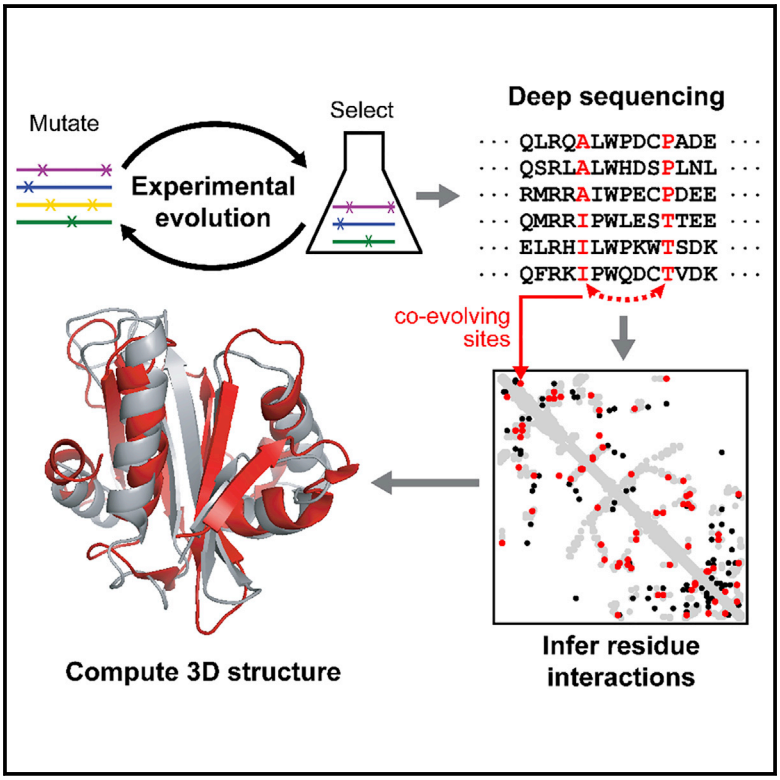
Article

Cell Systems

Protein Structure from Experimental Evolution

Graphical Abstract



Highlights

- Introduces a new experimental method for protein structure determination called 3Dseq
- Experimental evolution generates hundreds of thousands of functional sequences
- Analysis of sequence co-variation patterns yields accurate residue interactions
- Molecular dynamics with interaction constraints produces correct protein folds

Authors

Michael A. Stiffler, Frank J. Poelwijk, Kelly P. Brock, ..., Debora S. Marks, Nicholas P. Gauthier, Chris Sander

Correspondence

3Dseq.research@gmail.com (M.A.S.),
3Dseq.research@gmail.com (F.J.P.),
3Dseq.research@gmail.com (N.P.G.),
3Dseq.research@gmail.com (C.S.)

In Brief

We asked whether experimental evolution, similar to natural evolution, generates enough information about amino acid residue interactions to determine protein structure. We mimicked the natural evolution of two distinct antibiotic resistance proteins using an iterative mutation-selection process. From statistical transformation of co-evolution patterns in the evolved gene sequences we inferred important interactions in each protein, called evolutionary couplings, and from these accurately computed their 3D structures. This work introduces a new technology for 3D structure determination.

Stiffler et al., 2020, Cell Systems *10*, 1–10
January 22, 2020
https://doi.org/10.1016/j.cels.2019.11.008

CellPress

Please cite this article in press as: Stiffler et al., Protein Structure from Experimental Evolution, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.11.008

Cell Systems
Article

Protein Structure from Experimental Evolution

Michael A. Stiffler,[1,2,4,7,*] Frank J. Poelwijk,[1,2,4,7,*] Kelly P. Brock,[3] Richard R. Stein,[1,5] Adam Riesselman,[3] Joan Teyra,[6] Sachdev S. Sidhu,[6] Debora S. Marks,[3,4] Nicholas P. Gauthier,[1,2,4,8,*] and Chris Sander[1,2,4,8,9,*]
[1]cBio Center, Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
[2]Department of Cell Biology, Harvard Medical School, Boston, MA, USA
[3]Department of Systems Biology, Harvard Medical School, Boston, MA, USA
[4]Broad Institute, Cambridge, MA, USA
[5]Harvard School of Public Health, Boston, MA, USA
[6]Donnelly Centre, University of Toronto, Toronto, Canada
[7]These authors contributed equally
[8]Senior author
[9]Lead Contact
*Correspondence: 3Dseq.research@gmail.com (M.A.S.), 3Dseq.research@gmail.com (F.J.P.), 3Dseq.research@gmail.com (N.P.G.), 3Dseq.research@gmail.com (C.S.)
https://doi.org/10.1016/j.cels.2019.11.008

SUMMARY

Natural evolution encodes rich information about the structure and function of biomolecules in the genetic record. Previously, statistical analysis of co-variation patterns in natural protein families has enabled the accurate computation of 3D structures. Here, we explored generating similar information by experimental evolution, starting from a single gene and performing multiple cycles of *in vitro* mutagenesis and functional selection in *Escherichia coli*. We evolved two antibiotic resistance proteins, β-lactamase PSE1 and acetyltransferase AAC6, and obtained hundreds of thousands of diverse functional sequences. Using evolutionary coupling analysis, we inferred residue interaction constraints that were in agreement with contacts in known 3D structures, confirming genetic encoding of structural constraints in the selected sequences. Computational protein folding with interaction constraints then yielded 3D structures with the same fold as natural relatives. This work lays the foundation for a new experimental method (3Dseq) for protein structure determination, combining evolution experiments with inference of residue interactions from sequence information. A record of this paper's Transparent Peer Review process is included in the Supplemental Information.

INTRODUCTION

By continually generating random DNA sequence variation and selecting for survival, evolution has accumulated a coded record of the physicochemical constraints of the molecular components in evolving organisms. With advances in high-throughput sequencing technology, we now have access to extensive portions of this record in the form of DNA and protein sequence databases. Detecting sequence patterns in homologous proteins has allowed researchers to reconstruct phylogenetic trees and identify sets of functional amino-acid residues. A recent breakthrough in the computational protein structure prediction problem uses maximum entropy statistical analysis of co-evolution in protein and RNA families to enable the computation of important interactions between residues or bases and, from those, calculates accurate three-dimensional (3D) folds and complexes (Marks et al., 2011, 2012).

Here, we asked if evolution performed in the laboratory, with its simplified and controllable evolutionary dynamics, similarly can encode information on structural interactions. In contrast to experimental evolution, natural evolution is complex, occurring over long time periods, with highly variable population sizes (Wright, 1931), mutation rates (Itoh et al., 2002; Tanaka et al., 2003), and fluctuating environmental conditions (Bell, 2010; Haldane and Jayakar, 1963; Kawecki et al., 2012; Lande, 1976; Mustonen and Lässig, 2008; Poelwijk et al., 2011). Each of these factors may or may not be essential for deposition of structural constraints in the evolved sequences. For example, it has been suggested that co-evolutionary patterns arise by the continuous degradation and restoration of protein function, i.e., compensatory evolution (Bloom et al., 2006; DePristo et al., 2005; Tokuriki and Tawfik, 2009), driven by fluctuations in population sizes (Gillespie, 1999) or periods where functional selection is absent (Bell, 2010). Additionally, natural protein family members may vary in function, operate in various cellular environments, in different temperature regimes, and have a broad sequence diversity that is so far unattainable by experimental evolution. The motivation for this work is to use experimental evolution to elucidate the evolutionary determinants that give rise to co-evolutionary patterns and structural constraints of proteins.

RESULTS

We subjected two bacterial antibiotic resistance proteins—the *Pseudomonas* β-lactamase PSE1 and aminoglycoside acetyltransferase AAC6—to experimental evolution by repeated rounds of mutation and selection for preservation of function, a procedure also known as "laboratory drift" or "neutral drift experiments" (Bershtein et al., 2008; Gupta and Tawfik, 2008) (Figure 1, STAR Methods). To promote sequence divergence, we applied a high mutation rate using error-prone polymerase chain

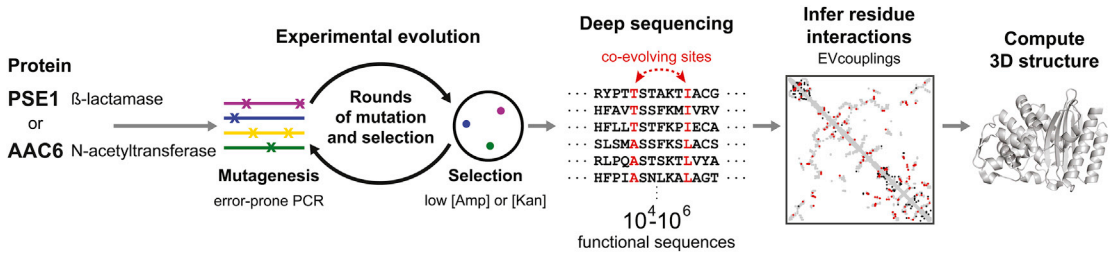


Figure 1. Approach: from Experimental Evolution to Residue Interactions and 3D Structures

The experiments involve repeated rounds of mutation and selection, starting from a single sequence (ß-lactamase PSE1, 266 residues; or aminoglycoside acetyltransferase AAC6, 148 residues). In each round, mutations are generated by error-prone PCR, followed by selection in *E. coli* for functional variants at relatively low antibiotic concentration (6 μg/mL ampicillin [Amp] for PSE1 and 10 μg/mL kanamycin [Kan] for AAC6). A large number of full-length sequences at various rounds are obtained by deep sequencing after selection; here, at rounds 10 and 20 for PSE1, and rounds 2, 4 and 8 for AAC6. Residue interactions are inferred from co-evolution patterns in the selected sequences using the evolutionary couplings (EVcouplings (Marks et al., 2011)) maximum entropy model, which are then used as distance constraints to compute 3D structures using distance geometry and simulated annealing molecular dynamics (Brunger, 2007).

reaction (epPCR; introducing approximately 3%–4% amino-acid substitutions per round) and selected for functional proteins under permissive selective conditions (6 μg/mL ampicillin for PSE1 and 10 μg/mL kanamycin for AAC6—slightly above the minimal inhibitory concentration, MIC, for *E. coli* lacking a resistance gene). These conditions generally resulted in survival of ~1% of the initial population post-selection (approx. 5×10^4 cells for PSE1 and 2×10^5 for AAC6) in each round. Successive rounds of mutation and selection were applied by using the selected sequences in one round as the template for mutations in the next round. We deep-sequenced the selected populations, obtaining 10^4–10^6 high-quality unique reads, at rounds 10 and 20 for PSE1 and rounds 2, 4, and 8 for AAC6 (STAR Methods).

Sequencing revealed that the mutation count relative to the ancestral sequence increases with the number of rounds of mutation and selection (Figure 2A). In the final round, the evolved sequences have an average of 34.2 missense mutations (12.9% of sequence length) in PSE1 and 8.7 missense mutations (5.9%) in AAC6. Thus, of the 3%–4% amino acid mutations introduced per round in each sequence, the functionally selected sequences end up with an average of 1.7 (0.6%) and 1.2 (0.8%) amino acid mutations for PSE1 and AAC6, respectively. In the later rounds, there was a trend toward fewer tolerated mutations; e.g., for PSE1, 1.9 mutations were added per round up to round 10, and 1.5 mutations were added per round between rounds 10 and 20.

The mutational distance to the ancestor is not by itself a measure of diversity, as the libraries could potentially consist of sets of very similar sequences. To assess diversity, we monitored the all-against-all pairwise sequence differences in each population. We observed an average of 19.8% and 10.9% pairwise sequence differences in the final round of PSE1 and AAC6 evolution, respectively (Figure 2B). This equates to an increase in pairwise sequence difference of 1.0% per round for PSE1 and 1.4% for AAC6—close to the maximum possible increase if the populations were freely expanding in sequence space; we conclude that our approach effectively generates and preserves a high level of sequence diversity. In contrast, the mean pairwise sequence difference within the set of known *natural* homologs of PSE1 or AAC6 is around 80% (Figure 2B), which is also evidenced by the higher level of mutational entropy at each sequence position (Figure 3). Projected onto a two-dimensional sequence space (STAR Methods), the experimentally evolved sequences increasingly disperse with increasing rounds of mutation and selection but otherwise occupy a small and dense area relative to natural sequences (Figure 2C).

Information about evolutionary constraints in iso-functional sequences increases both with sequence diversity and with the total number of non-identical sequences (Marks et al., 2011; Sheridan et al., 2015). Although the level of diversity and positional entropy in the experimentally evolved sequences is *lower* than in families of natural homologs (Figures 2C, 3C, and 3D), we have generated many *more* experimentally evolved sequences than are currently available for natural homologs (final experimental evolution rounds have 1.6×10^5 unique functional sequences derived from PSE1 and 1.3×10^6 for AAC6; the PFAM database has 3.7×10^4 homologs for PSE1 [PFAM: PF00144] and $\sim 1.2 \times 10^5$ for AAC6 [PFAM: PF00583]).

To quantify the extent to which evolution in the laboratory has encoded co-variation patterns that are informative of interactions between pairs of residue positions, we used a global probability model (EVcouplings) that has been successful in detecting such patterns in natural sequences (Figure 1, STAR Methods) (Lapedes et al., 1999; Marks et al., 2011; Sheridan et al., 2015; Stein et al., 2015). We compared the inferred interactions to actual contacts in published crystal structures closest in sequence to each of the two ancestral proteins (PDB: 1G68 for PSE1; PDB: 4EVY for AAC6). Comparison with crystal structures tests whether functional selection (i.e., enzymatic deactivation of antibiotic) conserves 3D structure—it is well established from analysis of natural sequences and structures that there is a high degree of structural conservation among iso-functional homologs even with highly diverged sequences (Sander and Schneider, 1991). However, in experimental evolution, there is a possibility that we generate more structural variability compared to natural evolution, as some aspects of protein fitness, such as those related to aggregation or thermodynamic stability (DePristo et al., 2005; Geiler-Samerotte et al., 2011), may be under weaker selection in the laboratory than in nature, given the much shorter timescales, smaller population sizes, and homogeneous environments (Gillespie, 1994).

In practice, we defined contact agreement as the percentage of top-ranked inferred interactions that are also contacts in the crystal structure, typically for $L/2$ inferred pairwise interactions where L is the length of the protein (STAR Methods). Agreement

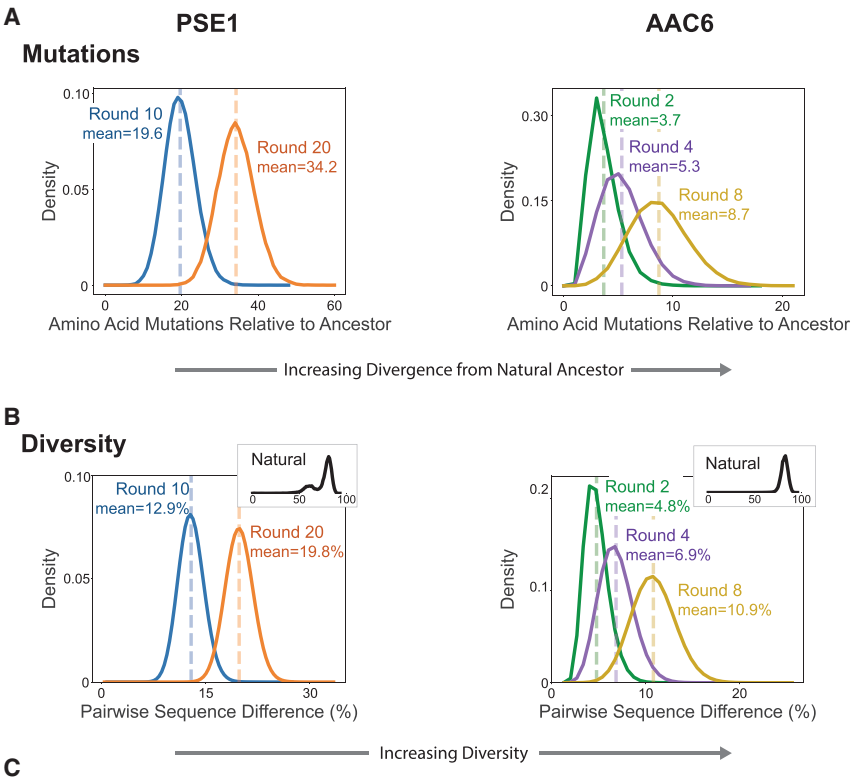


Figure 2. Divergence, Diversity, and Sequence Space Landscape of Experimentally Evolved Sequences

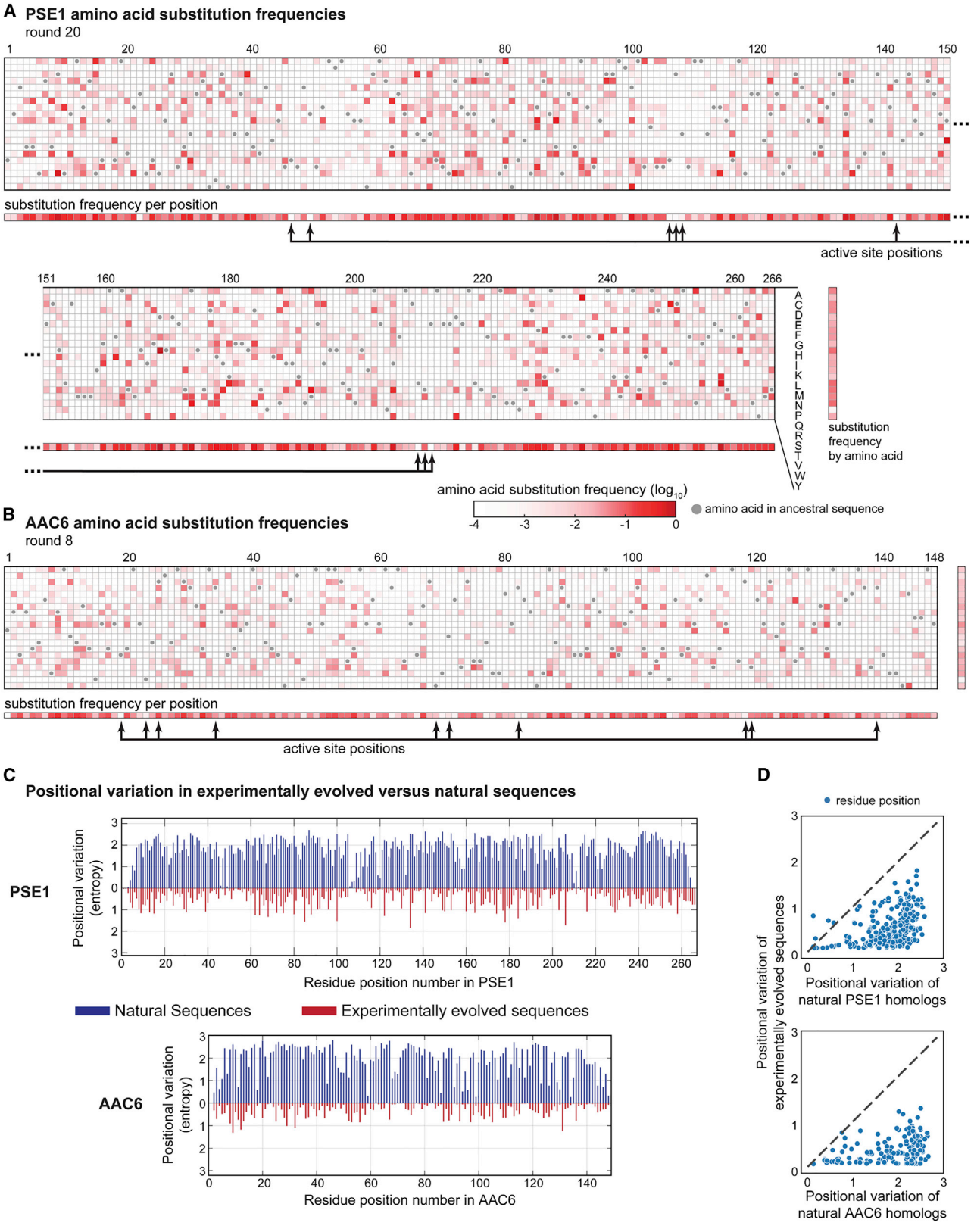
(A) Distributions of amino acid mutations relative to the starting (ancestral) sequence, per unique sequence obtained from experimental evolution of PSE1 (top) and AAC6 (bottom).

(B) Distributions of pairwise sequence diversity (percent positions with non-identical amino acids) calculated for 25 × 10^6 randomly chosen pairs of unique sequences for PSE1 (top) and AAC6 (bottom). Sequence diversity among natural homologs is substantially larger (inset at top right; see STAR Methods).

(C) Two-dimensional representation of sequence sets (each point is one protein sequence) in sequence space projected down from the N-dimensional space using a variational autoencoder (Riesselman et al., 2018) with two latent variables (z_1 and z_2) (STAR Methods). Sequences of natural homologs (current databases, colored by taxonomy) occupy a much larger space than those from our evolution experiments. The experimentally evolved sequences increasingly separate from the ancestral sequence with increasing rounds of mutagenesis (point cloud on right).

increases with successive rounds of mutation and selection (Figure 4A) as sequence diversity increases. For PSE1, agreement increased from 34% in round 10 to 49% in round 20 (for a subset of 4 × 10^4 unique sequences from each round). For AAC6 agreement was 22% in round 2, 30% in round 4, and 42% in round 8 (for an equal size subset of 10^5 unique sequences). These percentages are well above the random expectation of 1.9% and 4.1% for PSE1 and AAC6 (approximated by the ratio of crystal structure contacts over the total number of pairs), even at early rounds (see AAC6 round 2, Figure 4A). Inclusion of additional sequences also led to a higher agreement between inferred interactions and crystal structure contacts: for the last rounds, we obtained 1.6 × 10^5 unique functional sequences for PSE1 and 1.3 × 10^6 for AAC6, leading to a contact agreement of 54% for PSE1 and 51% for AAC6 (Figure 4). These results indicate that simplified evolutionary dynamics in the laboratory do generate functional sequences with co-evolutionary patterns that reflect constraints imposed by protein 3D structure.

The residue interactions inferred from experimental evolution include important structural features of both proteins (Figure 4B). The β-lactamase fold, for example, consists of two structural domains: an all-α-helical domain and a mixed α/β domain (Herzberg and Moult, 1987; Matagne et al., 1998). The polypeptide chain is interwoven between the two domains, leading to extremely sequence-distal contacts between the N- and C-terminal α-helices and between the N- and C-terminal β-strands. Consistent with the crystal structures, we likewise observe strongly co-evolving interactions between these sequence-distal structural elements in the PSE1 results (i.e., upper right and lower left corners of the contact map in Figure 4B). Proteins from the same subfamily of aminoglycoside acetyltransferases as AAC6 exist as homodimers, with the C-terminal β-strand of one protein chain inserted between two strands of the other chain (Stogios et al., 2017). We also discern these dimer interactions among the top-ranked inferred residue-residue interactions (Figure 4B).



(legend on next page)

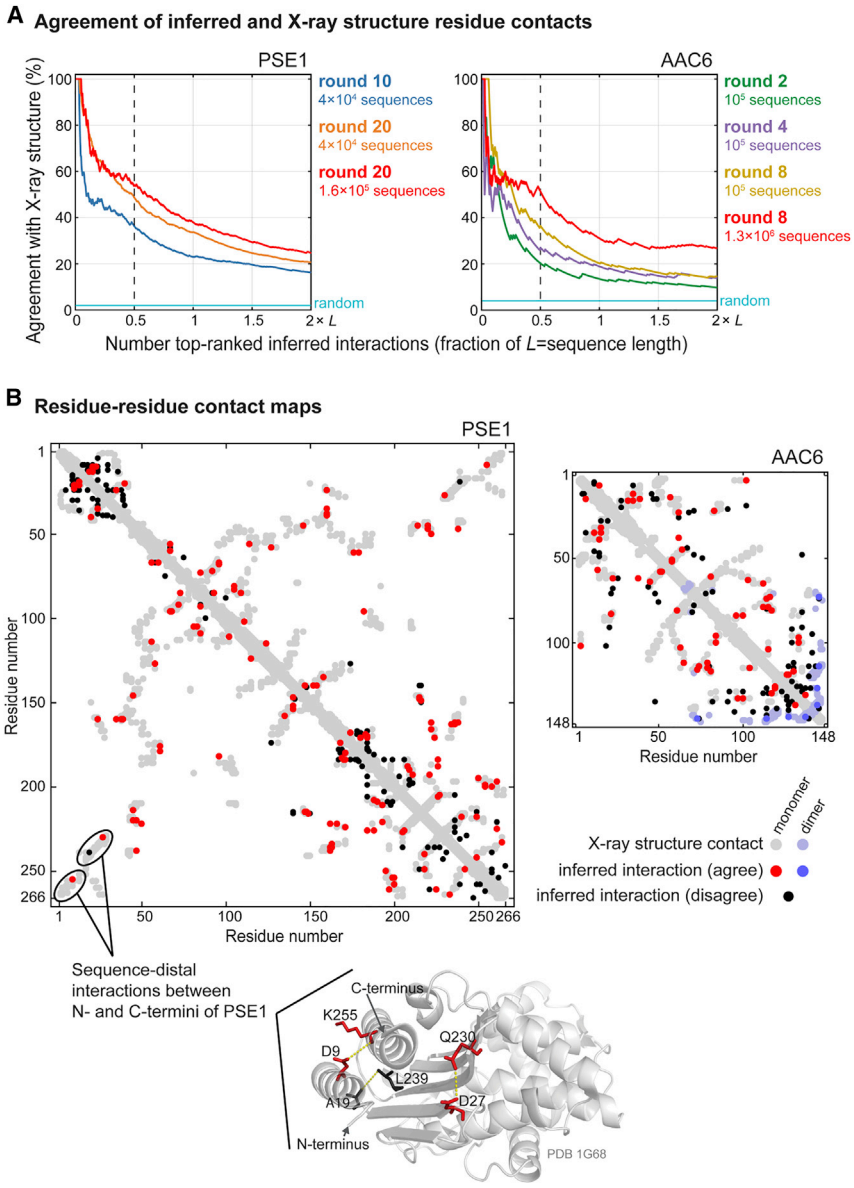


Figure 4. Agreement between Residue Contacts Inferred from Experimental Evolution and Contacts in Crystal Structures

(A) Agreement versus number of inferred interactions (as fraction of sequence length, *L*) during experimental evolution of PSE1 (left) and AAC6 (right). PSE1 results evaluated for an equal number (4×10^4) of randomly subsampled unique sequences from rounds 10 and 20 to illustrate change in agreement with increased rounds of mutation and selection, and all (1.5×10^5) unique sequences at round 20 to illustrate change with increased number of sequences. AAC6 similarly assessed for an equal number (10^5) of randomly subsampled unique sequences at rounds 2, 4 and 8, and all (1.3×10^6) unique sequences at round 8. Random is the average result obtained with randomly chosen residue pairs.

(B) Inferred interactions from PSE1 evolution at round 20 (left) and AAC6 evolution at round 8 (right), overlaid on contact maps of crystal structures. Inferred interactions either agree with monomer (red) or dimer (blue) contacts in the crystal structure (gray or light blue, respectively), or disagree (black). For PSE1, sequence-distal residue interactions between the N- and C-terminal α-helices and β-strands (lower left corner of contact map and indicated on crystal structure of PSE1) are particularly crucial constraints for the correct 3D fold via reduction of chain entropy. Dashed line in (A) and results in (B) are at *L*/2 inferred interactions; agreement of > 50% at *L*/2 often suffices to compute 3D structures (Hopf et al., 2012; Marks et al., 2012). In (A) and (B), residues in the known crystal structure are defined to be in contact if at least one atom-atom distance is < 5 Å; inferred residue-residue interactions are limited to a primary sequence distance > 5 residues.

Similar to previous work on natural sequences (Hopf et al., 2014), identification of these inter-protein contacts demonstrates that experimental evolution can also be informative of protein-protein interactions.

We next asked whether the inferred contacts from experimental evolution are sufficient to compute the 3D structure. For natural protein families, inferred residue interactions in the range of agreement with crystal structures of 50%–60% are typically sufficient to compute 3D folds that agree with those observed by crystallography or NMR (Hopf et al., 2012; Marks et al., 2011). To assess whether experimental evolution provides a similar level of information, we computed sets of structures using the inferred residue interactions as distance constraints in molecular dynamics with simulated annealing (Brunger, 2007; Hopf et al., 2012; Marks et al., 2011). The constraints are updated in an iterative process that filters the inferred interactions for geometric violations, i.e., interactions that are inconsistent with cooperatively folded 3D structures (STAR Methods). For AAC6, we folded only the monomeric

Figure 3. Sequence Variation per Residue Position in the Selected Sequence Libraries

(A and B) Frequencies of amino acid substitutions per residue position in the set of experimentally evolved sequences from (A) round 20 for PSE1 and (B) round 8 for AAC6. Columns: residue positions; rows: amino acids; gray dots: amino acids in the ancestral sequence; horizontal bars below each matrix: total substitution frequency per position (i.e., the fraction of non-ancestral amino acid). Active site positions (arrows) generally have a low substitution frequency. Vertical bars at right of matrices: substitution frequency of each amino acid averaged over all positions, indicating a broadly similar spectra for both proteins. The data to generate this figure are available at https://github.com/sanderlab/3Dseq.

(C and D) Positional variation, quantified as Shannon entropy using the substitution frequencies at each residue position, compared between experimentally evolved and natural sequences. Results for natural sequences are blue bars in (C) and along x axis in (D); results for experimentally evolved sequences are red bars in (C) and along y axis in (D). The level of positional variation observed in the natural sequences is generally not exceeded in the experimentally evolved sequences (empty upper left triangle (D)). Results for PSE1 at round 20 are at the top, and AAC6 at round 8 at the bottom (C and D).

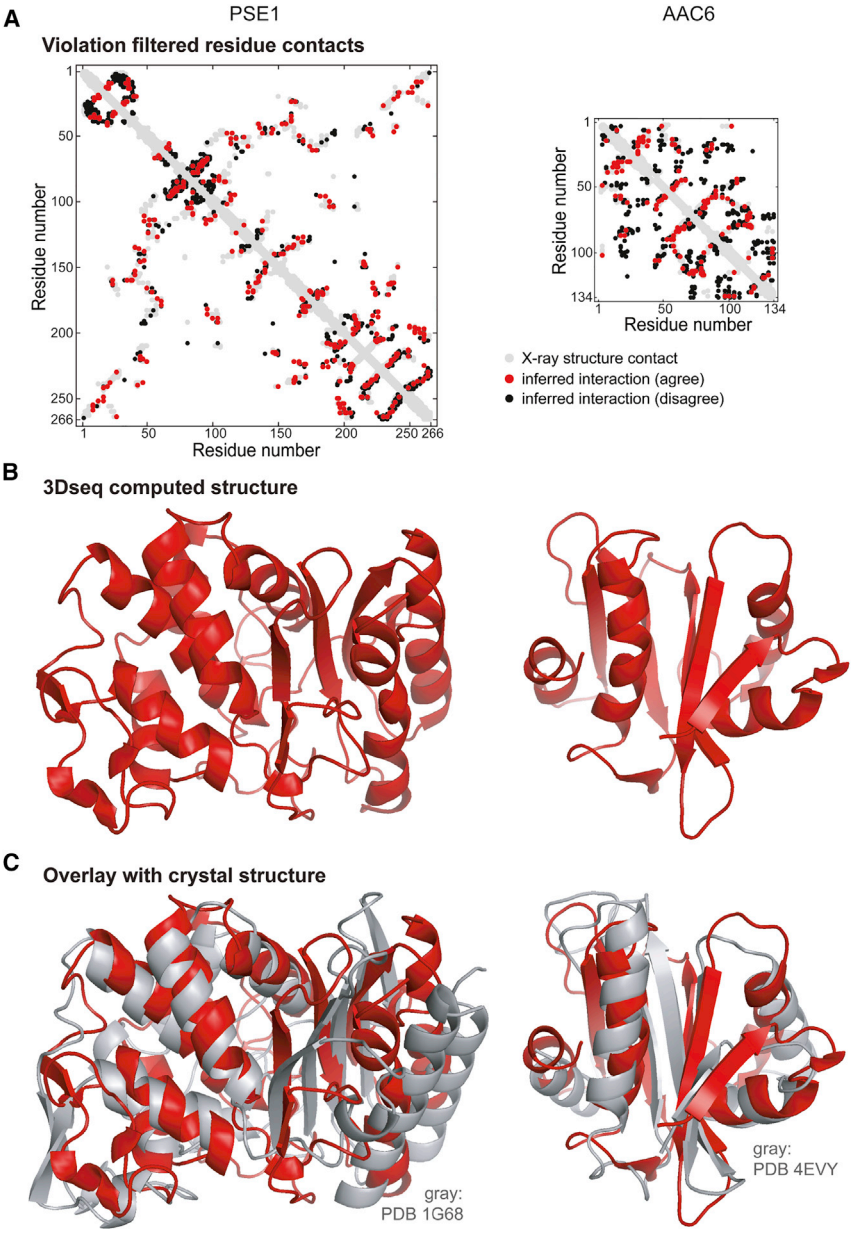


Figure 5. 3D Structures Computed from Experimental Evolution Compared to Those from X-ray Crystallography

(A) Inferred interactions iteratively filtered for geometric violations (STAR Methods); red interactions agree with contacts in the crystal structure, black ones disagree (results for $2 \times L$ inferred interactions, where L is the length of the protein sequence).

(B) 3D structures computed using the filtered inferred interactions as distance constraints (STAR Methods). Red ribbons are structures with the lowest Cα positional RMSD, evaluated over at least 90% of residues for either protein.

(C) 3Dseq computed structures (red ribbons) compared to crystal structures (gray ribbons): left for PSE1 (PDB: 1G68, Cα positional RMSD 4.5 Å over 240/266 residues, TM-score = 0.65); right for AAC6 (using a structural homolog of AAC6, PDB: 4EVY, Cα positional RMSD 3.8 Å over 122/130 residues, TM score = 0.59). For AAC6, the C-terminal β-strand known to be involved in dimer contact is excluded; we did not attempt to compute dimer structures.

unit consisting of residues 1–134, as residues 135–148 are directly involved in dimer contacts and computation of a dimeric molecule from a compounded contact map is beyond the scope of this work (STAR Methods). The removal of geometric violations led to improved agreement between inferred interactions and crystal structure contacts, from 54% to 65% for PSE1, and from 45% to 59% for AAC6 (Figures 5A and S3).

The final set of structures, computed from the inferred interactions after filtering for geometric violations, was assessed for agreement with the known crystal structure of the most sequence-similar homolog. Of the set of computed structures, 72% of PSE1 and 63% of AAC6 generated structures (of 690 models generated for PSE1 and 720 for AAC6 [STAR Methods]) have a template modeling score (TM score) of 0.5 or greater (Figure 6A); TM scores in excess of 0.5 are generally considered to indicate overall fold similarity (Xu and Zhang, 2010; Zhang and Skolnick, 2004, 2005). The structures with the lowest Cα positional root-mean-square deviation (RMSD) over more than 90% of residues and which do not contain knots (Kolesov et al., 2007; Virnau et al., 2006) have 4.5 Å RMSD for PSE1 (240/266 residues using PDB: 1G68) and 3.8 Å RMSD for AAC6 (122/130 residues using PDB: 4EVY) (Figures 5B and 5C). Overall, we conclude that the experimental evolution process mimics natural evolution in constraining residue interactions that conserve the 3D fold.

DISCUSSION

Here, we show for two genes that experimental evolution in the laboratory can generate an amount and type of genetic variation that informs about amino acid residue interactions in 3D protein structures, leading to an accurate determination of these protein structures via evolutionary coupling analysis and restrained molecular dynamics. For either gene, the experiments started out with a single DNA sequence, which makes the approach particularly useful to explore residue interactions in genes with few known homologs, such as so-called orphan genes (Tautz and Domazet-Lošo, 2011)—situations that preclude homology-based approaches (Marks et al., 2011).

The type of experimental evolution employed here, also known as neutral genetic drift or laboratory drift (Bershtein et al., 2008; Bloom et al., 2007a), has been employed before but with a different scientific intent: either to measure protein mutational tolerance and robustness (Bloom et al., 2007b; Rockah-Shmuel et al., 2015) or to diversify libraries prior to directed evolution

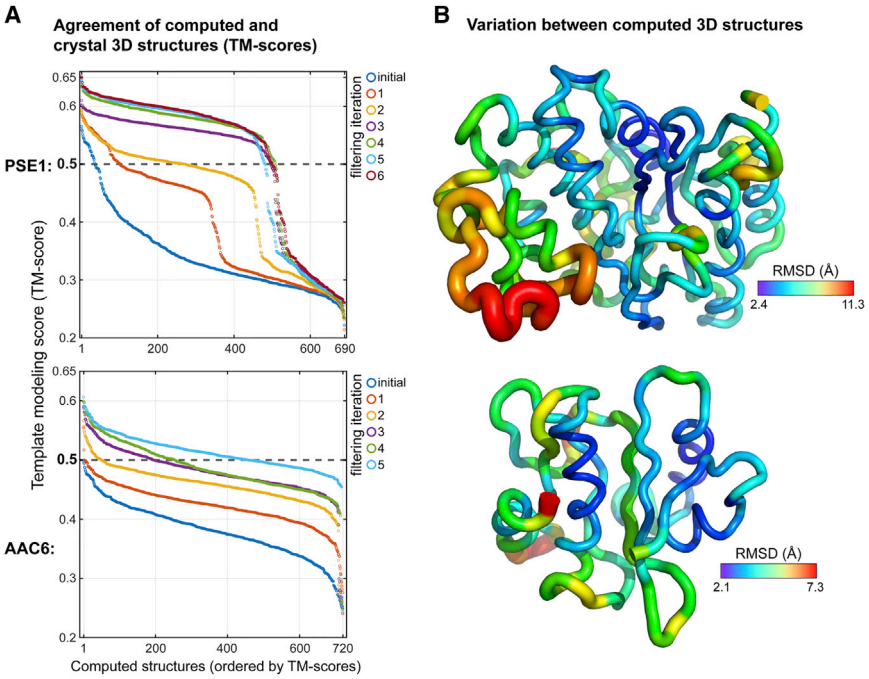


Figure 6. Variation among Computed 3D Structure Models

(A) Template modeling scores (TM-scores [Zhang and Skolnick, 2005]) for all computed models during geometric violation filtering iterations. Computed models are sorted along the x axis by TM-score within each iteration. Overall, structures computed with interactions inferred from experimentally evolved sequences have the same general fold as the crystal structure, with TM-scores of > 0.5 for 72% of PSE1 models (690 total models) and for 63% of AAC6 models (720 total models) in the final iteration for both proteins.

(B) Structural variation between computed models. The color and radius of each residue is monotonically related to the RMSD of Cα-Cα distances computed from all-versus-all pairwise superposition of models in the largest cluster (MaxCluster [Herbert and Sternberg, 2008]) from the final filtering iteration (STAR Methods).

(Bershtein et al., 2008; Bloom et al., 2007a). These studies found evidence for global suppressor mutations—single mutations that broadly compensate for the deleterious effects of a substantial number of other mutations (Bershtein et al., 2008)—but not for residue-residue interactions that are local in the protein structure. The fact that we do find evidence for contact interactions could be either due to differences in the experimental implementation—such as the number of applied mutations per round—or due to the fact that we reached a higher level of genetic diversity in combination with a larger number of gene variants sequenced over their full-length than previously attained.

This study lays the foundation for a new *ab initio* experimental method of determining protein structure, which we call 3Dseq. Using experimental evolution for structure determination is complementary to established methods such as X-ray crystallography, NMR, and cryo-electron microscopy—with several advantages and disadvantages. A major advantage is that determination of a structure by 3Dseq does not require biochemical purification of the protein. Similarly, beyond single proteins, inter-molecular interactions can be elucidated without purification and/or crystallization of complexes (see dimer contacts in AAC6, Figure 4B) using protein-protein interaction assays to select variant sequences, such as two-hybrid (Dove et al., 1997; Fields and Song, 1989; McLaughlin et al., 2012) or phage or yeast display assays (Boder and Wittrup, 1997; Sidhu and Geyer, 2015). Further, by controlling external conditions, one can directly infer which constrained interactions or structural variants are of functional importance under a given selection condition. A potential disadvantage arises from the fact that the distance constraints inferred from co-variation patterns in 3Dseq are an average property of the set of selected sequences and that single-sequence specificity is only implemented by constrained molecular dynamics. Thus, one would expect the precision of atomic positions to typically be less than that from, e.g., single-sequence, single-conformation crystallography. In the future, this difference in precision is likely to be reduced by improvements in the constrained molecular dynamics part of the 3Dseq method. One could use 3Dseq to fully explore structural variation within a set of sequences by computing single-sequence structures for all sequences in the experimentally selected libraries—rather than just the ancestral sequence as done here; however, executing many thousands constrained molecular dynamics runs is beyond the scope of this work. In any case, high precision of 3D coordinates for a single sequence does not necessarily imply a tight conformational ensemble of structures in physiological conditions. As in NMR spectroscopy, explicit data on conformational ensembles inferred from 3Dseq may provide detailed insight regarding structure-function relationships.

Future generalization of 3Dseq experimental technology would benefit from assays that directly select or screen for protein structural integrity (Cabantous et al., 2005; Foit et al., 2009; Waldo et al., 1999) and do not depend directly on selection for a particular cellular function (e.g., antibiotic resistance). A major efficiency gain would come from automated evolution systems, which combine mutation and selection in single cells and rely on proliferative advantage in pooled experiments (Badran and Liu, 2015; Esvelt et al., 2011; Ravikumar et al., 2018; Takahashi et al., 2015; Toprak et al., 2011).

In using experimental evolution to elucidate co-variation constraints one can be agnostic as to the detailed mechanism by which the constraints are encoded in the genetic sequences. Our recent work (Rollins et al., 2019) and related work (Schmiedel and Lehner, 2019) showed that two-way epistasis between amino acid mutations, derived from complete pairwise "deep" mutational scans with quantitative (non-binary) measurement of fitness, can reflect structural constraints sufficient for the computation of 3D structures, at least for small proteins. In general, with increasing ability to

control sequence diversity, external conditions, depth of sequencing and quantitation of fitness, future evolution experiments will provide further opportunities to unravel details of evolutionary pathways, to determine the level of cooperativity in sets of mutations, and to refine the discovery of sets of constraints essential for the implementation of particular functions.

In contrast to natural evolution, experimental evolution is typically performed in less diverse environments, over much shorter timescales, and with simpler population dynamics (Gillespie, 1994; Kawecki et al., 2012). Nonetheless, our results indicate that laboratory-based experimental evolution consisting of repeated cycles of random mutation and uniform selection can generate large and diverse sets of sequences with rich co-evolutionary interaction patterns. Experimental evolution approaches of this type can contribute to a better understanding of the complexities of natural evolution, to the design of useful proteins, and to the development of quantitative models, in molecular detail, of both retrospective and prospective evolutionary dynamics.

KEY CHANGES PROMPTED BY REVIEWER COMMENTS

In response to the reviewers, we clarified several points in the text—such as the scope of the project—and mentioned ongoing developments. We added references to laboratory drift studies and discussion of contrasts with our approach. We moved a figure from Supplemental Information into the Main Text (now Figure 3) and added a new figure about the variation between calculated structural models (Figure 6). We clarified some procedures, such as the evaluation of contact agreement. For context, the complete Transparent Peer Review Record is included within the Supplemental Information.

Note Added in Revision

Related work aiming to infer residue-residue interactions from laboratory evolved sequences has recently been reported (Fantini et al., 2019), though sufficiency for determining 3D structure was not reported.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - General
  - Plasmid Construction
  - Mutagenesis
  - Selection
  - Sequencing
  - Sequence Analysis
  - Alignment Conditioning
  - EVcouplings Analysis
  - Compute Structures from Inferred Interactions
  - Visualization of Sequence Space
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information is online at https://doi.org/10.1016/j.cels.2019.11.008.

ACKNOWLEDGMENTS

We thank our colleagues for their contributions. EVcouplings pipeline: Thomas Hopf, Christian Dallago, Benjamin Schubert; DeepSequence software package: John Ingraham; Sequencing: the Molecular Biology Core Facilities at Dana-Farber (Zack Herbert, Maura Berkeley), the MIT BioMicro Center, and the Genomics Core Facility at the Icahn Institute, Mt. Sinai (Melissa Smith, Diane Castillo, Gintaras Deikus); Pacbio process: Michael Weiand from Pacific Biosciences; Discussion: members of the Sander and Marks labs. We thank Barrett Rollins for guidance and support. Funding was obtained from Dana-Farber Cancer Institute and grant NIGMS-R01GM106303.

AUTHOR CONTRIBUTIONS

Project initiation: C.S.; First project phase: S.S.S., J.T., F.J.P., D.S.M., R.R.S., C.S., N.P.G.; Concepts: M.A.S., F.J.P., S.S.S., D.S.M., N.P.G., C.S.; Data curation & analysis: M.A.S., F.J.P., K.P.B., R.R.S., N.P.G.; Methodology: M.A.S., F.J.P., J.T., S.S.S., N.P.G., C.S.; Project management: N.P.G.; Software: M.A.S., F.J.P., K.P.B., A.R., R.R.S., N.P.G.; Visualization: M.A.S., F.J.P., N.P.G.; Writing: M.A.S., F.J.P., K.P.B., N.P.G., C.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 13, 2019
Revised: September 6, 2019
Accepted: November 20, 2019
Published: December 11, 2019

REFERENCES

Badran, A.H., and Liu, D.R. (2015). Development of potent in vivo mutagenesis plasmids with broad mutational spectra. Nat. Commun. *6*, 8425.

Bell, G. (2010). Fluctuating selection: the perpetual renewal of adaptation in variable environments. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 87–97.

Bershtein, S., Goldin, K., and Tawfik, D.S. (2008). Intense neutral drifts yield robust and evolvable consensus proteins. J. Mol. Biol. *379*, 1029–1044.

Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. (2006). Protein stability promotes evolvability. Proc. Natl. Acad. Sci. USA *103*, 5869–5874.

Bloom, J.D., Romero, P.A., Lu, Z., and Arnold, F.H. (2007a). Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. Biol. Direct *2*, 17.

Bloom, J.D., Lu, Z., Chen, D., Raval, A., Venturelli, O.S., and Arnold, F.H. (2007b). Evolution favors protein mutational robustness in sufficiently large populations. BMC Biol. *5*, 29.

Boder, E.T., and Wittrup, K.D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. Nat. Biotechnol. *15*, 553–557.

Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.C., Heyneker, H.L., Boyer, H.W., Crosa, J.H., and Falkow, S. (1977). Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene *2*, 95–113.

Brunger, A.T. (2007). Version 1.2 of the Crystallography and NMR system. Nat. Protoc. *2*, 2728–2733.

Cabantous, S., Terwilliger, T.C., and Waldo, G.S. (2005). Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. Nat. Biotechnol. *23*, 102–107.

DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. *6*, 678–687.

Dove, S.L., Joung, J.K., and Hochschild, A. (1997). Activation of prokaryotic transcription through arbitrary protein-protein contacts. Nature *386*, 627–630.

Esvelt, K.M., Carlson, J.C., and Liu, D.R. (2011). A system for the continuous directed evolution of biomolecules. Nature *472*, 499–503.

Fantini, M., Lisi, S., De Los Rios, P., Cattaneo, A., and Pastore, A. (2019). Protein Structural Information and Evolutionary Landscape by In Vitro Evolution. Mol. Biol. Evol. msz256.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245–246.

Foit, L., Morgan, G.J., Kern, M.J., Steimer, L.R., von Hacht, A.A., Titchmarsh, J., Warriner, S.L., Radford, S.E., and Bardwell, J.C.A. (2009). Optimizing protein stability in vivo. Mol. Cell *36*, 861–871.

Gatti-Lafranconi, P. (2014). Pymol script: loadBfacts.py (Figshare).

Geiler-Samerotte, K.A., Dion, M.F., Budnik, B.A., Wang, S.M., Hartl, D.L., and Drummond, D.A. (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc. Natl. Acad. Sci. USA *108*, 680–685.

Gillespie, J.H. (1994). The Causes of Molecular Evolution (Oxford University Press).

Gillespie, J.H. (1999). The role of population size in molecular evolution. Theor. Popul. Biol. *55*, 145–156.

Gupta, R.D., and Tawfik, D.S. (2008). Directed enzyme evolution via small and effective neutral drift libraries. Nat. Methods *5*, 939–942.

Haldane, J.B.S., and Jayakar, S.D. (1963). Polymorphism due to selection of varying direction. J. Genet. *58*, 237–242.

Herbert, A., and Sternberg, M.J.E. (2008). MaxCluster: a tool for protein structure comparison and clustering.

Herzberg, O., and Moult, J. (1987). Bacterial resistance to beta-lactam antibiotics: crystal structure of beta-lactamase from Staphylococcus aureus PC1 at 2.5 A resolution. Science *236*, 694–701.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. Cell *149*, 1607–1621.

Hopf, T.A., Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M.J.J., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife *3*, 1–19.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. *33*, 1635–1638.

Itoh, T., Martin, W., and Nei, M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc. Natl. Acad. Sci. USA *99*, 12944–12948.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195–202.

Kawecki, T.J., Lenski, R.E., Ebert, D., Hollis, B., Olivieri, I., and Whitlock, M.C. (2012). Experimental evolution. Trends Ecol. Evol. *27*, 547–560.

Kolesov, G., Virnau, P., Kardar, M., and Mirny, L.A. (2007). Protein knot server: detection of knots in protein structures. Nucleic Acids Res. *35*, W425-8.

Lande, R. (1976). Natural Selection and Random Genetic Drift in Phenotypic Evolution. Evolution *30*, 314–334.

Lapedes, A.S., Giraud, B.G., Liu, L.C., and Stormo, G.D. (1999). Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. Lect. Notes Monogr. Ser. *33*, 236–256.

Lim, D., Sanschagrin, F., Passmore, L., De Castro, L., Levesque, R.C., and Strynadka, N.C. (2001). Insights into the molecular basis for the carbenicillinase activity of PSE-4 beta-lactamase from crystallographic and kinetic studies. Biochemistry *40*, 395–402.

Lutz, R., and Bujard, H. (1997). Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Res. *25*, 1203–1210.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS ONE *6*, e28766.

Marks, D.S., Hopf, T.A., and Sander, C. (2012). Protein structure prediction from sequence variation. Nat. Biotechnol. *30*, 1072–1080.

Matagne, A., Lamotte-Brasseur, J., and Frère, J.M. (1998). Catalytic properties of class A beta-lactamases: efficiency and diversity. Biochem. J. *330*, 581–598.

McLaughlin, R.N., Jr., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Mustonen, V., and Lässig, M. (2008). Molecular evolution under fitness fluctuations. Phys. Rev. Lett. *100*, 108101.

Poelwijk, F.J., de Vos, M.G.J., and Tans, S.J. (2011). Tradeoffs and optimality in the evolution of gene regulation. Cell *146*, 462–470.

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Res. *46* (W1), W200–W204.

Ravikumar, A., Arzumanyan, G.A., Obadi, M.K.A., Javanpour, A.A., and Liu, C.C. (2018). Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. Cell *175*, 1946–1957.e13.

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat. Methods *15*, 816–822.

Rockah-Shmuel, L., Tóth-Petróczy, Á., and Tawfik, D.S. (2015). Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. PLoS Comput. Biol. *11*, e1004421.

Rollins, N.J., Brock, K.P., Poelwijk, F.J., Stiffler, M.A., Gauthier, N.P., Sander, C., and Marks, D.S. (2019). Inferring protein 3D structure from deep mutation scans. Nat. Genet. *51*, 1170–1176.

Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins *9*, 56–68.

Schmiedel, J.M., and Lehner, B. (2019). Determining protein structures using deep mutagenesis. Nat. Genet. *51*, 1177–1186.

Sheridan, R., Fieldhouse, R.J., Hayat, S., Sun, Y., Antipin, Y., Yang, L., Hopf, T., Marks, D.S., and Sander, C. (2015). EVfold.org: Evolutionary Couplings and Protein 3D Structure Prediction. bioRxiv ●●●, 021022.

Sidhu, S.S., and Geyer, C.R. (2015). Phage Display. In Biotechnology and Drug Discovery (CRC Press).

Stein, R.R., Marks, D.S., and Sander, C. (2015). Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. PLoS Comput. Biol. *11*, e1004182.

Stogios, P.J., Kuhn, M.L., Evdokimova, E., Law, M., Courvalin, P., and Savchenko, A. (2017). Structural and Biochemical Characterization of Acinetobacter spp. Aminoglycoside Acetyltransferases Highlights Functional and Evolutionary Variation among Antibiotic Resistance Enzymes. ACS Infect. Dis. *3*, 132–143.

Takahashi, C.N., Miller, A.W., Ekness, F., Dunham, M.J., and Klavins, E. (2015). A low cost, customizable turbidostat for use in synthetic circuit characterization. ACS Synth. Biol. *4*, 32–38.

Tanaka, M.M., Bergstrom, C.T., and Levin, B.R. (2003). The evolution of mutator genes in bacterial populations: the roles of environmental change and timing. Genetics *164*, 843–854.

Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. Nat. Rev. Genet. *12*, 692–702.

Tokuriki, N., and Tawfik, D.S. (2009). Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. *19*, 596–604.

Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D.L., and Kishony, R. (2011). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat. Genet. *44*, 101–105.

Virnau, P., Mirny, L.A., and Kardar, M. (2006). Intricate knots in proteins: Function and evolution. PLoS Comput. Biol. *2*, e122.

Waldo, G.S., Standish, B.M., Berendzen, J., and Terwilliger, T.C. (1999). Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. *17*, 691–695.

Wright, S. (1931). Evolution in Mendelian Populations. Genetics *16*, 97–159.

Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889–895.

Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins *57*, 702–710.

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
NEB 10-beta	NEB	C3019I
MC1061	Avidity Inc	AVB100
Critical Commercial Assays		
Genemorph II random mutagenesis kit	Agilent	200550
MiSeq Reagent Kit v3 (600-cycle)	Illumina Inc.	MS-102-3003
PacBio SMRTbell preparation kits	Pacific Biosciences	100-465-800, 100-465-900, 100-466-100
Deposited Data		
Sequencing Reads (FASTQ files)	This paper	Sequence Read Archive BioProject PRJNA578762
Sequences	This paper	https://github.com/sanderlab/3Dse
Sequence alignments	This paper	https://github.com/sanderlab/3Dseq
3D structure model files	This paper	https://github.com/sanderlab/3Dseq
Cα-Cα distances for all pairs of superposed models	This paper	https://github.com/sanderlab/3Dseq
Recombinant DNA		
pBR322_KanR_AgeI_AvrII_PSE1	This paper	Addgene 135229
pBR322_ZA_ampR_PSEAB_2A	This paper	Addgene 135230
Software and Algorithms		
EVcouplings: generates a global probability model from a multiple sequence alignment via maximum entropy reduction with pseudo-likelihood maximization (plm) or mean field approximation. We used this algorithm extensively to extract informative interactions between pairs of residue positions, called evolutionary couplinges, as well as to compute 3D structures via restrained molecular dynamics with the CNS software system (Brunger, 2007)	https://evcouplings.org/	
Deepsequence: scripts to generate a latent variable model on biological sequences. This software was used to project natural homologs and experimentally evolved sequences into a two-dimensional sequence space.	Riesselman et al., 2018	https://github.com/debbiemarkslab/DeepSequence
MaxCluster: this tool compares a set of proteins structures. We applied this program to cluster computed models for comparison of structural variability.	Herbert and Sternberg 2008	http://www.sbg.bio.ic.ac.uk/~maxcluster/
Custom code was used to analyze PacBio and Illumina sequencing reads, generate substitution and mutation and entropy statistics, perform dimensionality reduction, and calculate and visualize variation between 3D structure. Code is available at github.	This paper	https://github.com/sanderlab/3Dseq

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Chris Sander (3Dseq.research@gmail.com).

Plasmids generated in this study that contain ancestral PSE1 and AAC6 sequences have been deposited at Addgene: 135229 and Addgene: 135230, respectively. This study did not generate other unique reagents. Requests for further information, resources, and reagents should be sent to 3Dseq.research@gmail.com, which will reach all principal authors (M.S., F.P., N.G., C.S.), with the Lead Contact (C.S.) responsible for fulfillment of the requests.

Please cite this article in press as: Stiffler et al., Protein Structure from Experimental Evolution, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.11.008

METHOD DETAILS

General

The two proteins used in this work, aminoglycoside acetyltransferase AAC6 and β-lactamase PSE1 are substantially different in length, and we therefore used different selection strategies and sequencing methods. The larger protein (PSE1) was selected on plates and sequenced using PacBio. The smaller protein (AAC6) allowed sequencing by the shorter read length, higher throughput Illumina method, and larger library sizes and was therefore selected in liquid media. Below, details of the procedures followed are given separately for the two proteins as needed.

Plasmid Construction

AAC6

Starting with pBR322 (Bolivar et al., 1977), we replaced the tetracycline resistance gene by the ampicillin resistance gene *bla*, and replaced the native *bla* gene by the pZA promoter (Lutz and Bujard, 1997) driving the expression of the aminoglycoside antibiotic resistance protein 6'-N-acetyltransferase AAC(6')-I from *Pseudomonas sp.* ABAC61 (UniProt sequence identifier A0A0W0NPD4_9PSED). The plasmid map is shown in Figure S1A, the sequence in GenBank format is given in the Supporting Information. The relevant X-ray crystal structure is PDB: 4EVY (Stogios et al., 2017), a homolog of the AAC6 used here with 50% sequence identity (74/148 identical residues).

PSE1

The pBR322 plasmid was modified by replacing the tetracycline resistance gene with a kanamycin resistance gene. An AgeI restriction site was created by making two synonymous mutations at amino acid positions within the periplasmic signal peptide region of the existing TEM-1 *bla* gene, and an AvrII restriction site was created directly downstream of the existing *bla* gene. The existing TEM-1 *bla* gene was replaced by that for *Pseudomonas aeruginosa* PSE1 (Uniprot sequence identifier Q03170), retaining the periplasmic signal peptide sequence of TEM-1 (i.e., only the mature β-lactamase sequences were interchanged). The sequence in GenBank format is provided in the Supporting Information (Figure S1B). The corresponding X-ray crystal structure is PDB: 1G68 (Lim et al., 2001), a homolog of PSE1 (PSE4) that differs by a single amino acid.

Mutagenesis

Random mutagenesis was performed by error-prone PCR (epPCR) using the Genemorph II Random Mutagenesis Kit (Agilent) following the manufacturer's protocol.

AAC6

The input concentration for the epPCR reaction, which determines the rate of mutagenesis, is set at 0.8 ng plasmid DNA, corresponding to 0.1 ng per reaction of the *aac6* coding sequence. The PCR product is gel purified and digested with restriction enzyme *BsaI* in a single reaction together with gel-purified non-mutated PCR product of the vector backbone, followed by ligation of the fragments to form a circular plasmid.

PSE1

Error-prone PCR was performed using 1 ng of the plasmid vector (corresponding to 0.2 ng of the PSE1 gene) as template. In order to reduce the possibility of amplifying contaminating β-lactamases, primers specific to the mature PSE1 protein region were created which overlap with the N-terminal four and C-terminal two amino acids of the mature PSE1 protein sequence. Specific primers were used in generating libraries at all rounds, except in those prior to sequencing in which primers annealing outside the mature PSE1 protein sequence were used to allow some variation at these amino acid positions. PCR products were gel-purified, digested with restriction enzymes AgeI and AvrII, and ligated into the modified pBR322 plasmid vector previously digested with AgeI and AvrII.

Selection

AAC6

Mutant libraries were transformed by electroporation into *E. coli* strain MC1061. After recovery in 1 mL LB for one h at 37°C, a small aliquot (typically 5 μL) was taken out for determination of the transformation and selection efficiencies, and the remainder is grown overnight under selective conditions at 37°C in 100 mL LB and 10 μg/mL kanamycin, under vigorous shaking. A plasmid miniprep is prepared from 6 mL of this culture, and is used for epPCR of the next round. Typical population size pre-selection is 10^7, of which about 1% survives the kanamycin selective conditions.

PSE1

Mutation libraries were transformed by electroporation into NEB 10-beta DH10B *E. coli* cells. After one h recovery in 1 mL of outgrowth medium, a small aliquot was taken for determination of transformation efficiency on LB-agar Petri dishes containing 30 μg/mL kanamycin. The remaining recovery culture was plated onto large Petri plates containing LB-agar and 6 μg/mL ampicillin as the sodium salt; at rounds prior to sequencing, plates contained 10 μg/mL ampicillin to better ensure non-functional sequences are eliminated. Plates were incubated at 37°C overnight (approximately 15 h). Ampicillin selection plates were scraped of colonies using approximately 40 mL of water, cells pelleted by centrifugation, then resuspended in 10 mL of water by vortexing. Plasmid DNA was purified by miniprep for use in another round of epPCR or for deep-sequencing. Typical population size pre-selection was approximately 5×10^6, of which about 1% survive the ampicillin selection conditions.

Sequencing

The platforms for high-throughput sequencing used in this study are different for PSE1 and AAC6, because of the difference in length of their coding regions (798 and 444 nucleotides, respectively) and the requirement of co-evolutionary analysis to observe pairs of amino acid mutations across the entire gene. For PSE1 we exploited the greater read length of the PacBio platform, achieving about 300,000 raw reads of about 10,000 bases, while for AAC6 we were able to make use of Illumina's higher throughput, with up to 30 million raw reads of length 600. Note that the very long length of individual reads in the PacBio platform allows a large improvement in read quality scores by circular consensus sequencing (see also below).

AAC6

Libraries were prepared by two consecutive PCR reactions, adding the adapters for Illumina sequencing. In order to increase the sample heterogeneity necessary for Illumina sequencing, both forward and reverse primers are of staggered length, as a result of a small random base section of variable length (N4-N10). In later sequencing procedures the gene is incorporated both in forward and in reverse direction, to increase sample heterogeneity even more. Sequencing was performed using Illumina MiSeq with a paired-end 300 kit at the Dana-Farber Molecular Biology Core Facilities.

PSE1

Miniprep plasmid DNA from selection libraries was digested with AgeI and AvrII and the fragment at the size of the PSE1 gene (approximately 800 bp) gel-purified. Samples for PacBio Sequel SMRT sequencing were prepared according to the manufacturer's protocol for end-repair and ligation of multiplex barcode adapters, DNA damage repair, exonuclease digestion, and purification of SMRTbell templates. Annealing of primer and polymerase binding to templates, and PacBio Sequel SMRT sequencing, was performed by the Genomics Core Facility at the Icahn Institute, Mt. Sinai; or by the MIT BioMicro Center with assistance from the Dana-Farber Molecular Biology Core Facilities.

Sequence Analysis

AAC6

The resulting forward and reverse read fastq files were stitched together using the FLASH program (Magoč and Salzberg, 2011). After this, sequences were quality filtered: first, the minimum Q score of each base has to be at least Q15, and second, the compound Q score, which is an expression for the overall quality of a read, $Q_{\text{comp}} = -10 \cdot \log_{10}\left(1 - \prod_{i}^{L}\left(1 - 10^{-\frac{Q_i}{10}}\right)\right)$, where i is the sequence position and L the sequence length, is required to be at least 10, implying that for each read there is a 90% probability it contains no read errors. After this, sequences are translated and only full-length sequences are retained.

PSE1

Barcode demultiplexing and generation of fastq files of circular consensus sequences were performed by the Icahn Institute Genomics Core Facility at Mount Sinai, or by the MIT BioMicro Center. Reads from fastq files were first filtered for containing upstream and downstream nucleotides corresponding to those used in the mutagenesis primers which anneal outside the PSE1 gene, then filtered for a minimum Q score of at least 30. Resulting sequences were then translated; only full-length sequences of 271 amino acids are retained, which are then truncated to the length of that in the crystal structure 1G68, 266 amino acids (removing two N-terminal and three C-terminal amino acids).

Alignment Conditioning

AAC6

Since we observed that the selected pools contain a small fraction of contaminant sequences that correspond to a homolog of *aac6* that has been used in the same laboratory (not uncommon in long-term evolution experiments with antibiotic resistance genes), we applied a filtering procedure that effectively removes all sequences that contain stretches larger than three amino acids that are more similar to the contaminant than to the gene under consideration (MATLAB script in Supporting Information). Eventually this procedure only removes a very small fraction of sequences (typically 0.1%–0.2%).

PSE1

The selected libraries contain a small fraction of contaminant sequences corresponding to those of β-lactamases used in the laboratory. We applied a filtering procedure that removes all sequences containing more than six adjacent amino acid mutations from the ancestor PSE1 sequence. This procedure removes a very small fraction of sequences (approximately 0.1%).

AAC6 and PSE1

We subsampled the alignment to only contain sequences with more than the mean number of mutations. This is done in order to remove lowly mutated sequences that would cause the standard filtering step in EVcouplings (see immediately below) to downweight informative sequences. Note that removing sequences at the lower end of the mutational distribution only removes a small fraction of the total number of mutations from the pool. Moreover, as the number of pairs of mutations in a sequence is drastically higher for more highly mutated sequences, this procedure removes a negligible fraction of pair information from the library.

EVcouplings Analysis

EVcouplings analysis (evolutionary couplings) is performed according to (Marks et al., 2011; Sheridan et al., 2015) (implementation on https://www.evcouplings.org, source code and sample config file available at https://github.com/debbiemarkslab/EVcouplings). The inputs are the conditioned alignments as described above and a small number of parameters, mainly regarding filtering, weighting of sequences, and regularization for the calculation of fields and coupling terms (Marks et al., 2011). In the current study, parameters remained unchanged from their default values used in the application of EVcouplings to alignments of natural homologous proteins (Marks et al., 2011), except for a single parameter, θ, that specifies the stringency at which similar sequences (with pairwise sequence identity $\geq \theta$) are down-weighted. The downweighting prevents spurious evolutionary signals due to uneven clusters of very similar sequences in the alignment (Marks et al., 2011). As the sequence-space distribution of natural and laboratory-generated alignments is different, the optimal values for the parameter θ typically differ for these two cases (see below for the numerical choices of θ used here).

AAC6

Results are fairly robust to choices of θ around the mean pairwise identity of the analyzed populations of sequences (Figure S2B). The optimum, with respect to recovering contacts also seen in crystal structure, is slightly above this value. Too low θ will reduce the effective number of sequences in the alignment, at which point the inference of interactions is of lower quality.

PSE1

The results depend slightly more on the choice of θ than they do for AAC6 (Figure S2A). The optimal value, with respect to agreement with residue contacts in the crystal structure, is above that of the mean pairwise identity of the analyzed sequences and slightly below the mean fractional sequence identity to wild-type PSE1.

Compute Structures from Inferred Interactions

A set of computed structures is generated using the distance geometry and simulated annealing protocol in the Crystallography and NMR System package (CNS) (Brunger, 2007), using up to 1.5 × L top-ranked inferred interactions as distance constraints, following the detailed procedure in ref (Marks et al., 2011).

Filtering Inferred Contacts for Geometric Violations

One reason we can compute well-folded structures at the level of agreement with contacts in related crystal structures as low as 50% is that a set of interactions in a folded protein structure has to be mutually consistent, given the connectivity of the polypeptide chain. We make use of the consistency requirement by an iterative algorithm that removes residue interactions that are not satisfied in a subset of folded structures.

Since a fraction of inferred interactions are incompatible with 3D protein fold, which in general decreases the fold accuracy, we developed an iterative filtering approach using additional constraints from 3D geometry to remove incompatible inferred interactions ('geometric violations'). This both improves the correspondence between inferred interactions and X-ray structure contacts (Figures 5A and S3), and also improves the computation of 3D structure using CNS (Figures 5B and 6), which can be frustrated by a too large number of incorrect constraints (Marks et al., 2011). The procedure: (i) A set of computed structures is generated with CNS, using up to 1.5 × L top-ranked inferred residue interactions as distance constraints, plus constraints from secondary structure prediction (Jones, 1999). (ii) Inferred interactions not present in at least 3%–5% of the thus computed structures are considered geometric violations and masked out from the set of inferred interactions. (iii) Steps i and ii are repeated 5-6 times. The procedure is robust to choice of parameters in these ranges. The structures presented in Figure 5B are those with the lowest Cα positional RMSD over more than 90% of residues and which do not contain knots (Kolesov et al., 2007; Virnau et al., 2006).

Dimer Interactions in AAC6

The statistical approach used in EVcouplings is able to identify coevolving amino acid residues agnostic to whether these residues involve two positions within a single protein, or two positions between separate proteins, i.e., protein-protein interactions. For the obligate dimer AAC6 we obtain *inter*-monomer inferred interactions with roughly equal strengths as *intra*-monomer inferred interactions. This also means that in order to compute an *ab-initio* structure, the need arises to disentangle inferred monomer and dimer interactions. Solving this problem is not trivial and is left for future work. In the current manuscript, inferred interactions are identified for both monomer and dimer contacts in AAC6 (Figure 4), but the computed folded structure only considers a single monomer without the dimerization tail (residues 135-148).

Structural Variability

To examine structural variability among the computed models, we first clustered the set of models computed using the inferred contacts from the final iteration of filtering for geometric violations (MaxCluster; (Herbert and Sternberg, 2008)). To remove outliers, only those models in the largest cluster were subsequently analyzed (516 of 690 models for PSE1 and 717 of 720 for AAC6). Both of these clusters contained the highest TM-score computed structure. Structural variability is computed at each residue position as the root-mean-square deviation (RMSD) of Cα-Cα distances for all pairs of superposed models; models were superposed using the MATLAB pdbsuperpose function. Structural variability is visualized in "sausage" representation (Figure 6B) by changing the B-factors in the computed structure PDB file to the RMSD values using the loadBfacts.py pymol script (Gatti-Lafranconi, 2014).

Visualization of Sequence Space

To visualize sequence space, we used the DeepSequence software package (Riesselman et al., 2018) to generate a nonlinear latent-variable model of natural homologs, and then projected all sequences (experimentally evolved and natural) into 2-dimensional latent

space (z_1 and z_2). For both proteins, an alignment was generated with a jackhmmer search (Potter et al., 2018) by EVcouplings using a bitscore of 0.4. These alignments were conditioned by removing sequences that were more than 15% gapped and filtered to only retain sequences with less than 95% pairwise identity. A random subsampling of 5000 natural sequences from these alignments was used to generate the model. Parameters for model training:

batch_size = 100
num_updates = 300000
encoder_architecture = [1500,1500]
decoder_architecture = [100,500]
n_latent = 2
n_patterns = 4

Taxonomy ids for individual sequences were downloaded from uniprot and uniref. The NCBITaxa module in the ETE Toolkit python package (Huerta-Cepas et al., 2016) was used to translate taxonomy ids into class labels.

DATA AND CODE AVAILABILITY

Code and scripts are available on GitHub: https://github.com/sanderlab/3Dseq. The accession number for the sequencing reads (FASTQ files) is [Sequence Read Archive]: [BioProject PRJNA578762]. Alignments, model files, including 3D all-atom structures, and the all-versus-all positional distances are linked to on Github: https://github.com/sanderlab/3Dseq; backup: http://www.3dseq.org