Protein structure without structure determination: direct coupling analysis based on in vitro evolution

Marco Fantini[1], Simonetta Lisi[1], Paolo De Los Rios[2], Antonino Cattaneo[1,3*], Annalisa Pastore[1,4,5*]

[1]Scuola normale superiore (SNS), Pisa, Italy
[2]École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3]European Brain Research Institute, Roma, Italy
[4]Maurice Wohl Institute, King's College London, London, UK
[5]The Francis Crick Institute, London, UK

*Co-corresponding authors

Keywords
β-lactamase, beta lactamase, Amp^R, DCA, direct coupling analysis, evolutionary couplings, sequel, pacbio, 3rd generation sequencing, SMRT sequencing, mutagenesis, error prone PCR, molecular evolution.

Running title
In vitro evolution to obtain structural information

Abstract

Direct Coupling Analysis (DCA) is a powerful technique that enables to extract structural information of proteins belonging to large protein families exclusively by in silico analysis. This method is however limited by sequence availability and various biases. Here, we propose a method that exploits molecular evolution to circumvent these limitations: instead of relying on existing protein families, we used in vitro mutagenesis of TEM-1 beta lactamase combined with in vivo functional selection to generate the sequence data necessary for evolutionary analysis. We could reconstruct by this strategy, which we called CAMELS (**C**oupling **A**nalysis by **M**olecular **E**volution **L**ibrary **S**equencing), the lactamase fold exclusively from sequence data. Through generating and sequencing large libraries of variants, we can deal with any protein, ancient or recent, from any species, having the only constraint of setting up a functional phenotypic selection of the protein. This method allows us to obtain protein structures without solving the structure experimentally.

Introduction

Deleterious mutations can damage the fold and the function of proteins. These mutations are usually rescued, in the course of evolution, by compensatory mutations at spatially close sites that restore contacts and thus preserve structure and function. This creates a correlation between protein contacts and the mutational space of the residues involved, that can be compared to shackles. These shackles, that are called evolutionary couplings, can be observed by looking at the covariation between positions in a multiple sequence alignment. Through them, it is possible to predict the network of contacts that determine protein fold. Recently, direct coupling analysis (DCA) and other techniques based on the interpretation of evolutionary couplings have emerged as a powerful novel methodology that enables to predict protein architecture, fold and interactions[1–7]. These techniques have immensely increased the arsenal of tools at the scientists' disposal to obtain structural information[8–10]. The use of web servers able to apply these methods (EVfold[1,11] evfold.com, GREMLIN[4,5] gremlin.bakerlab.org and MetaPSICOV[12,13] bioinf.cs.ucl.ac.uk/MataPSICOV) has quickly spread throughout the scientific community since these servers provide a inexpensive and fast way to obtain structural information only using sequence information. One of the several advantages of an evolution-based approach is also the possibility to obtain structural information of proteins notably difficult to crystalize and/or model, such as membrane[14] or disordered proteins[11]. DCA has for instance been successfully applied at the proteome scale obtaining milestone results and has led, for instance, to the successful prediction of all the binary protein interactions in E. coli[15] and the retrieval of the structures of entire protein families and subfamilies present in the PFAM database[16].

However, despite being powerful, DCA has severe limitations that restrict its general applicability: public sequence databases such as UniProt are enriched with sequences from model or common organisms. This makes it difficult to avoid a skewed data representation or phylogenetic biases that can be only partially corrected (for instance by average product correction)[17]. DCA requires very large protein families to yield meaningful results. Protein evolutionary age is therefore another typical limiting factor, since older protein families have higher chances to be represented in the majority of the species and had more time to generate orthologs by duplication. Ancient bacterial proteins are thus overrepresented whilst eukaryotic or species-specific protein sequences are usually scarce. Finally, only family-level information is achievable since phylogeny-based algorithms are fed with homologous sequences found by data mining. This makes it impossible to focus on a specific protein: we have to rely only on large protein families.

To overcome these limitations, we developed a general methodology based on experimental techniques coupled to computational analysis. Our method goes all the way from an original ancestor gene sequence, to the generation and collection of sequences, to data analysis using molecular evolution. We generated a large library of variants of a target gene, followed by in vivo phenotypic selection to isolate functional variants of the ancestor protein. The plasmid library carrying the mutants was then sequenced and analyzed by DCA. By this method, we were able to demonstrate that we can retrieve evolutionary constraints and reconstruct protein fold. By substituting natural with in vitro evolution, we explored a brand new application of the DCA concept which overcomes the limitations that have so far hindered the generality and

scalability of the method. As a proof of concept, we chose TEM-1 beta-lactamase, a member of the Beta lactamase family of enzymes that confer to bacteria the ability to destroy the beta lactam ring of penicillin[18] and derivatives such as ampicillin. Resistance allows bacteria to grow in the presence of these antibiotics, a function that is easily amenable to a phenotypic selective pressure. Our data clearly demonstrate the feasibility of our approach to retrieve structural information on the target protein in the absence of structural data and provide a new tool to all branching fields of evolutionary coupling research.

Results

Experimental design

A schematic representation of the pipeline from the ancestor gene to the construction of the libraries and the retrieval of contact predictions is shown in **Figure 1**. We employed random mutagenesis from error prone PCR[19] to generate a large library of variants of the target gene, followed by transformation into bacterial cells and *in vivo* phenotypic selection to isolate functional variants of the ancestor protein. The plasmid library carrying the mutants was then collected from the surviving bacteria and subjected to Pacific Bioscience' single molecule real time (SMRT) sequencing[20]. We used the TEM-1 beta-lactamase of the pUC19 plasmid[21]. TEM lactamases are encoded by fairly long genes (~900 bp) and are present in several natural variants[22]. Their structure consists in a three-layer ($\alpha\beta\alpha$) sandwich. As a reference for the mutational landscape, a collection of beta lactamase sequences (named "UniProt" dataset) was obtained from the UniProt database. To obtain a heavily mutagenized beta lactamase without damaging the survival rate, the library was subjected to consecutive cycles of mutations, selection and amplification through the use of error prone PCR[19] and growth in selective semisolid media[23,24]. We will refer hereafter to the library at the end of each cycle as a generation of molecular evolution. In total, we performed twelve generations. The first, fifth and twelfth generations were sequenced with the Pacific Bioscience (PacBio) Sequel platform and analyzed.

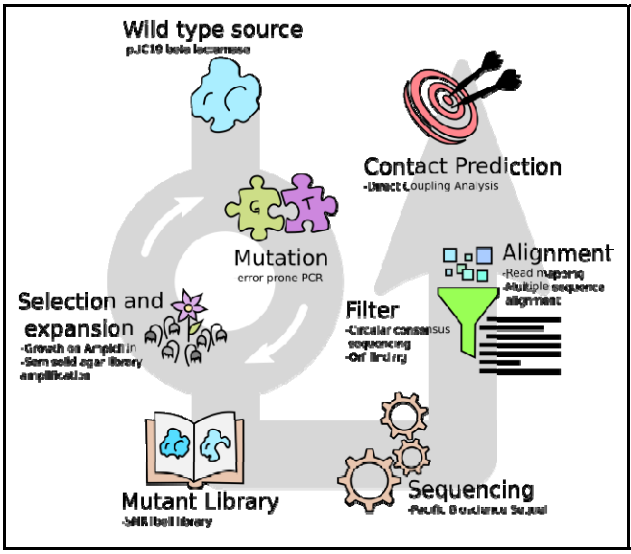

Figure 1: Schematic representation of the pipeline

Molecular evolution libraries mimic natural variability

Random mutagenesis is able to quickly and cheaply create a heavily mutagenized library. We first thought to adopt the protocol suggested by Rollins[25] and Schmiedel[26] which uses complete combinatorial two-residues deep mutational scanning to create a 50 amino acids receptor domain library[27]. However, this approach would have been prohibitively expensive for common proteins since the number of required double mutants grows with the square of protein length. We used instead error prone PCR to drive mutagenesis. To verify the progress of molecular evolution and maintain libraries with a fair amount of complexity, we controlled three parameters throughout the twelve generations: the number of transformants in the bacterial growth, the number of mismatching amino acids in a small sample of clones and the information entropy a each amino acid position.

Since each bacterial colony in the selection medium expresses a single functional variant of the protein, the number of transformants poses the theoretical ceiling to the library diversity. We kept the number of transformants at least in the same order of magnitude of the sequencing capacity of the next generation (NGS) platform (between 100 thousand and 1 million) to guarantee a good library complexity. In the last few generations we raised this limit to 400 thousand clones to increase the probability to sequence unique variants. After each generation a small sample of clones underwent sequencing to retrieve an estimation of the number of mismatching nucleobases and amino acids with respect to the ancestor sequence (**Figure 2A**).

To complement this information, the same parameter was estimated from the sequencing results of the three sequenced generations. The distribution of the number of mismatches per sequence fitted the theoretical Poissonian model expected for a mutagenesis (gen1: λ=5.12 s=0.0054; gen5: λ=12.54 s=0.0084; gen5: λ=26.9 s=0.0159) (**Figure 2B**). The median number of mutated residues observed when the colonies were picked matched perfectly that obtained from next generation sequencing (**Figure 2A**) and what was expected from a Poissonian model, proving that the handful of colonies picked provided a good representation of the mutations present in the library. We could conclude from the observed steady increase in the number of mutations throughout the molecular evolution that the final mutation rate of the evolved protein library can be regulated by increasing the number of generations. Sequencing data also allowed us to calculate the mutation rate per amino acidic position, defined as the frequency of the observed mismatching amino acids compared to the original pUC19 beta lactamase sequence. After 12 generations of molecular evolution we started to observe several instances of genetic drifts, where a mutation became more common than the original residue at a given position (**Suppl Figure S1**). This phenomenon makes the mutation rates less informative, since they involve a comparison to the original residue that is now a minority. To circumvent the problem, we measured the Shannon information entropy of each residue, obtaining an approximation of the impact of the mutagenesis for each position, without the need of a reference sequence (**Figure 2C, Suppl Figure S2**). The proportion of mutants and the information entropy of each residue were strongly correlated one to each other and with those observed from the UniProt dataset (mutant frequency: rho 0.624, $p < 1e\text{-}15$; entropy: rho 0.632, $p < 1e\text{-}15$) (**Suppl Figure S3**). We also observed that both the entropy and the mutation frequency compared to the reference pUC19 TEM-1 lactamase at each position of the molecular evolution libraries are almost always lower than the corresponding evolutionary one (**Figure 2D, Suppl Figure S3**),

supporting the idea that the latter poses a limit to which a molecular evolution library would tend, given enough mutagenesis rounds.

The observed molecular evolution took the same pathways of natural evolution, proving that the former could effectively substitute the latter as a source of genetic variants. The cycles of mutations and selections determine the population bottlenecks that alter the genetic variability of the system. This will inevitably affect the genetic variability of the population because the drastic reduction in the population size after catastrophic events will establish a founder effect and because in a smaller community genetic drift and fixation are far more common.

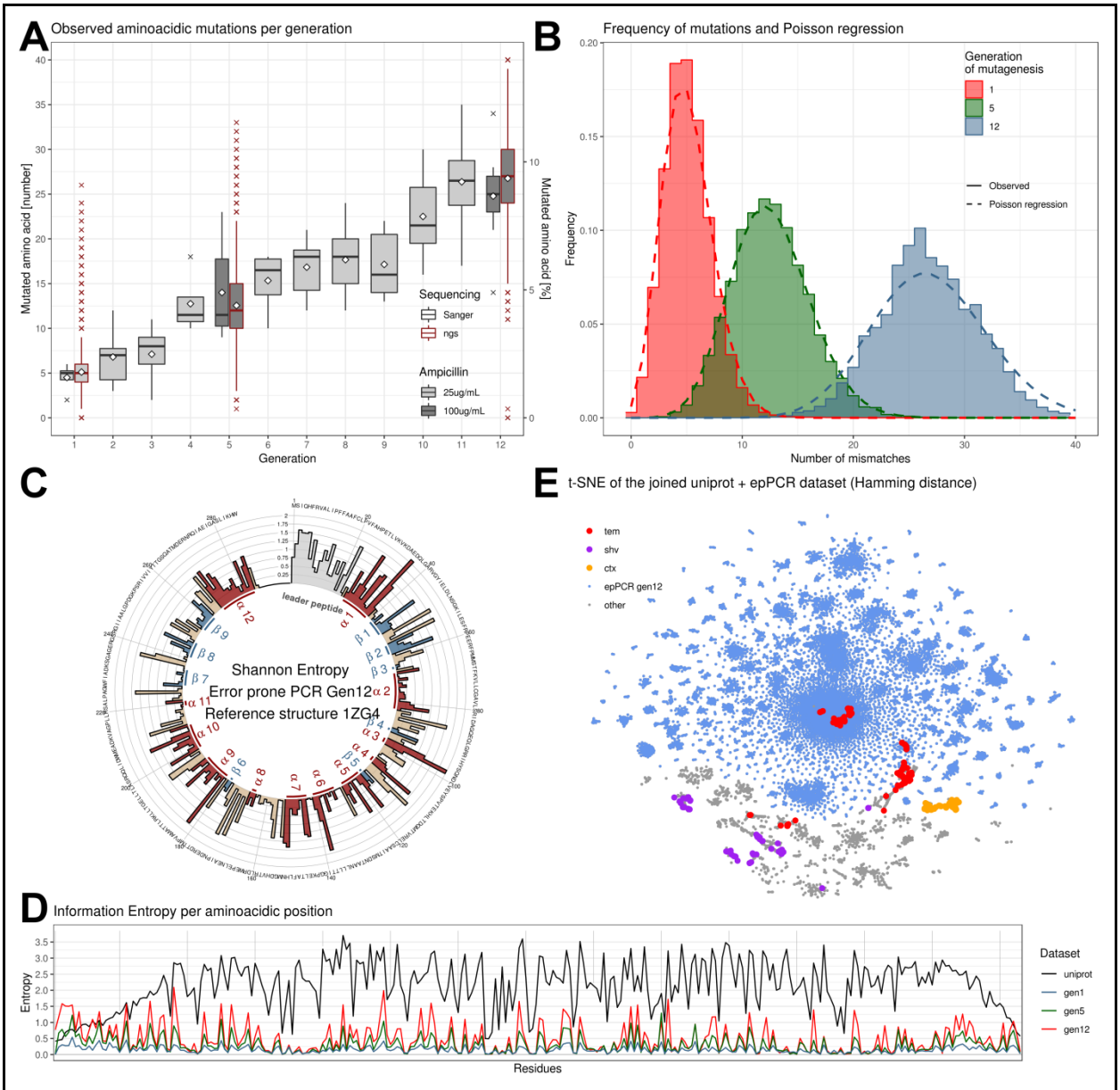

Figure 2: Sequencing and molecular evolution results

A) Boxplot showing the number of amino acid mutations (mismatches) observed in the sample of clones sequenced after each generation (Sanger sequencing, black border)

and after NGS (red border). The white diamond dots indicate the mean.

B) Frequency distribution of the number of aminoacidic mutations observed in the sequenced libraries (solid lines) and their respective Poissonian regressions (dotted lines: gen1: λ=5.12; gen5: λ=12.54; gen5: λ=26.9).

C) Shannon information entropy (H) per residue position of the sequenced 12th generation library. The colors and annotations follow the secondary structure classification present in the PDB structure 1ZG4 (red: alpha helices, blue: beta sheets, tan: coils). The leader peptide sequence (light gray) is missing in the structure.

D) Comparison of the Shannon information entropy between the UniProt and the in vitro evolved datasets.

E) t-SNE dimensionality reduction applied to the joined UniProt / error prone PCR 12th generation library dataset. Hamming distance between sequences was used as distance metric. Gray and cyan represent the original dataset (gray UniProt, cyan epPCR library). Overlaid on top, the UniProt sequence membership to one of the three main families of type A beta lactamases retrieved from the corresponding UniProt annotation are displayed in bright colors. The original pUC19 beta lactamase before molecular evolution is classified as a TEM beta lactamase (red).

Single molecule sequencing overcomes library restrictions.

We used the PacBio single molecule real time (SMRT)[20] sequencing platform (Sequel) that can obtain up to a million readings per sequencing cell[28] and is compatible with the complexity of a molecular evolution library. The total number of transformants for the three sequenced libraries, that pose a limit to the library complexity, were 200K, 260K and 400K colony forming units (CFUs), respectively, while each sequencing run generated 192K, 289K and 157K raw readings after quality filtering. The sequenced DNA fragment was over 800 base pairs. Other more common next generation sequencing platform like Illumina HiSeq or MiSeq are instead characterized by decreasing quality with increasing base position[29] and thus cannot sequence more than few hundreds base pairs. The number of sequencing cells, the amount of grown bacteria that undergo the selection process and the fragment size were all designed several times under the limits of the employed techniques and could be easily scaled up to meet the needs of any target protein. Our mutational library is the first molecular evolution library sequenced in a third generation sequencer, thus guaranteeing a high volume of high quality single molecule data. This library is also one of the most mutated molecularly evolved TEM beta lactamase libraries ever produced, where its elements diverge from the ancestral protein for around 1/10 of their original amino acidic composition.

The mutational landscape of the evolved library reflects the structural features of TEM beta lactamases

The beta lactamase structure 1ZG4[30] from PDB (https://www.rcsb.org) was used as a reference structure to assess the contact prediction and the accuracy of the prediction analysis. TEM1 beta lactamase is a globular protein with a roughly ellipsoidal shape (**Figure 3A**)[31]. It can be divided into two subdomains, one composed of a five stranded beta sheet plus the N-terminal and the 2 last C-terminal helices, the second is a big helical subdomain located on the other side of the beta sheet. The protein contains a large hydrophobic core between the beta sheet

and the helical subdomain, and a second hydrophobic region in the core of the helical domain. The innermost helix of this domain, H2, contains both structural and catalytic residues. The PDB structure lacks the first 23 amino acids, corresponding to the leader sequence for secretion, which is cleaved during protein maturation to allow protein release.

The profiles of the mutation rate and entropy per residue observed in our molecular evolution libraries is conserved and increases across generations, in line with what is observed in the UniProt dataset of the naturally evolved beta lactamase family (**Figure 2D**). This profile reflects the different mutation propensities of the various residues as well as the interactions with the solvent and the polarity of the local environment. We observed a high degree of conservation in the presence of bulky nonpolar residues like tryptophans and methionines and in cysteines involved in the sulfur bridge, whilst small residues show in general an increased variability (**Suppl Figure S4**). A periodic alternating pattern of high to low entropy can be seen in the long alpha helices H1, H9, H10 and H12. This reflects the nature of the two halves of the helices, one being hydrophilic partially exposed to the solvent, the other containing hydrophobic residues packed against the protein core.

H2 is different from the other helices since it is located deeply inside the hydrophobic core of the protein and mediates most of the hydrophobic interactions of the protein. This parallels the lower mutation frequency and entropy observed in all our libraries (**Figure 2C, Suppl Figures S1 and S2**), since mutations in the hydrophobic core have a high chance to damage the fold and thus impair the function of the protein.

Noteworthy is also the correlation (spearman correlation: rho 0.53, $p < 1e\text{-}15$) observed between the mean crystallographic B factor of residues in the reference structure and the information entropy retrieved from the evolved library (**Suppl FIG S5**). This correlation likely reflects the tendency of residues that are part of ordered structures to be averse to mutation.

While the mutational landscape of TEM-1 beta lactamase covers a broad range of substitutions, four mutations in particular became by genetic drift more frequent than the original sequence in the last generation of molecular evolution: M180T, E195D, L196I, S281T (**Suppl FIG S1**). Among these M180T, that corresponds to M182T in the standard numbering scheme of class A beta lactamases (ABL)[32], is a well-documented mutation known to contribute to the protein stability and found very commonly both in natural variants[33,34] and in mutagenesis experiments[35]. E195D and L196I (E197D and L198I in ABL) are mutations in the H8/H9 turn which are commonly found during mutagenesis[36]. Significantly, D197 is the consensus amino acid for this position (197) in the original alignment of class A beta lactamase[32].

We next used principal component analysis (PCA) on the Shannon entropies associated to every position of each dataset, to evaluate the evolution of the mutagenized libraries towards the natural diversity (**Suppl Figure S6**). We also applied PCA (**Suppl Figure S7**)[37] and t-SNE (**Figure 2E**) on the sequences themselves, to evaluate the degree of dispersion for each generation in comparison to the natural variants. These analyses suggest that the subsequent mutagenesis cycles consistently evolve the sequences in a concerted direction that is similar to that observed in the natural dataset. t-SNE also proved that the cluster of the evolved lactamase is only an extension of the TEM family and do not represents other members of class A beta lactamase (**Figure 2E**). Thus the molecular evolution libraries describe the mutational space of a specific protein and not of a protein family.

From this analysis we may conclude that the library has retained most of the characteristics of a collection of natural beta lactamase variants and represents only the mutational landscape around the protein of interest. This means that the library provides a snapshot of the early stages of the evolution of a protein, neither too similar nor too diverse from the original version, but exploring the landscape of mutational substitutions in a direction dictated by natural selection.

The predominant Direct Coupling Analysis predictions are short range interactions where the co-evolution effect is stronger

After the several generations, we extracted the longest open reading frame from each of the 150.000 circular consensus reads obtained after sequencing the last generation of mutagenesis and removed those shorter than the wild type protein. We built a multiple sequence alignment (MSA) from the remaining 106487 (68.9%) translated peptides and kept only the original 286 positions related to the wild type enzyme. To predict which residue pairs interact, we applied a custom implementation of DCA that applies a pseudo-likelihood approximation[38] to this MSA as well as to the MSA obtained similarly from the other two sequenced generations of mutagenesis (see Materials and Methods).

We retained the 286 residue pairs (0.72% of the total possible contacts) which showed the highest DCA score and were more than five residue apart in the MSA and compared them to the contact map of the reference structure (**Figure 3B**). Despite removal of the predicted contacts where the position pairs were too close to each other in the MSA, we could observe an enrichment of proximal interactions (near the diagonal of the contact map), at the expense of long range contacts. This is a good indicator that the analysis is extracting co-evolving positions since proximal residues using standard evolutionary dataset are typically associated with a high DCA score. In general, the predicted contact distribution was non-random and contacts tended to crowd at both extremities of the helices ignoring highly conserved areas like H2 (residues 67-83). Other minor crowding could be observed around two big looping regions (90-100 and 160-170). N-terminal crowding is likely due to the degeneration and duplication around the starting site that was already observed during Sanger sequencing, while the C-terminal density is probably due to sequential mutated positions in sequences where a C-terminal frameshift creates a block of strongly correlated positions without significantly affecting the functionality of the protein. The propensity to avoid conserved areas like H2 instead reflects the difficulty in creating a robust prediction when observing an inadequate number of mutations. We thus face an interesting problem: the more contacts a residue mediates the more harmful a mutation becomes and thus we will observe a limited set of variations. However, since the mutational space at each position is tightly linked to the prediction power, the contacts formed by the most important residues are also the ones harder to predict.

Improving the prediction power in key areas and retrieval of long range interactions

To improve the accuracy and the spread of the predictions we applied a correlation-based approach identical to that proposed for fitness by Schmiedel and Lehner[26]. Residues in structural proximity are often deeply interconnected and likely to share the same environment, consequently producing similar interaction patterns with other positions. Exploiting this similarity,

we could obtain interactions from conserved positions by calculating partial correlation of the protein positions on the DCA patterns, because highly interconnected positions will have a characteristic association pattern across the protein easily recognizable by partial correlation, even if the original DCA predictions are fairly inaccurate.

To validate this approach, we calculated the partial correlation with the UniProt dataset (**Suppl Figure S8**). As expected, the predictions from partial correlation are very similar to the coupling score obtained by DCA (**Figure 3C**), and are in general less densely packed around the diagonal albeit showing a few more incorrect predictions.

The partial correlation approach applied to the molecular evolution dataset gave instead very different results compared to the original coupling score (**Figures 3B,D**) and resulted more similar to what observed in the UniProt dataset, where the predicted interactions were more broadly distributed and both the terminals and the diagonal were far less crowded. The accuracy of the prediction was relatively low (**Figure 3E**), even if several times bigger than the random expectation. However, this inaccuracy was caused by low precision and not by a low trueness to the underlying values as proven by the low value of shortest path from a true contact observed for the predicted pairs (**Figure 3F**). Along the contact map diagonal we observed densities in correspondence to strong secondary structure interactions, like the proximity between N terminal sheets and the H1 helix represented in the graph with the cluster of points found near residues 25 to 60. Other subdiagonal crowding (around residues 90-170) could be observed in the helical domain in correspondence to the interactions formed by the bending of the peptide chain in a turn. These interactions and similar ones, formed between the C-terminal half of the five stranded sheet and the C-terminal helices of the protein (200-285), could also be seen in the original DCA score (**Figure 3B**) and were the most evident areas along the diagonal of the UniProt dataset where the predicted interactions clustered (**Figure 3C**). Long range contacts, represented in the contact map by data points in regions far from the diagonal, were significantly different if we evaluated the interactions obtained by partial correlation and those predicted by the original DCA score. Partial correlation prediction showed several off-diagonal prediction points, mainly associated with highly interconnected regions or between elements of the hydrophobic core. In particular, we could observe several interactions of the H2 helix (67-83) with other elements of the helical domain (H2 to H10, residues 65-210), demonstrating the centrality of the helix, even though the region *per se* is characterized by a very small mutational landscape (**Figure 2C**). The analysis identified also another cluster in the helical domain describing the proximity of helix H10 (199-210) to helix H5 (117-126).

Overall, we were able to obtain a predicted contact map that matches effectively the contact map of the reference crystal structure without any prior structural information. Our analysis demonstrated the possibility to obtain evolutionary couplings from a collection of sequences evolved in vitro. DCA highlighted the strongest evolutionary signal of proximal interactions (around the diagonal of the contact map) while partial correlation extracted information on the centrality of the H2 helix and the relations between secondary structure elements. These results demonstrate that molecular evolution can be used as an easy and powerful tool for structural analysis, by compressing the millions of years of natural selection into the couple of months of in vitro mutagenesis and selection.

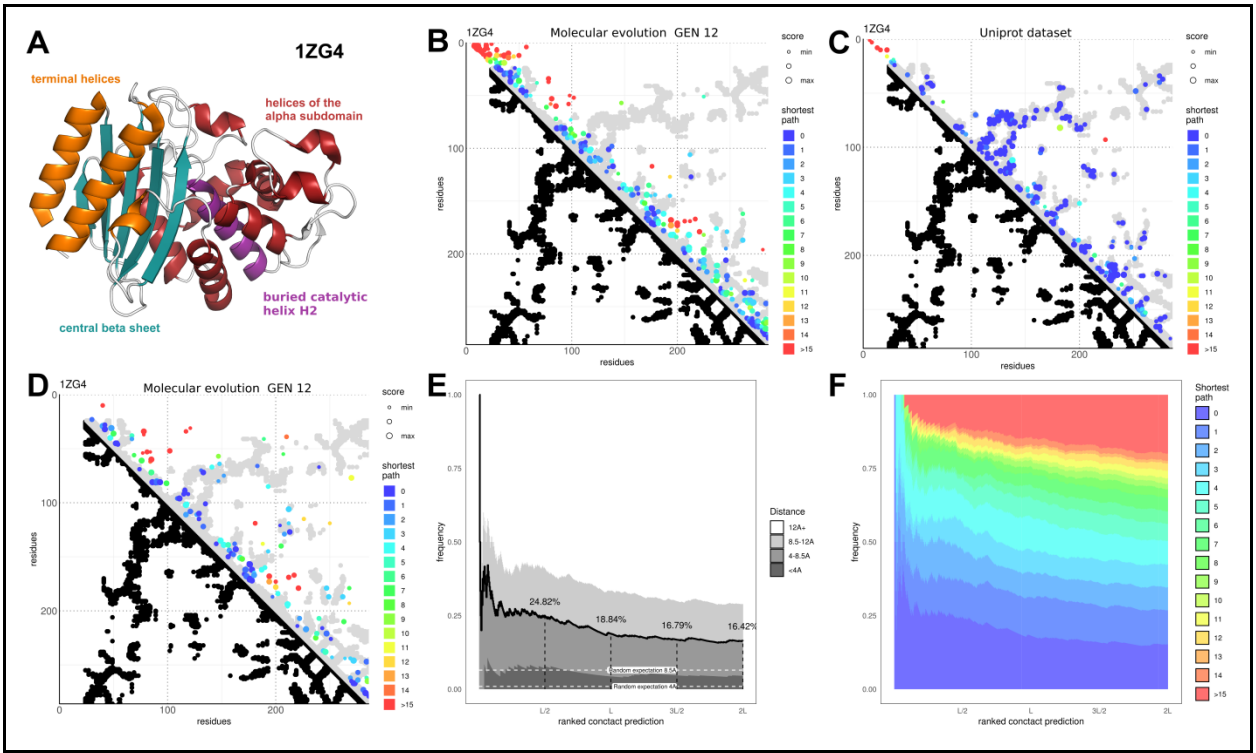

Figure 3:

A) Experimental structure and main features of TEM-1 beta lactamase (derived from PDB 1ZG4). The N- and C-terminal helices (orange) form a subdomain with the five stranded central beta sheet (blue) linked to the helical subdomain by two small hinge regions (red) on the opposite side of the sheet. The catalytic pocket resides at the interface between the beta sheet and the helical domain. Helix H2 (purple) is the innermost helix of the helical domain and comprises both a catalytic and a structural function.

B) DCA plot showing the top L (L = 286, the length of the protein amino acid chain) contact predictions by DCA obtained from the 12th generation of molecular evolution. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the gray mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment are excluded from this ranking to promote visualization of long range interactions. In the graph, the dot size indicates the ranking of the prediction score while the color indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure).

C) Plot of the top L DCA predictions of the UniProt dataset.

D) Plot of the top L/2 partial correlations of residue positions on DCA score obtained from

the 12th generation of molecular evolution.

E) Partial correlations between positions of the DCA score at the 12th generation of molecular evolution sorted by their value. From the top: the first x elements were extracted, x increasing along the X axis. The graph represents the fraction of residue pairs that have atoms less than 4, 8.5 or 12 Å apart in the reference structure. Dotted white lines represent the frequency of position pairs expected to be under 4 and 8.5 Å apart in a random sampling (random expectation).

F) Partial correlations at the 12th generation of molecular evolution. The correlations were sorted and the top elements extracted as described in panel E. The graph represents the frequency of shortest path distances from a true contact (non-hydrogen atoms that are less than 8.5 Å apart in the reference structure) observed in the contact pairs of these samples.

Discussion

Despite the well acclaimed success that structural biology is currently undergoing, experimental determination of macromolecular structures, with particular regard to membrane proteins[39,40] and protein-protein complexes[41], remains a costly and time consuming endeavor. The possibility to gather structural information without the need of experimental structural determination represents therefore an important goal.

The study of evolutionary couplings is an emerging frontier for structural biology, able to retrieve the network of interactions that dictate protein fold and function[1–7]. The innovation brought by the technique is the ability to produce structural information without the need of experimental structure determination, relying only on the traces left by the evolution of the protein sequence. The correlations are obtained from the continuous polishing process that the flow of time exerts on sequence to optimize/retain function. This makes any structural information retrieved by the analysis like a fossil imprint of an *in vivo* interaction.

However, the current computational techniques based on evolutionary couplings require thousands of sequences to provide statistically meaningful results[2,3,42]. Thus, current evolutionary coupling methods are limited to ancient and universal protein families, for which sequence data are available across a huge variety of species. This is a major limitation: a large number of human proteins, for instance, do not have ancient phylogenetic origin[43]. They are therefore not amenable to evolutionary coupling methods based on phylogenetic databases and can only be approached by experimental approaches.

Here, we presented a strategy (CAMELS) that overcomes this limitation and lays the bases to develop a general method to gather structural information on protein contacts without performing experimental structural studies. We provided a unique pipeline from the molecular to the computational level using all the most advanced techniques, and have solved a number of crucial technical problems, along the way.

Our CAMELS method is based on molecular evolution and on the power of phenotypic selection. We produced one of the largest and most diversified molecular evolution libraries that shows high single molecule sequencing quality and sequence divergence from the original

ancestral protein of nearly 10% (i.e. 25 amino acid mutations). It is also the first to have been sequenced at the single molecule level by third generation NGS.

We used the library to obtain structural information, by creating sequence diversity through mutation and analyzing artificial evolutionary couplings. We showed that the predicted contact map matches successfully the contact map of the reference crystal structure. By generating hundreds of thousands of mutagenic functional variants, CAMELS permits to focus on any protein and builds the foundation for a targeted structural analysis. This may allow to investigate by DCA-like methods evolutionary younger proteins, like eukaryotic-only or vertebrate-only proteins or human proteins of neurobiological interest, ultimately solving species-specific questions that need species-specific answers.

Pilot work on structure prediction from molecular evolution experiments have been published in preprint form during the development of this study[25,26,44]. However, the strategy used to collect their data can only be applied to proteins strictly under 200 amino acids and can thus be used solely on a small fraction of the proteome from all three domains of life[45]. One of the key advantage of CAMELS is the absence of protein length constraints, since both the mutagenesis strategy and the sequencing allow the processing of proteins of any length.

One important property of CAMELS is the requirement for a phenotypic selection, a property that makes the method truly evolutionary. For some proteins (such as TEM1 beta lactamase), a phenotypic selection scheme is readily designed. More generally, selection schemes for the correct folding of any protein can be easily designed. The structure of protein complexes and of protein-protein interaction interfaces is also an area of enormous interest but experimentally difficult and time consuming. CAMELS may be modified to the study of protein-protein surface interactions, exploiting selection schemes for interacting proteins, coupled to SMRT sequencing, which would allow the observation of both paired proteins in a single sequencing read.

The next obvious step will be to exploit standard and generic selection methods that rely on the folding and binding properties of the mutant proteins in the library, regardless of their functional activity. We have, for instance, already envisaged to use two hybrid selection schemes to select for interacting partners, using a strategy we already pioneered in recent papers[46,47] to select more stable antibodies against a given target. Thus, in a next step we could apply CAMELS to two covariant interacting proteins, which could then be co-selected by a two hybrid scheme for preserving their mutual binding. This strategy will, for instance, provide information on the direct or indirect structural determinants for protein-protein interacting domains. This is a revolutionary breakthrough that is not restricted to any specific case.

Another advantage of mutational libraries with respect to the classical phylogenetic data is the representation of a sequence instead of a family, since the Markovian models that retrieve the sequences for the alignments in the standard analysis do not differentiate close paralogs from true orthologs. This poses a serious limitation for protein families rich in paralogs like 7TM receptors or globulins, for which it is nearly impossible to obtain information for a specific member of the family. The ability to represent a protein instead of a family is a new feature that can enable to distinguish a different level of details during the biological interpretation of the data.

Finally, one of the biggest obstacles to an *in vitro* evolution approach was the precarious equilibrium between mutagenic strength and selection survival rate. We solved this issue with a

generational approach. We can further envisage future applications of the method to a continuous evolution in a specialized bioreactor. Overall, the method provides a solid methodology that bypasses the most limiting factors of evolutionary coupling analysis techniques and opens a new page in structural biology and evolution.

Acknowledgements

We thank our colleagues from Scuola Normale Superiore Federico Cremisi and Alessandro Cellerino for valuable comments. We are also grateful to Martina Goracci and Ottavia Vitaloni for times when, thanks to their wisdom and expertise, we were able to make major breakthroughs. We are also indebted to Duccio Malinverni for providing the software implementation of the asymmetric version of the DCA and to Alessandro Viegi for administrative support throughout the project. The project was funded by institutional funds from Scuola Normale Superiore and by a BlueSky grant from University of Pavia.

Materials and methods

Plasmid construction & cloning

The backbone plasmid vector pUC19[21] (ATCC 37254) from ThermoFisher Scientific (SD0061) was modified to add flanking XhoI and NheI restriction sites to the already present Amp^R ORF to be able to easily clone in later steps the mutagenized Amp^R,. To construct the plasmid, both the β-lactamase gene and the complementary plasmid vector fragments were amplified with oligonucleotides carrying the XhoI and NheI restriction sites (XhoI_bla_fw: tgaaaactcgaggaagagtATGAGTATTCA, NheI_bla_rv: acttgggctagctctgacagTTACCAATGC; NheI_backbone_fw: gtcagagctagcccaagtttactcatatat, XhoI_backbone_rv: ctcttcctcgagttttcaatattattgaag). They were then digested with the restriction enzymes and ligated with T4 ligase (Suppl FIG S9). The 5' restriction site was placed just behind the Shine-Dalgarno sequence and the ability to metabolize ampicillin was assessed by growth of *E.coli* carrying the plasmid in selective media. The new plasmid is named pUC19a. (Suppl FIG S10). The Amp^R gene of pUC19 (GenBank: M77789.2) expresses a TEM-1 (class A) β-lactamase whose structure can be viewed in the 1ZG4 PDB entry.

Error prone PCR

Mutagenesis of the Amp^R gene was achieved with error prone PCR[19] in a mutation prone buffer with manganese ions, low magnesium, unbalanced dNTPs concentrations and a low fidelity DNA polymerase. Both low magnesium and the presence of manganese ions affect the efficiency of magnesium ions as cofactors of the polymerase by competition or by sheer low availability, while the unbalanced dNTP concentration favors mutations by scarcity of substrate and the deliberate usage of a low fidelity polymerase further increases the mutation rate. The reaction mix contained Tris-HCl pH 8.3 10 mM, KCl 50 mM, $MgCl_2$ 7 mM, dCTP 1 mM, dTTP 1 mM, dATP 0.2 mM, dGTP 0.2 mM, 5′ primer (bla_mut_fw: tgaaaactcgaggaagagtATG) 2 µM, 3′ primer (bla_mut_rv: acttgggctagctctgacagTTA) 2 µM, template DNA 20 pg/µl, $MnCl_2$ 0.5 mM (added just before reaction starts), Taq G2 DNA polymerase (Promega M784A) 0.05 U/µl

(added just before reaction starts). The error prone PCR was carried out in serial reactions of 4 cycles in 100 µL in the recommended supplier reaction conditions and with an annealing temperature of 62°C. In the first reaction tube, the DNA template was a gel purified XhoI/NheI digested β-lactamase fragment 20 pg/µl, while subsequent reactions were fed with 10 µL of the previous PCR product.

Library construction

The purification and digestion protocols before library construction changed slightly among generations. However, the optimized version of the pipeline employed in the last generations proceeded as follows: gel purify ~80 µL of the PCR reaction mixture underwent a cumulative amount of 20 cycles of error prone PCR, avoiding carrying over other reaction byproducts as much as possible. PCR was performed in standard reaction conditions to amplify the product and guarantee that the two strands of the amplicons did not contain mismatching base pairs. This step helped reducing the ambiguity in base calling during the circular consensus analysis. The purified PCR product was digested with XhoI and NheI restriction enzymes for 3h in CutSmart buffer (NEB). One hour before the end of the reaction, an appropriate amount of calf intestinal phosphatase (CIP) (NEB M0290S) was added following the supplier's instruction. Adding CIP during insert digestion strongly reduced the formation of insert concatemers, guaranteeing a single β-lactamase variant per plasmid. After gel purification to remove the CIP, the ligation between the fragment and the XhoI/NheI digested backbone of pUC19a was performed in a 1:1 insert:vector ratio. Formation of backbone concatemers was expected and unavoidable, but did not hinder the selection efficiency.

Selection

The ligated library was purified and then transformed by electroporation in ElectroMAX DH5α-E competent cells (Invitrogen #11319019). We employed ultralow gelling agarose (SeaPrep, Lonza #50302) 0.3% in Luria Broth (LB) medium with ampicillin 25 µg/mL to grow the bacteria[23,24], obtaining between 0.4 and 3 million surviving colonies per liter. The bacterial growth in the fifth and twelfth generation was performed in LB medium with ampicillin 100 µg/mL to increase the stringency of the selection before the sequencing. After 40 h growth, the bacterial pellet was retrieved by centrifugation 7500RPM at RT and the plasmids extracted by a maxi prep.

Sequencing

Construction of the libraries and sequencing on PacBio Sequel platform were carried out by Arizona Genomics Institute (AGI). After sequencing, the library was processed with the PacBio official analysis software SMRTlink to obtain the circular consensus (using ccs2) of the reads. In this step, the sequences where the consensus was built from less than 10 sequencing polymerase passes or when the predicted accuracy was less than 100 ppm (Phred 40) were filtered out from the dataset. The result was then mapped to the wild type β-lactamase XhoI-NheI digestion fragment of pUC19a with bowtie2[48] to retrieve the coding strand and the start site of the lactamase. After *in silico* translating the dataset, protein collection was further refined

keeping only the elements coding a protein of 286 amino acids (as the wild type) and then aligned using MAFFT (http://mafft.cbrc.jp/alignment/software/)[49] to construct the MSA. The 12th generation had issues with the *in silico* translation step caused by degeneration of the N-terminus as well as the starting site, resulting in a big amount of sequence with a premature termination codon. To circumvent the problem, the longest open reading frame, identified with a custom script, was considered the correct genetic sequence and the translated products were filtered to keep the sequences coding for proteins of at least the wild type length. This procedure was required to remove from the alignment bad quality reads, unrelated sequences and protein variants carrying a frameshift which would generated a strong correlation noise between adjacent amino acid positions.

To compare our data to the natural occurring mutations of TEM beta lactamase, we created a reference dataset by running a small seed of TEM beta lactamases in Hmmer[50] on the UniProt database (https://www.uniprot.org).

Direct Coupling Analysis

The predicted contact pairs were obtained using a custom implementation[51] of the asymmetric version of the DCA[2,7] that applies the Pseudo-likelihood method to infer the parameters of the Potts model[3,38]: $P(X) = \frac{1}{Z} exp[\sum_i^N \quad h_i(Xi) + \sum_{i,j}^{N,N} \quad J_{ij}(X_i, X_j)]$ (1)

where X is a sequence of the MSA and Z is the partition function.

Sequences were reweighed using an identity threshold that reflects the mutation rate of the generation analyzed to remove parental inheritance (intended as “phylogenetic” bias created during mutagenesis) and sampling biases in the MSA. The first generation was too similar to the wild type to apply any sampling correction without unreasonably reducing the number of effectively non-redundant sequences[2]. The fifth generation used a 95% identity threshold and the twelfth generation a 90%. A standard L2 regularization was added following the original regularization described in Ekeberg et al., 2013[3] ($\lambda = 0.01$). The code used the scoring scheme for contacts proposed in Markley et al.[52] where the DCA scores were computed as the Frobenius norm of the local coupling matrices of the Potts model. Dunn et al. average product correction (APC) was subtracted to remove background correlation[17]. The N top scoring contact predictions (N equals the MSA sequence length) were compared with the contact map of the reference structure (1ZG4) constructed considering two residues to be in contact if at least a pair of their respective heavy-atom (non hydrogens) was less than 8.5 Å apart[3]. As it is standard practice, we removed predictions along the diagonal of the contact map if the residue pairs were less than five positions apart to promote enrichment of long-range predictions. We used the shortest-path (SP) distance[53] defined as the L_1 norm in the contact map lattice to join DCA predictions and the closest structural contact to visualize the agreement between predictions and empirical observations.

Partial Correlation

To obtain the partial correlation matrix from the symmetrical DCA score matrix, we first set all the diagonal elements of the matrix to 1 and then approximated the partial correlation between rows with the pcor.shrink function of the corpcor R package[54].

Other bioinformatic tools

Graph generation was performed with R version 3.2.3 (2015-12-10). Poisson regression was performed with the fitdistrplus R package, while PCA was performed with the base R prcomp function. Mutation rates and Shannon information entropies were calculated with custom scripts. t-SNE was performed in Matlab version R2018b using the Hamming distance as metric.

Bibliography

1. Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. **6,** (2011).
2. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **108,** E1293-301 (2011).
3. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **87,** (2013).
4. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 15674–9 (2013).
5. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **2014,** 1–21 (2014).
6. Ovchinnikov, S. *et al.* Protein structure determination using metagenome sequence data. *Science* **355,** 294 LP-298 (2017).
7. Weigt, M., White, R. a, Szurmant, H., Hoch, J. a & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 67–72 (2009).
8. Altschuht, D., Leskx, A. M., Bloomer, A. C. & Klug, A. Correlation of Co-ordinated Amino Acid Substitutions with Function in Viruses Related to Tobacco Mosaic Virus. 693–707 (1987).
9. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins Struct. Funct. Genet.* **18,** 309–317 (1994).
10. Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, a. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271,** 511–523 (1997).
11. Toth-Petroczy, A. *et al.* Structured States of Disordered Proteins from Genomic Sequences. *Cell* **167,** 158–170.e12 (2016).
12. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28,** 184–190 (2012).
13. Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31,** 999–1006 (2015).
14. Hopf, T. A. *et al.* Theory Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **149,** 1607–1621 (2012).
15. Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3,** e03430 (2014).
16. Uguzzoni, G. *et al.* Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E2662–E2671 (2017).

17. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24,** 333–340 (2008).
18. Abraham, E. P. & Chain, E. An enzyme from bacteria able to destroy penicillin [1]. *Nature* **146,** 837 (1940).
19. Wilson, D. S. & Keefe, A. D. in *Current Protocols in Molecular Biology* **51,** 8.3.1-8.3.9 (John Wiley & Sons, Inc., 2001).
20. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323,** 133–138 (2009).
21. Norrander, J., Kempe, T. & Messing, J. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* **26,** 101–106 (1983).
22. Bush, K. Nomenclature of TEM beta-lactamases. *J. Antimicrob. Chemother.* **39,** 1–3 (1997).
23. Fantini, M. *et al.* Assessment of antibody library diversity through next generation sequencing and technical error compensation. *PLoS One* **12,** e0177574 (2017).
24. Elsaesser, R. & Paysan, J. Liquid gel amplification of complex plasmid libraries. *Biotechniques* **37,** 200–202 (2004).
25. Rollins, N. J. *et al.* 3D protein structure from genetic epistasis experiments. *bioRxiv* 320721 (2018). doi:10.1101/320721
26. Schmiedel, J. & Lehner, B. Determining protein structures using genetics. *bioRxiv* 303875 (2018). doi:10.1101/303875
27. Olson, C. A., Wu, N. C. & Sun, R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Curr. Biol.* **24,** 2643–2651 (2014).
28. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **34,** 666–681 (2018).
29. Kircher, M., Stenzel, U. & Kelso, J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* **10,** R83 (2009).
30. Stec, B., Holtz, K. M., Wojciechowski, C. L. & Kantrowitz, E. R. Structure of the wild-type TEM-1 β-lactamase at 1.55 Å and the mutant enzyme Ser70Ala at 2.1 Å suggest the mode of noncovalent catalysis for the mutant enzyme. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **61,** 1072–1079 (2005).
31. Jelsch, C., Mourey, L., Masson, J.-M. & Samama, J.-P. Crystal structure of Escherichia coli TEM1 β-lactamase at 1.8 Å resolution. *Proteins Struct. Funct. Bioinforma.* **16,** 364–383 (1993).
32. Ambler, R. P. *et al.* A standard numbering scheme for the class A beta-lactamases. *Biochem. J.* **276,** 269–70 (1991).
33. Wang, X., Minasov, G. & Shoichet, B. K. The Structural Bases of Antibiotic Resistance in the Clinically Derived Mutant β-Lactamases TEM-30, TEM-32, and TEM-34. *J. Biol. Chem.* **277,** 32149–32156 (2002).
34. Huang, W. & Palzkill, T. A natural polymorphism in β-lactamase is a global□suppressor. *Proc. Natl. Acad. Sci.* **94,** 8801 LP-8806 (1997).
35. Goldsmith, M. & Tawfik, D. S. Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proc. Natl. Acad. Sci.* **106,** 6197 LP-6202 (2009).
36. De Visser, J. A. G. M., Salverda, M. L. M. & Barlow, M. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiol. Rev.* **34,** 1015–1036 (2010).
37. Wang, B. & Kennedy, M. A. Principal components analysis of protein sequence clusters. *J. Struct. Funct. Genomics* **15,** 1–11 (2014).
38. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning

generative models for protein fold families. *Proteins* **79,** 1061–78 (2011).
39. Lacapère, J.-J., Pebay-Peyroula, E., Neumann, J.-M. & Etchebest, C. Determining membrane protein structures: still a challenge! *Trends Biochem. Sci.* **32,** 259–270 (2007).
40. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **18,** 581–586 (2008).
41. McPherson, A. & Gavira, J. A. Introduction to protein crystallization. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **70,** 2–20 (2014).
42. Marks, D. S., Hopf, T. a & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30,** 1072–80 (2012).
43. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
44. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Mol. Biol. Evol.* **33,** (2016).
45. Zhang, J. Protein-length distributions for the three domains of life. *Trends Genet.* **16,** 107–109 (2000).
46. Visintin, M., Tse, E., Axelson, H., Rabbitts, T. H. & Cattaneo, A. Selection of antibodies for intracellular function using a two-hybrid in vivo system. *Proc Natl Acad Sci U S A* **96,** 11723–11728 (1999).
47. Chirichella, M. *et al.* Post-translational selective intracellular silencing of acetylated proteins with de novo selected intrabodies. *Nat. Methods* **14,** (2017).
48. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
49. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).
50. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39,** (2011).
51. Fantini, M., Malinverni, D., De Los Rios, P. & Pastore, A. New techniques for ancient proteins: Direct coupling analysis applied on proteins involved in iron sulfur cluster biogenesis. *Front. Mol. Biosci.* **4,** (2017).
52. Markley, J. L. *et al.* Metamorphic protein IscU alternates conformations in the course of its role as the scaffold protein for iron-sulfur cluster biosynthesis and delivery. *FEBS Lett.* **587,** 1172–1179 (2013).
53. Malinverni, D., Marsili, S., Barducci, A. & de Los Rios, P. Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. *PLoS Comput. Biol.* **11,** 1–15 (2015).
54. Schäfer, J. & Strimmer, K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology* **4,** (2005).