Published in final edited form as:

Angew Chem Int Ed Engl. 2018 May 14; 57(20): 5674–5678. doi:10.1002/anie.201713220.

Co-evolutionary fitness landscapes for sequence design

Pengfei Tian, John M. Louis, James L. Baber, Annie Aniana, and Robert B. Best*

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520 (USA)

Abstract

Efficient and accurate models to predict the fitness of a sequence would be extremely valuable in protein design. We have explored the use of statistical potentials for the coevolutionary fitness landscape, extracted from known protein sequences, in conjunction with Monte Carlo simulations, as a tool for design. As proof of principle, we created a series of predicted high-fitness sequences for three different protein folds, representative of different structural classes: the GA (all- α) and GB (α/β) binding domains of streptococcal protein G, and an SH3 (all- β) domain. We found that most of the designed proteins can fold stably to the target structure, and a structure for a representative of each for GA, GB and SH3 was determined. Several of our designed proteins were also able to bind to native ligands, in some cases with higher affinity than wild-type. Thus, a search using a statistical fitness landscape is an remarkably effective tool for finding novel stable protein sequences.

Entry for the Table of Contents

An efficient computational method is introduced to design proteins guided by a coevolutionary fitness landscape. A series of predicted high fitness sequences for three different protein folds are created. Most designs fold to the target structure, and many bind native ligands.

Keywords

direct coupling analysis; coevolution; epistatic; computation; multiple sequence alignment

* robert.best2@nih.gov.

Design of protein sequences to fold to a given structure and perform a targeted function is of enormous theoretical^[1], medical and industrial interest^[2]. This can be achieved experimentally, by screening methods such as directed evolution^[3], as well as computationally by structure-based de novo protein design^[4]. A variety of successful computational algorithms exist, of which the best known is ROSETTA, that exploit both the information in structural databases together with a combination of physics-based and knowledge-based energy functions^[5]. However, as we have shown, the number of sequences which fold to a given structure is far larger than could ever be sampled by algorithms that explicitly consider protein structure^[6]. Therefore, an effective and reliable method for sampling protein sequences of high fitness could be a valuable addition to the design toolbox.

Maintaining the functions and structural stability of the folded state is an important constraint on natural selection in protein evolution^[7]. Therefore, proteins from the same family, which share the same fold, should contain common features in their sequences, both in the propensities of residues to be at certain positions, as well as the covariation between different sites which are in contact in the native state, illustrated by the multiple sequence alignment (MSA) fragment in Figure 1A. These propensities and sequence correlations can be approximately captured using a simple statistical model, whose parameters are fitted to a multiple sequence alignment representing a set of evolutionarily-related proteins. Capturing the sequence correlations is essential: for example, none out of 43 sequences designed using only single site propensity to fold to a WW domain structure were folded^[8].

Residues that are proximal in the three-dimensional structure, tend to show correlated patterns of amino-acid occurrences, but correlations calculated by mutual information tend to be inflated by indirect effects^[9]. The Direct-Coupling Analysis (DCA) has been proposed to disentangle such indirect effects from direct (i.e. epistatic) couplings^[10], allowing prediction of native contacts of protein structures, protein-protein interactions and RNA structures,^[11] and the likelihood model underlying them has also been shown to be correlated with protein stability^[12]. Indeed, this type of analysis has been used to help design targeted single-point sequence variants^[13]. However, the reliability of such likelihood functions for finding completely novel foldable sequences was less clear. Our own work had suggested that DCA models could be predictive of protein stability for a sequence of designed WW domain variants^[6], but sequences designed using DCA models had never been directly tested. In the present proof of principle study, we demonstrate that the direct coupling analysis method can be used to build a sequence fitness landscape which could be an effective guide for rational protein engineering purposes.

The fitness model is described by single site amino acid propensities $h_i(X)$ for amino acid type X to be at position i , and residue-residue coupling parameters $J_{ij}(X, Y)$, describing the coupling between residue types X, Y at positions i, j . A Potts-like likelihood function $P(A_1, A_2, \dots, A_L)$ is used to represent the likelihood of a given amino acid sequence $A_1, A_2, \dots, A_L$ for a particular protein fold^[10d, 10e],

$$P(A_1, A_2, \dots, A_L) = \frac{1}{Z} \exp \left\{ \sum_{i < j} J_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\} \quad (1)$$

where Z is a normalization constant. The optimal parameters $h_i(X)$ and $J_{ij}(X, Y)$ are learned from an MSA (Figure 1A) using a pseudo-likelihood optimization procedure. An “Evolutionary Hamiltonian” energy E_{EH} can then be associated with sequence x via

$$E_{EH}(x) = -\ln P(x) \quad (2)$$

Guided by this model, Monte Carlo simulation in sequence space can generate artificial sequences with similar statistical properties to natural sequences. The single-site propensities are in good agreement with those from the raw MSA (SI Figure S1). We built an evolutionary Hamiltonian for each protein family (GA, GB and SH3): E_{GA} and E_{GB} and E_{sh3} , using Equations 1 and 2, based on MSA's generated with query sequences of wild type GA, GB and SH3 respectively (pdb codes: 2FS1^[14] and 1PGA^[15], 3THK^[16]), using the Jackhmmer method^[17]. Evidence that the model also captures the couplings between residues comes from the consistency between the contacts predicted from the coupling parameters and the interactions in the native structure (Figure 1C and B). Protein stability is usually an important constraint on natural selection, therefore we can test the model by comparing E_{EH} with the experimental stability data for designed variants under different conditions^[18]. As we can see for the mutants of GA (Figure 1D), GB and SH3 (SI Figure S2) the evolutionary statistical energy correlates very well with the experimental melting temperature/stability, as seen for other protein families^[6].

To explore the sequence energy landscape and find sequences with E_{EH} near the global minimum, simulated annealing Monte Carlo simulations are carried out. In each Monte Carlo iteration, the amino acid of one randomly chosen residue is proposed to mutate to another type of amino acid. The mutation is accepted/rejected based on the changes of E_{EH} using a Metropolis criterion. Further details of the simulation can be found in the SI. The generated sequences whose E_{EH} is low are considered to be stable sequences according to our model.

To test whether our protein design method is feasible in practice, we designed a series of sequence variants of three different protein folds: the GA (all- α) and GB (α/β) binding domains of streptococcal protein G and an SH3 (all- β) domain. By sampling the sequence space using Monte Carlo simulation, we find there is a large number of mutants which are considered to be stable ($E_{EH}(\text{mutant}) - E_{EH}(\text{WT})$) in each fold, in agreement with our earlier work^[6]. We selected 12 of them (GA Seq1–5, GB Seq1–2 and SH3 Seq1–5) for further experimental characterization. The exact sequences are listed in SI Table S1. They were chosen based on four criteria: 1) The evolutionary statistical energy is as low as possible. 2) The fractional sequence identity to wild type was low (Table 1). For instance, the sequence identity between SH3 Seq 1–5 and SH3 WT3thk is below 50% in each case; the fraction sequence identity is similar if only buried residues are considered (Table 1). This was

achieved in practice by adding a penalty term proportional to sequence similarity in the Monte Carlo simulation to enhance sequence diversity.

For each fold, these cut-offs were chosen such that it was still possible to obtain sufficient sequences predicted to be stable. 3) To ensure diversity, we also required that the fractional sequence identity between any pair of designed sequences is below 85%. 4) To obtain additional confidence, we estimated the stability of the designed sequences based on their unfolding time in molecular dynamics (MD) simulations, starting from native-like structures built by homology modelling. Homology modelling was carried out using the MODELLER^[19] package, based on the pdb structure of the wild type GA (2FS1^[14]), GB (1PGA^[15]) and SH3 (3THK^[16]) as template, respectively. The sequences whose estimated unfolding time using molecular dynamics unfolding simulations was comparable with the wild type were selected for further experimental verification. More details can be found in the SI.

The sequences were expressed and purified, together with the wild-type, and experimental conditions were optimized for each, based initially on those used previously for each wild-type sequence. The far-UV CD spectra of the mutants are generally very similar to the wild type (Figure 2 A, C and E), suggesting a similar fold. An exception was SH3 Seq2 whose spectrum suggests a high helical population. SH3 Seq5 showed poor expression and yielded mostly insoluble protein. The stability of each sequence was measured by determining thermal melting curves using the circular dichroism (CD) signal at 222 nm as a reporter on folding (Figure 2 B, D and F). The thermal denaturation curves of SH3 Seq1 and SH3 Seq3 are close to the wild type while the flat curves for SH3 Seq2 and SH3 Seq4 indicate that they might be unstable under the current experimental conditions (details in SI). The thermal denaturation midpoints (T_m) of the two GB mutants are very close to that of the wild type. For GA, all mutants are stable (Figure 2B) The melting temperatures are also consistent with relative stability predicted by the evolutionary statistical score E_{GA} (Fig. 2B inset).

We had originally included the MD selection step (4) in our design protocol to give added confidence to our predictions prior to experimental testing. Given the success of the design in producing folded sequences, we asked whether inclusion of this step was necessary, since it is by far the most computationally expensive. We therefore designed 3 further GA sequences (GA Seq 6*–8*) based on E_{GA} alone (without performing MD simulation). These also showed very similar folds to the wild type (Figure 2A) although the melting curves (Figure 2B) indicate that GA Seq8* is significantly less stable than the others). While this is a limited test, it does suggest that inclusion of the selection based on MD folding simulations is not essential. Indeed, the unfolding times estimated from simulation are strongly correlated with E_{ch} (SI Figure S9).

To further probe the fold of the designed sequences, we also studied one sequence of each fold by nuclear magnetic resonance (NMR) spectroscopy (Spectra are shown in SI)^[20]. ¹H-¹⁵N HSQC spectra of each variant (Figure 3 and SI Figure S10–12) show well-dispersed spectra indicative of folded structure. For selected representatives of each fold (GA Seq5, GB Seq2 and SH3 Seq3), we were able to determine high resolution structures using CS-

Rosetta^[21] with restraints derived from chemical shifts and residual dipolar couplings. The resulting structures (Figure 3) are highly consistent with the target GA, GB and SH3 folds.

The structural analysis of the designed sequences demonstrates that direct evolutionary couplings can specify a network of global interactions between amino acids defining a given protein fold. We were interested whether these folded artificial sequences also retain the function of their natural counterparts. The wild type GA domain is known to bind Human serum albumin (HSA)^[22], which is one of the most abundant plasma proteins. Cell surface proteins interacting with these molecules are found in different pathogenic bacterial species. Here, we have used isothermal titration calorimetry (ITC) to determine the complete binding thermodynamics of GA Seq 1–5 to HSA. The binding isotherms obtained for GA/HSA binding at room temperature are given in Figure 4 (and SI Figure S13), from which one can see GA Seq1 binds more strongly than the wild type (binding dissociation constants K_D for GA WT, GA Seq1 are 6.0 and 1.4 nM respectively, SI Table S2). GA Seq2 and GA Seq5 bind to HSA more weakly, with the K_D of 70.0 and 213.0 nM, respectively. K_D could not be determined for GA_Seq3 and GA_Seq4 since the binding affinity is very low. We found that the K_D of the designed sequences is highly correlated with the number of mutations at the binding interface between GA WT2fs1 and HSA (SI Table S2). This is expected since the specific HSA used in the ITC experiment is the natural binding partner of GA WT2fs1. However, one ought be able to design GA mutants which have strong binding affinity to a given HSA sequence by building a combined evolutionary landscape of the two proteins, including also the coevolution information between GA and HSA. We also measured K_D of GB WT and mutants binding to its partner IgG^[23]. GB Seq2 shows stronger binding affinity than the wild type while GB Seq1 does not bind significantly (Figure 4, SI Figure S14 and SI Table S3).

To our knowledge, the only previous complete design of protein sequences based on coevolution is the pioneering work by Ranganathan and co-workers, who used residue-residue coevolution information to create a stable and functional 35 amino acid WW domain^[8], using a method known as statistical coupling analysis (SCA)^[8]. Their method results in stable folds with similar sequence properties to those designed here, as well as the ability to bind ligands^[24]. In their design algorithm, a natural MSA is initially scrambled by random permutation of the amino acids at each position in the alignment. The design proceeds by moves in which pairs of residues are swapped between sequences, in such a way as to optimize a score measuring the statistical similarity of the designed set of sequences to the natural alignment. There are many conceptual similarities between the SCA-based design and ours. Perhaps the main difference is that rather than designing a set of sequences, we use a likelihood function for individual sequences which captures single site conservation and direct coupling effects between pairwise residues^[6, 10e]. This makes it easier to explore the fitness landscape of a given fold using the established sampling methods of statistical mechanics^[6]. The evolutionary score can also be used as a proxy for the stability of the folded state^[6].

Here, we have shown that a simple evolutionary energy function derived from known sequences can be a useful tool for both protein stability prediction and sequence design. The prediction is robust with the size of MSA which is used to build the model (SI Figure S15).

In particular, we have designed a number of sequences for three representative protein folds from different structural classes. The majority of our designed sequences are folded, and for each fold, we have determined an experimental folded structure for a representative sequence. Analysis of the binding of GA and GB mutants to their native partners, HSA and IgG respectively, suggests that many also retain native binding function. Although this method is of course limited to redesign of known folds for which a sufficient number of existing sequences are known, our results suggest that it is sufficiently accurate to be useful as a design tool. An interesting future possibility could be the combination of DCA-based design with cutting-edge template-based design methods, such as the RosettaDesign^[5b] or MODELLER^[19], in order to increase sequence diversity or to improve the quality of prediction. Another could be the design of “bridge” sequences that are able to switch between different folds^[18].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the Intramural Research Program of NIDDK, NIH. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD. (<http://biowulf.nih.gov>). We acknowledge use of the NIDDK Advanced Mass Spectrometry Core Facility and particularly John Lloyd for his timely support. JLB thanks Yang Shen for assistance in using the CS-Rosetta program and Jinfa Ying and Mengli Cai for helpful discussions. We thank Catherine Hefferan for creating the cover art.

References

- [1]. Bornberg-Bauer E, Chan HS, Proceedings of the National Academy of Sciences 1999, 96, 10689–10694.
- [2]. Brekke OH, Sandlie I, Nature Reviews Drug Discovery 2003, 2, 52–62. [PubMed: 12509759]
- [3]. Romero PA, Arnold FH, Nature reviews. Molecular cell biology 2009, 10, 866. [PubMed: 19935669]
- [4]. Chevalier A, Silva D-A, Rocklin GJ, Hicks DR, Vergara R, Murapa P, Bernard SM, Zhang L, Lam K-H, Yao G, Nature 2017.
- [5]. Song Y, DiMaio F, Wang RY-R, Kim D, Miles C, Brunette T, Thompson J, Baker D, Structure 2013, 21, 1735–1742 [PubMed: 24035711] Dantas bG., Kuhlman B, Callender D, Wong M, Baker D, Journal of molecular biology 2003, 332, 449–460. [PubMed: 12948494]
- [6]. Tian P, Best RB, Biophysical Journal 2017, 113, 1719–1730. [PubMed: 29045866]
- [7]. England JL, Shakhnovich BE, Shakhnovich EI, Proceedings of the National Academy of Sciences 2003, 100, 8727–8731.
- [8]. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R, Nature 2005, 437, 512–518. [PubMed: 16177782]
- [9]. De Juan D, Pazos F, Valencia A, Nature Reviews Genetics 2013, 14, 249–261.
- [10]. aMarks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C, PloS one 2011, 6, e28766 [PubMed: 22163331] bMorcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M, Proceedings of the National Academy of Sciences 2011, 108, E1293–E1301 cSchug A, Weigt M, Onuchic JN, Hwa T, Szurmant H, Proceedings of the National Academy of Sciences 2009, 106, 22124–22129 dKamisetty H, Ovchinnikov S, Baker D, Proceedings of the National Academy of Sciences 2013, 110, 15674–15679 eEkeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E, Physical Review E 2013, 87, 012707.

- [11]. aToth-Petroczy A, Palmedo P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS, Cell 2016, 167, 158–170. e112 [PubMed: 27662088] bWeinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS, Cell 2016, 165, 963–975 [PubMed: 27087444] cDe Leonardis E, Lutz B, Ratz S, Cocco S, Monasson R, Schug A, Weigt M, Nucleic acids research 2015, 43, 10444–10455. [PubMed: 26420827]
- [12]. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, Marks DS, Nature biotechnology 2017, 35, 128.
- [13]. Cheng RR, Haglund E, Tiee N, Morcos F, Levine H, Adams JA, Jennings PA, Onuchic JN, bioRxiv 2017, 116947.
- [14]. He Y, Rozak DA, Sari N, Chen Y, Bryan P, Orban J, Biochemistry 2006, 45, 10102–10109. [PubMed: 16906768]
- [15]. Gallagher T, Alexander P, Bryan P, Gilliland GL, Biochemistry 1994, 33, 4721–4729. [PubMed: 8161530]
- [16]. Gushchina LV, Gabdulkhakov AG, Nikonov SV, Filimonov VV, Journal of Biomolecular Structure and Dynamics 2011, 29, 485–495. [PubMed: 22066535]
- [17]. Eddy SR, in Genome Inform, Vol. 23, 2009, pp. 205–211. [PubMed: 20180275]
- [18]. Alexander PA, He Y, Chen Y, Orban J, Bryan PN, Proceedings of the National Academy of Sciences 2009, 106, 21149–21154.
- [19]. Webb B, Sali A, Protein Structure Prediction 2014, 1–15.
- [20]. Bax A, Grzesiek S, Accounts of Chemical Research 1993, 26, 131–138.
- [21]. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A, Proceedings of the National Academy of Sciences 2008, 105, 4685–4690.
- [22]. Falkenberg C, Bjoerck L, Akerström B, Biochemistry 1992, 31, 1451–1457. [PubMed: 1737003]
- [23]. Myhre EB, Kronvall G, Infection and immunity 1977, 17, 475–482. [PubMed: 409673]
- [24]. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R, Nature 2005, 437, 579–583. [PubMed: 16177795]

Figure 1. Statistical fitness landscape inferred from homologous Pfam family. (A) Single site amino acid propensity and epistatic couplings. (B) Structure of GA. (C) Comparison of native contacts (grey) with 40 top-scoring contacts (yellow) predicted from residue coevolution information for GA. Top contacts are also indicated in yellow in B. (D) Correlation between evolutionary statistical energy of GA (E_{GA}) and the folding temperature of designed GA variants.

Figure 2.

Analysis of secondary structure and thermal denaturation by CD. The far-UV CD spectra for wild-type and designed sequences are shown for (A) GA, (C) GB and (E) SH3 folds.

Temperature melts monitored at 222 nm are plotted in (B), (D) and (F) for the respective proteins, and fitted to a two-state folding model. E_{GA} vs T_m for GA wt and mutants is plotted in the inset of B).

Figure 3.

NMR characterization of selected sequences. ^1H - ^{15}N HSQC spectra are shown for A) GA Seq5, C) GB Seq2 and E) SH3 Seq3. Spectra labelled with peak assignments are given in SI Figures S10–12. The peaks connected by black lines are the side chain amide resonances of Q or N. The red peaks for SH3 Seq3 E) are folded sidechain R resonances that only appear if they are hydrogen-bonded. Hence, there are only 2 peaks for the 3 R's in the sequence. The unlabeled peaks at $> \sim 10$ ppm ^1H are due to W sidechain resonances. NMR structures have been determined for B) GA Seq5, D) GB Seq2 and F) SH3 Seq3. The first column of B), D) and F) shows ribbon diagrams of each mutant structure (in color) and the corresponding wild type (in grey). The second and third columns give a ribbon diagrams of wild-type and mutant structures respectively, with side chains of those residues which differ between them shown in magenta.

Figure 4.

Binding isotherm for the interaction of GA with HSA (A) and GB with IgG (B) in 50 mM sodium phosphate buffer (pH 7) at 28 °C. The data were best fit using a single binding constant to calculate the thermodynamic parameters. Data for GA Seq3, GA Seq4 and GB Seq1 are not included because of low binding affinity.

Table 1.

Properties of designed proteins. PDB code corresponding to each sequence considered as wild-type are given in parentheses.

	Mutants /WT	IDwt ^[^]	IDns ^[#]	IDave ^[‡]	IDbur ^[&]	EEH (kBT)	Tm (°C) [+]
GA fold	WT (2fs1)					-127.1	86
	Seq1	79%	79%	22%	94%	-129.0	86
	Seq2	54%	71%	19%	50%	-117.0	63
	Seq3	50%	57%	25%	62%	-114.7	73
	Seq4	50%	52%	24%	62%	-111.4	66
	Seq5	50%	64%	20%	50%	-111.5	59
	Seq6 [*]	50%	64%	23%	75%	-115.6	55
	Seq7 [*]	50%	50%	25%	69%	-115.0	40
	Seq8 [*]	50%	66%	21%	50%	-114.9	-
GB fold	WT (1pga)					-106.5	77
	Seq1	75%	79%	14%	79%	-94.2	73
	Seq2	75%	79%	18%	71%	-93.9	75
SH3 fold	WT (3thk)					-72.3	70
	Seq1	45%	57%	34%	64%	-96.4	64
	Seq2	45%	59%	34%	64%	-96.5	-
	Seq3	48%	59%	35%	71%	-97.9	63
	Seq4	46%	57%	34%	64%	-95.0	-
	Seq5	45%	57%	34%	64%	-94.6	-

[^] Identity to wide type.

[#] Identity to closest natural sequences.

[‡] Average identity to natural sequences from the MSA.

[&] Sequence identity of buried residues to the wild type.

[+] Folding Temperature measured by fitting CD experiments using a two-state model.

[*] Selected without testing by MD simulation.