are auxotrophic for Tyr and Phe, with both the degree of supplementation of these amino acids and the expression level of a reintroduced CM gene quantitatively determining the growth rate (*23*). Structurally, AroQ$_\alpha$ subfamily members exemplified by EcCM, the CM domain of the CM-prephenate dehydratase from *E. coli*, form a domain-swapped dimer of relatively small protomers (~100 amino acids) (Fig. 1E) (*26*, *27*). Their size, essentiality for bacterial growth, and the existence of good biochemical assays make AroQ CMs an excellent target for testing the power of statistical models inferred from MSAs.

We used DCA to make a statistical model (Eq. 1) for an alignment of 1259 natural AroQ protein domains that broadly encompasses the diversity of bacterial, archaeal, fungal, and plant lineages. Deducing the exact parameters (h_i, J_{ij}) from the observed statistics in the MSA (f_i, f_{ij}) for any protein is computationally intractable, but a number of approximation algorithms are available (*1*). Here, we used bmDCA, a computationally quite expensive but highly accurate method based on Boltzmann machine learning (*28*). For example, sequences generated by MC sampling from the model reproduced the empirical first- and second-order statistics of natural sequences used for fitting (Fig. 2, A and B). We also found that the model recapitulates higher-order statistical features in the MSA that were never used in inferring the model (see materials and methods). These features include three-way residue correlations (Fig. 2C) and the inhomogeneously clustered phylogenetic organization of the protein family in sequence space (Fig. 2D). Our findings suggest that the statistical model goes beyond just being a good fit to the first- and second-order statistics to capture the essential rules governing the divergence of natural CM sequences through evolution. By contrast, a simpler profile model that retains only the intrinsic propensities of amino acids at sites [$h_i(a)$, fitted to reproduce frequencies of amino acids f_i^a] (fig. S1A) but leaves out pairwise couplings fails to reproduce even the second-order statistics of the MSA (fig. S1B) and does not account well for the pattern of sequence divergence in the natural CM proteins (fig. S1D).

These findings raise the possibility that bmDCA may be a generative model, meaning that natural sequences and sequences sampled from the probability distribution $P(\boldsymbol{a})$ are, despite considerable divergence, equivalent. To test this, we developed a high-throughput, quantitative in vivo complementation assay to monitor CM activity in *E. coli* that is suitable for studying large numbers of natural and designed CMs in a single experiment. Briefly, libraries of CM variants (natural and/or artificial, see below) were made using a custom de novo gene synthesis protocol that is capable

[1]University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Sorbonne Université, CNRS, Institut de Biologie Paris Seine, Laboratoire de Biologie Computationnelle and Quantitative, Paris, France. [3]Laboratory of Organic Chemistry, ETH Zurich, Switzerland. [4]Biozentrum, University of Basel, Basel, Switzerland. [5]Center for Physics of Evolving Systems, Biochemistry and Molecular Biology and the Pritzker School for Molecular Engineering, University of Chicago, Chicago, IL, USA. [6]Laboratoire de Physique de l'Ecole Normale Supérieure, PSL and CNRS, Paris, France.
*These authors contributed equally to this work.
†Corresponding author. Email: ranganathanr@uchicago.edu (R.R.); martin.weigt@upmc.fr (M.W.)

Corrected 3 December 2020. See full text.

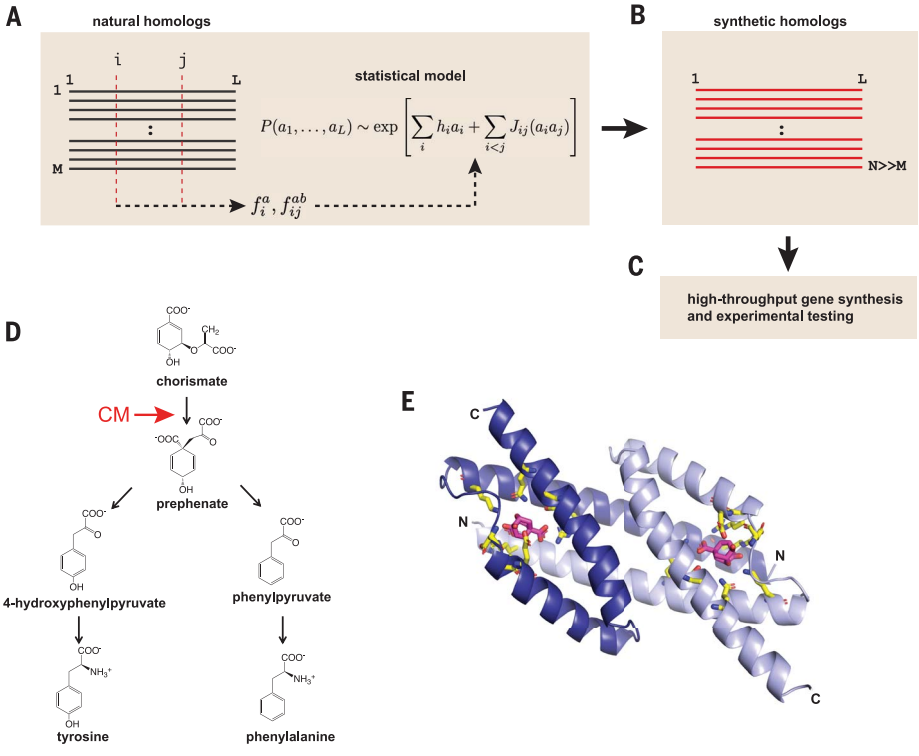

Fig. 1. Evolutionary data-driven protein engineering. (**A**) MSA of *M* natural homologs provides empirical first- and second-order statistics of amino acids (f_i^a, f_{ij}^{ab}), which are used to infer a statistical model with the bmDCA method. The probability of sequence $\boldsymbol{a} = (a_1, \ldots, a_L)$ is an exponential function of a Hamiltonian, or statistical energy, parameterized by intrinsic fields $h_i(a)$ and couplings $J_{ij}(a, b)$ acting on amino acids. (**B** and **C**) The model is used to generate $N \gg M$ artificial sequences that can be tested in a high-throughput assay for desired functions. (**D**) CM is an enzyme occurring at the central branch point in the shikimate pathway that leads to the synthesis of Tyr and Phe. (**E**) Members of the AroQ$_\alpha$ and AroQ$_\delta$ families of CMs fold into a domain-swapped dimer (PDB ID 1ECM). Active site residues are shown with yellow stick bonds and arise from both subunits (dark and light blue). A bound substrate analog is shown in magenta.

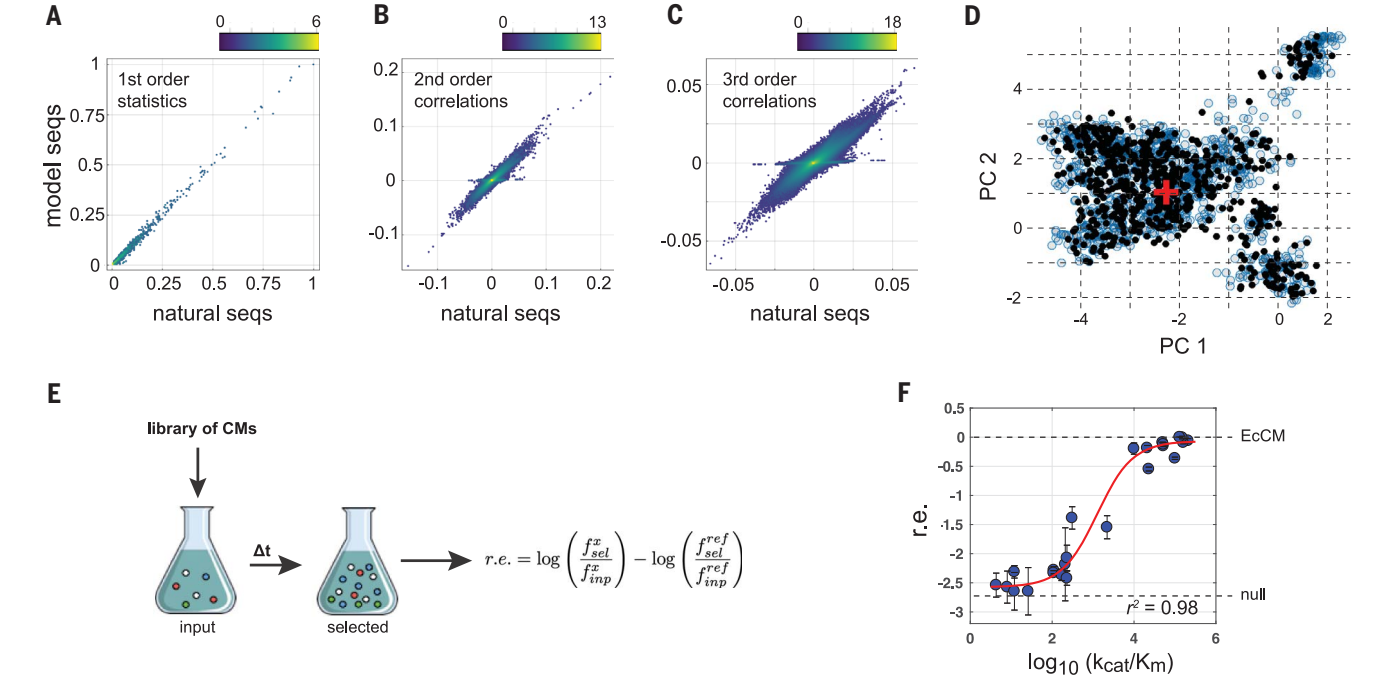

Fig. 2. Design and testing of artificial CM sequences. (**A** to **C**) 2D histograms showing the relationship of first- (A), second- (B), and third-order (C) statistics of natural and bmDCA-designed sequences. The color scale indicates the number of counts per bin. (**D**) The top two principal components of the pairwise sequence distance matrix of natural homologs (blue circles) overlaid with a projection of artificial CM sequences (black circles); the position of EcCM from *E. coli* is marked with a red plus sign. Artificial sequences both recapitulate data used for fitting (A and B) and also account for statistical features of natural data not used for fitting (C and D). (**E**) Workflow for functional characterization of CM activity. CM-deficient *E. coli* cells carrying libraries of variants were grown under selective conditions in minimal medium, after which we performed deep sequencing of input and selected populations and calculation of the r.e. of each variant. (**F**) The relationship of r.e. to catalytic power [$\log_{10}(k_{cat}/K_m)$] for a number of CM variants yields a "standard curve." The assay shows a hyperbolic relationship over the range from complete lack of CM activity to wild-type EcCM activity.

Corrected 3 December 2020. See full text.

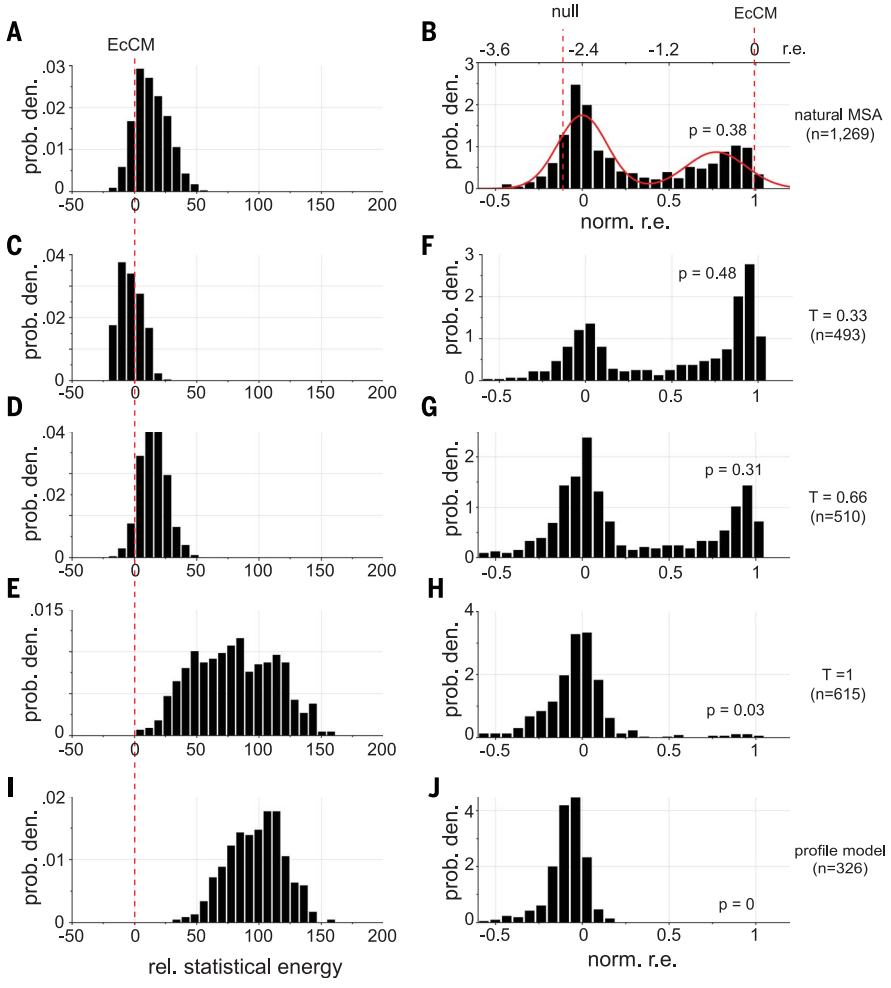

Fig. 3. Functional analysis of natural and artificial CMs. (**A**) Distribution of bmDCA statistical energies for 1130 tested natural AroQ homologs, relative to the value for EcCM. The data show a unimodal distribution centered close to EcCM. (**B**) The distribution of functional complementation by natural AroQ sequences is bimodal, with ~38% of sequences in one mode near that of EcCM and the rest in another mode close to the r.e. of the null allele. The bimodality is used to normalize the raw r.e. scores between zero and the mean of the near-null mode for all libraries in panels B, F to H, and J. (**C** to **E**) Distributions of statistical energies for artificial sequences. (**F** to **H**) Distributions of corresponding r.e. values sampled at $T \in \{0.33, 0.66, 1\}$, respectively. (**I** and **J**) Distributions of statistical energy and r.e. for artificial sequences retaining the intrinsic propensities of amino acids at positions but leaving out all correlations. Taken together, the data show that bmDCA is a generative model.

of fast and relatively inexpensive assembly of novel DNA sequences at large scale (*29*) (see materials and methods). For example, we made gene libraries comprising nearly every natural CM homolog in the MSA (1130 out of 1259 total) and more than 1900 artificial variants by exploring various design parameters of the bmDCA model (see materials and methods). These libraries were expressed in a CM-deficient bacterial host strain (KA12/pKIMP-UAUC) (*23*), and all transformants were grown together as a single population in selective medium lacking Phe and Tyr to select for those variants exhibiting CM activity (Fig. 2E). Deep sequencing of the population before and after selection allowed us to compute the log frequency of each allele relative to that of wild-type EcCM. This quantity is called the relative enrichment (r.e.), which under particular conditions of gene induction, growth time, and temperature quantitatively and reproducibly reports the catalytic CM activity (Fig. 2F and fig. S2). This "select-seq" assay is monotonic over a broad range of catalytic power and serves as an effective tool to rigorously compare large numbers of natural and artifical variants for functional activity in vivo, in a single internally controlled experiment.

The first study examined the performance of the natural CM homologs in the select-seq assay as a positive control for bmDCA-designed sequences. Natural sequences showed a unimodal distribution of bmDCA statistical energies centered close to the value of EcCM (defined as zero, Fig. 3A), but it was not obvious how they would function in the particular *E. coli* host strain and experimental conditions used in our assay. For example, members of the CM family may vary in unknown ways with regard to activity in any particular environment, and the MSA includes some fraction of paralogous enzymes that carry out a related but distinct chemical reaction (*30*, *31*). The select-seq assay showed that the 1130 natural CM homologs exhibit a bimodal distribution of complementation in the assay, with one mode comprising ~38% of the sequences centered close to the level of wild-type EcCM and the remainder comprising a mode centered close to the level of the null allele (Fig. 3B). A green fluorescent protein–tagged version of the library suggests that the bimodality of complementation is not obviously related to differences in expression levels compared to the *E. coli* variant (fig. S3); instead, the bimodality presumably originates from nonlinearities linking sequence to growth rate in the host strain and from the inclusion of some functionally distinct paralogous sequences. For the purpose of this study, the bimodality allows normalization of r.e. scores by the two modes and by Gaussian mixture modeling to meaningfully group sequences into those that are functionally either like wild-type EcCM (norm. r.e. > 0.42) or like the null allele in our assay (Fig. 3B). The standard curve shows that this quantity is a stringent test of high CM activity (Fig. 2F).

To evaluate the generative potential of the bmDCA model, we used MC sampling to randomly draw sequences from the model that span a wide range of statistical energies relative to the natural MSA (Fig. 3, C to E), with the hypothesis that sequences with low energy (i.e., high probability) may be functional CMs. To sample sequences with low energy, we introduced a formal computational "temperature" of $T \leq 1$ in our model:

$$P_T(a, \cdots, a_L) \sim exp[-H(a_1, \cdots a_L)/T]$$

which, in exact analogy to temperature in statistical physics, serves to decrease the mean energy when set to values below unity. For example, sampling at $T \in \{0.33, 0.66\}$ produced sequences with statistical energies that closely reflected the natural distribution (Fig. 3D) or reached even lower values (Fig. 3C). By contrast, sequences sampled at $T = 1$ showed a broad distribution of statistical energies that deviated significantly from the natural distribution (Fig. 3E) toward higher energies. This deviation is, among other factors, due to statistical adjustments [regularization (see materials and methods)] used during model inference for compensating for the limited sampling of sequences in the input MSA.

Table 1. Biochemical properties of natural and artificial CM enzymes. Construct numbers correspond to the numbering presented in tables S1 to S3, which provide additional information about these CM proteins.

Construct	Closest natural	ID to EcCM	Top ID	k_{cat} [s^{-1}]	K_m [μM]	k_{cat}/K_m [$M^{-1}s^{-1}$]
			Natural			
1	*Escherichia coli*	1.00	1.00	64	390	1.6×10^5
	Methanococcus jannaschii	0.30	1.00	5.7	41	1.4×10^5
125	*Nitratiruptor* sp. SB155-2	0.32	1.00	110 ± 10	600 ± 120	$2.0 \pm 0.6 \times 10^5$
230	*Geobacter uraniireducens*	0.30	1.00	45 ± 4	37 ± 7	$1.2 \pm 0.1 \times 10^6$
449	*Acetoanaerobium sticklandii*	0.18	1.00	19 ± 1	190 ± 30	$1.0 \pm 0.2 \times 10^5$
528	*Ilyobacter polytropus*	0.24	1.00	19 ± 2	38 ± 5	$5.1 \pm 1.2 \times 10^5$
553	*Collinsella intestinalis*	0.23	1.00	7.6 ± 3.2	110 ± 20	$6.8 \pm 1.8 \times 10^4$
			Designed			
1950	*Sulfurisphaera tokodaii*	0.25	0.67	8.7 ± 1.3	120 ± 20	$7.7 \pm 2.1 \times 10^4$
2021	*Desulfovibrio* sp.	0.27	0.68	2.8 ± 0.1	330 ± 10	$8.5 \pm 0.1 \times 10^3$
2026	*Methanocella arvoryzae*	0.25	0.69	30 ± 10	220 ± 30	$1.4 \pm 0.6 \times 10^5$
2051	*Methanococcus vannielii*	0.22	0.66	3.0 ± 0.1	230 ± 3	$1.3 \pm 0.0 \times 10^4$
2064	*Stroptococcus sanguinis*	0.22	0.69	21 ± 13	75 ± 25	$2.7 \pm 0.8 \times 10^5$

We made and tested libraries of 493 to 615 artificial sequences sampled at $T \in \{0.33, 0.66, 1.0\}$ from bmDCA models inferred at two regularization strengths (Fig. 3, F to H). The data showed that, overall, these sequences also displayed a bimodal distribution of complementation, with many complementing function near the level of the wild-type EcCM sequence. Consistent with our hypothesis, the probability of complementation is well-predicted by the bmDCA statistical energy, with low-energy sequences drawn from $T \in \{0.33, 0.66\}$ essentially recapitulating or even somewhat exceeding the performance of natural sequences (Fig. 3, F and G). By contrast, sequences drawn from $T = 1$ showed poor performance, consistent with deviation in bmDCA statistical energies (Fig. 3H). Overall, 481 artificial sequences out of 1618 total (~30%, norm r.e. > 0.42) rescued growth in our assay, comprising a range of top-hit identities to any natural CM of 42 to 92% (table S1 and fig. S4, A and B). These included 46 sequences with <65% identity to proteins in the MSA, corresponding to at least 33 mutations away from the closest natural counterpart. Sequence divergence from EcCM ranged from 14 to 82% (fig. S4, C and D). A representation of the positions in the EcCM protein that contribute most to the bmDCA statistical energy highlights residues distributed both within the active site and extending through the AroQ tertiary structure to include the dimer interface (blue spheres, Fig. 4F).

Are the abilities of artificial CM sequences to rescue just a function of their sequence distance from their natural counterparts? To test this, we made 326 sequences with the same distribution of top sequence identities as bmDCA-designed sequences but preserving only the first-order statistics (position-specific conservation) and leaving out correlations. These sequences expectedly showed high bmDCA energies and displayed no complementation at all (Fig. 3, I and J). Thus, enzyme function is not simply about the magnitude of sequence variation and not even about conservation of sites taken independently; instead, it fundamentally depends on the pattern of correlations imposed by the couplings in J_{ij} (see Eq. 1).

We selected five natural and five artificial CMs that complement growth in our assay for in-depth biochemical studies. The natural sequences occur in organisms representing a broad range of phylogenetic groups and diverse environments, and the artificial sequences were chosen to sample regions that reflect this diversity (Fig. 4, B and C). The selected CM genes were expressed in an *E. coli* strain that is deficient for endogenous CM to eliminate any possibility of contamination with the wild-type enzyme, and catalytic parameters of the purified proteins were determined using a spectrophotometric assay following the consumption of the substrate chorismate (*23*). The data showed that all 10 CM genes were expressed similarly and that the natural CMs displayed catalytic parameters similar to those of the previously characterized *E. coli* (*32*) and *Methancoccus jannaschii* (*33*) enzymes (Table 1). Consistent with their natural-like complementation, the artificial CMs showed catalytic parameters that closely recapitulated those of natural CMs. Thus, we conclude that the bmDCA-designed CMs are bona fide synthetic orthologs of the CM family.

Putting all the data together, we found a steep relationship between bmDCA statistical energy and CM activity (Fig. 4A and fig. S5). Forty-five percent of artificial sequences rescued the CM null phenotype when the statistical energy was below a threshold value set by the distribution of statistical energies observed for natural sequences ($E_{DCA} < 50$) (Fig. 4A), and essentially no sequences (<3%) were functional above this value. Thus, bmDCA infers an effective generative model, capable of designing natural-like enzymatic activity with considerable sequence diversity if statistical energies are within the range of natural homologs. The extent of sequence variation from natural homologs highlights the sparsity of the essential constraints on folding and biochemical function.

The bmDCA model captures the overall statistics of a protein family and does not focus on specific functional activities of individual members of the family. Thus, just like natural CM homologs, most bmDCA-designed sequences do not complement function under the specific conditions of our assay (Fig. 4A). But, might it be possible to improve the generative model to deduce the extra information that makes a protein sequence optimal for a specific phenotype? A bit of insight comes from the study of how sequences that rescue function in our assay occupy the sequence space spanned by natural CM sequences. Natural CMs that complement function in our *E. coli* host strain are distributed in several diverse clusters (Fig. 4B), but functional artificial sequences also follow the same pattern (Fig. 4C). This suggests that information about CM function in the specific context of the *E. coli* assay conditions exists in the statistics of natural sequences and might be learned. If so, knowledge gained in the first experimental trial might be added to formally train a classifier to predict artificial sequences that encode particular protein phenotypes and organismal environments.

Corrected 3 December 2020. See full text.

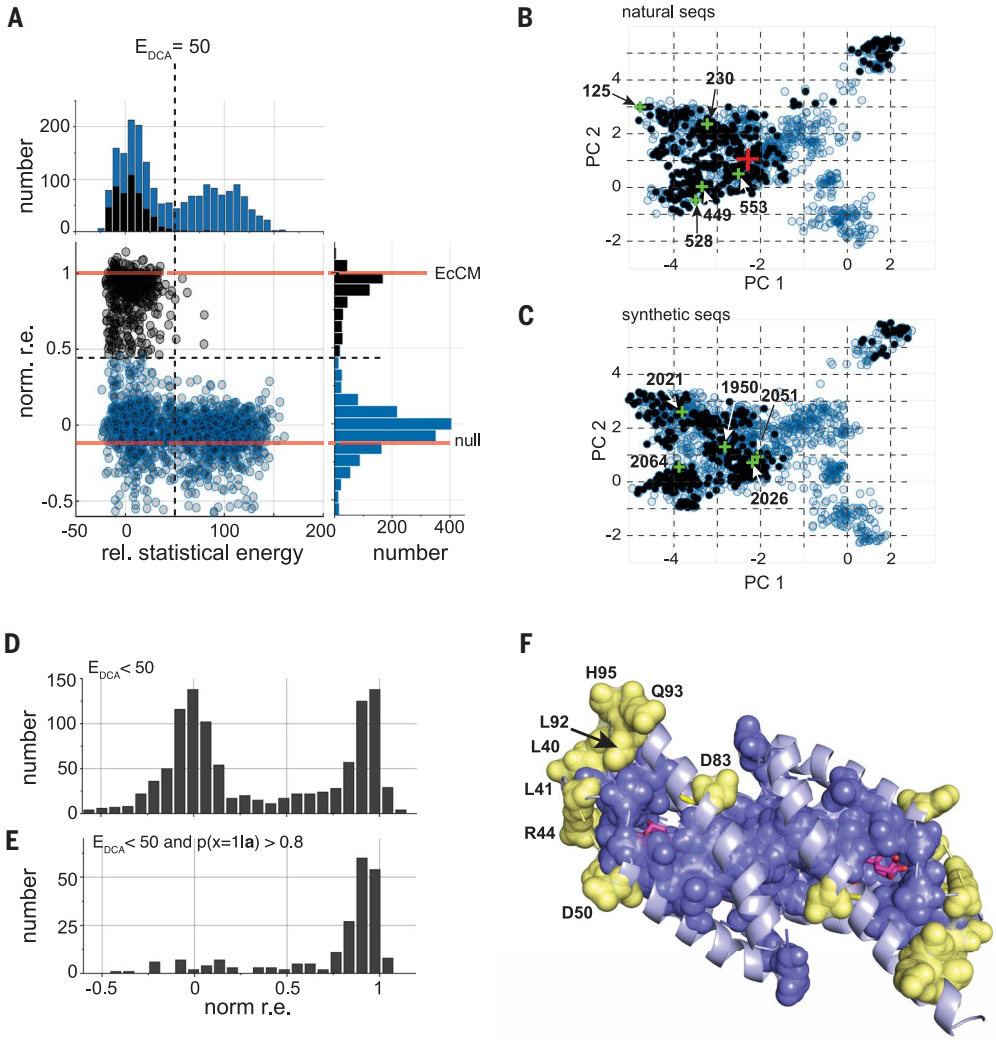

Fig. 4. General and system-specific constraints in CM. (**A**) The overall relationship of bmDCA statistical energies and catalytic function, with functional sequences in black and nonfunctional sequences in blue. The data expose a strong nonlinear relationship, with functional sequences strictly below a sharp threshold ($E_{DCA} < 50$). This value is the limit of statistical energies for natural homologs (Fig. 3A). (**B**) The two top principal components of sequence variation in natural homologs, with sequences complementing the *E. coli* CM auxotroph in black. (**C**) The same as for panel b, but for artificial sequences, showing a similar pattern. Sequences chosen for in-depth biochemical characterization are indicated in panels B and C (see Table 1). (**D** and **E**) The r.e. distributions for all artificial sequences with $E_{DCA} < 50$ (D) or with an additional statistical constraint derived from the pattern of rescue of natural homologs ($P(x = 1|\boldsymbol{a})$; see text for details) (E). The additional constraint identifies sequences functional in *E. coli* in our selection assay. (**F**) Spatial architecture of functional constraints in CM enzymes mapped onto the EcCM structure. Blue spheres show positions constrained in the bmDCA model. Yellow spheres show the extra constraints required for *E. coli*–specific function, highlighting a peripheral shell around active site residues important for CM catalysis.

To test this idea, we annotated the sequences in the natural MSA with a binary value x indicating their ability to function in our assay ($x = 1$ if functional, $x = 0$ if not). Due to its formal similarity to the DCA framework, we used logistic regression on the annotated MSA to develop a model providing a probability for any sequence $\boldsymbol{a} = (a_1, \cdots, a_L)$ to function in the *E. coli* select-seq assay; that is, $P(x = 1|\boldsymbol{a})$. Figure 4, D and E, shows that for low-energy CM-like artificial sequences sampled from the naïve, unsupervised bmDCA model (Fig. 4D), the extra condition $P(x = 1|\boldsymbol{a}) > 0.8$ efficiently predicted the subset that complements in the context of our assay (83%) (Fig. 4E). These results support an iterative design strategy for specific protein phenotypes in which the bmDCA model is updated with each round of selection to optimize desired phenotypes.

What structural principles underlie the general constraints on CM function and the extra constraints for system-specific function? Mapping the positions that contribute most significantly to *E. coli*–specific function of CM sequences showed an arrangement of amino acids peripheral to the active site, within a poorly conserved secondary shell around active site positions (Fig. 4F). Thus, these positions work allosterically or otherwise indirectly to control catalytic activity, a mechanism to provide context-dependent fine-tuning of reaction parameters.

The results described here validate and extend the concept that pairwise amino acid correlations in practically available sequence alignments of protein families suffice to specify protein folding and function (*19*, *20*). The bmDCA model is one approach to capture these correlations, but there is more work to be done to fully understand these models. Currently, the interpretation of DCA is focused on the relatively few highest-magnitude terms in the matrix of couplings (J_{ij}), because these terms identify direct structural contacts between amino acids in protein tertiary structures (*34*). Indeed, the top terms in J_{ij} for CMs do nicely correspond to contacts in the tertiary structure (fig. S6). However, contact terms in J_{ij} alone do not suffice to reproduce either the alignment statistics in the AroQ family (fig. S7) or the functional effects of mutations (*17*). Instead, function in proteins seems to depend also on many weaker, noncontacting terms in J_{ij} that currently have no simple physical interpretation. Similar findings have been made in the case of predicting protein-protein interaction specificity (*35*). The weaker terms in J_{ij} seem to describe the collective evolution of amino acids within the structure, a property

that may be related to patterns elucidated by other approaches to sequence coevolution (*1*, *36*–*40*). A key next goal is to further refine the topology and best representation of sequence correlations that underlie the physics of protein structure and function.

However, even pending these necessary refinements, the data presented here provide the foundations for a general data-driven approach to protein engineering. This approach is similar to directed evolution in that it works without the use of physics-based potentials or atomic structures, but the computational models access a sequence space of functional proteins that is vastly larger than currently understood. Indeed, the results presented here permit a lower-bound estimate of the size of the sequence space consistent with the evolutionary rules for specifying members of a protein family. For example, at $T = 0.66$, we conservatively compute a total space of 10^{25} sequences that could be synthetic homologs of the AroQ family (see materials and methods). Given that ~30% of sequences randomly sampled from this pool rescue CM function under our assay conditions, this amounts to more than 10^{24} sequences that can operate in a specific genomic and experimental context. These numbers are enormous in absolute terms but are infinitesimally unlikely in a sequence space searched without any model (~10^{125}) or with models capturing only first-order constraints (10^{85}; see materials and methods). These considerations suggest that it will be of great interest to use evolution-based statistical models to guide the search for functional proteins with altered or even novel chemical activities.

REFERENCES AND NOTES

1. S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, M. Weigt, *Rep. Prog. Phys.* **81**, 032601 (2018).
2. K. Davidsen *et al.*, *eLife* **8**, e46935 (2019).
3. D. Repecka *et al.*, Expanding functional protein sequence space using generative adversarial networks. bioRxiv 789719 [Preprint]. 4 October 2019. https://doi.org/10.1101/789719.
4. K. A. Reynolds, W. P. Russ, M. Socolich, R. Ranganathan, *Methods Enzymol.* **523**, 213–235 (2013).
5. S. Ovchinnikov *et al.*, *Science* **355**, 294–298 (2017).
6. D. S. Marks *et al.*, *PLOS ONE* **6**, e28766 (2011).
7. D. S. Marks, T. A. Hopf, C. Sander, *Nat. Biotechnol.* **30**, 1072–1080 (2012).
8. F. Morcos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
9. A. F. Bitbol, R. S. Dwyer, L. J. Colwell, N. S. Wingreen, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12180–12185 (2016).
10. Q. Cong, I. Anishchenko, S. Ovchinnikov, D. Baker, *Science* **365**, 185–189 (2019).
11. T. Gueudré, C. Baldassi, M. Zamparo, M. Weigt, A. Pagnani, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12186–12191 (2016).
12. S. Ovchinnikov, H. Kamisetty, D. Baker, *eLife* **3**, e02030 (2014).
13. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
14. R. R. Cheng *et al.*, *Mol. Biol. Evol.* **33**, 3054–3064 (2016).
15. M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, M. Weigt, *Mol. Biol. Evol.* **33**, 268–280 (2016).
16. R. M. Levy, A. Haldane, W. F. Flynn, *Curr. Opin. Struct. Biol.* **43**, 55–62 (2017).
17. V. H. Salinas, R. Ranganathan, *eLife* **7**, e34300 (2018).
18. T. A. Hopf *et al.*, *Nat. Biotechnol.* **35**, 128–135 (2017).
19. W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, R. Ranganathan, *Nature* **437**, 579–583 (2005).
20. M. Socolich *et al.*, *Nature* **437**, 512–518 (2005).
21. P. Tian, J. M. Louis, J. L. Baber, A. Aniana, R. B. Best, *Angew. Chem. Int. Ed.* **57**, 5674–5678 (2018).
22. D. H. Calhoun, C. A. Bonner, W. Gu, G. Xie, R. A. Jensen, *Genome Biol.* **2**, H0030 (2001).
23. P. Kast, M. Asif-Ullah, N. Jiang, D. Hilvert, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5043–5048 (1996).
24. G. MacBeath, P. Kast, D. Hilvert, *Science* **279**, 1958–1961 (1998).
25. P. R. Andrews, G. D. Smith, I. G. Young, *Biochemistry* **12**, 3492–3498 (1973).
26. A. Y. Lee, P. A. Karplus, B. Ganem, J. Clardy, *J. Am. Chem. Soc.* **117**, 3627–3628 (1995).
27. K. Roderer, P. Kast, *Chimia (Aarau)* **63**, 313–317 (2009).
28. M. Figliuzzi, P. Barrat-Charlaix, M. Weigt, *Mol. Biol. Evol.* **35**, 1018–1027 (2018).
29. N. Eroshenko, S. Kosuri, A. H. Marblestone, N. Conway, G. M. Church, *Curr. Protoc. Chem. Biol.* **2012**, ch110190 (2012).
30. C. Gaille, P. Kast, D. Haas, *J. Biol. Chem.* **277**, 21768–21775 (2002).
31. D. E. Künzler, S. Sasso, M. Gamper, D. Hilvert, P. Kast, *J. Biol. Chem.* **280**, 32827–32834 (2005).
32. A. Mandal, D. Hilvert, *J. Am. Chem. Soc.* **125**, 5598–5599 (2003).
33. G. MacBeath, P. Kast, D. Hilvert, *Biochemistry* **37**, 10062–10073 (1998).
34. I. Anishchenko, S. Ovchinnikov, H. Kamisetty, D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 9122–9127 (2017).
35. A. F. Bitbol, *PLOS Comput. Biol.* **14**, e1006401 (2018).
36. K. Shimagaki, M. Weigt, *Phys. Rev. E* **100**, 032128 (2019).
37. J. Tubiana, S. Cocco, R. Monasson, *eLife* **8**, e39397 (2019).
38. L. J. Colwell, M. P. Brenner, A. W. Murray, *PLOS ONE* **9**, e107723 (2014).
39. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, *Cell* **138**, 774–786 (2009).
40. S. W. Lockless, R. Ranganathan, *Science* **286**, 295–299 (1999).
41. P. Barrat. ranganathanlab/bmDCA, Version v0.8.12, Zenodo (2020); https://doi.org/10.5281/zenodo.3825613.

ACKNOWLEDGMENTS

We thank O. Rivoire, C. Nizak, A. Ferguson, C. Feinauer, G. Uguzzoni, A. Pagnani, and members of the Ranganathan lab for discussions, A. George for computational work, and the high-performance computing (BioHPC) and flow cytometry facilities at UT Southwestern Medical Center. **Funding:** This work was supported by NIH grant RO1GM12345 (R.R.), Robert A. Welch Foundation grant I-1366 (R.R.), a Data Science Discovery award from the University of Chicago Center for Data and Computing (R.R.), the Green Center for Systems Biology at UT Southwestern Medical Center (R.R.), the EU H2020 Research and Innovation Programme MSCA-RISE-2016, Grant Agreement 734439 (InferNet) (M.W.), grant number CE30-0021-01 RBMpro from the Agence Nationale de la Recherche (R.M. and S.C.), and Swiss National Science Foundation grants 310030M_182648 (P.K.) and 310030B_176405 (D.H.). **Author contributions:** W.P.R., M.W., R.M., S.C., and R.R. conceived the idea for the project; M.F., P.B.-C., S.C., and M.W. built the computational models; W.P.R., with contributions from P.K. and D.H., developed the in vivo CM assay; W.P.R. built the synthetic gene libraries and carried out the in vivo assays; C.S., with contributions from M.S., P.K., and D.H., obtained the kinetic parameters in Table 1; S.C. and R. M. carried out the sequence entropy calculations; and W.P.R., M.W., and R.R. wrote the paper with contributions from all authors. **Competing interests:** R.R. is a cofounder and scientific advisor of Evozyne and is an inventor on a patent application filed by the University of Chicago on technologies for data-driven molecular engineering. **Data and materials availability:** Materials are available from the authors by request, and codes for bmDCA are available at https://github.com/ranganathanlab/bmDCA and at Zenodo (*41*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6502/440/suppl/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S3
References (*42*–*54*)

View/request a protocol for this paper from *Bio-protocol*.

24 November 2019; accepted 13 May 2020
10.1126/science.aba3304