RESEARCH ARTICLE

Long-Range Epistasis Mediated by Structural Change in a Model of Ligand Binding Proteins

Erik D. Nelson*, **Nick V. Grishin**

Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Room ND10.124, Dallas, Texas, United States of America

* nelsonerikd@gmail.com

OPEN ACCESS

Citation: Nelson ED, Grishin NV (2016) Long-Range Epistasis Mediated by Structural Change in a Model of Ligand Binding Proteins. PLoS ONE 11(11): e0166739. doi:10.1371/journal.pone.0166739

Editor: Yaakov Koby Levy, Weizmann Institute of Science, ISRAEL

Received: August 12, 2016

Accepted: November 2, 2016

Published: November 21, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported in part by the National Institutes of Health (GM094575 to NVG) and the Welch Foundation (I-1505 to NVG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Recent analyses of amino acid mutations in proteins reveal that mutations at many pairs of sites are epistatic—i.e., their effects on fitness are non—additive—the combined effect of two mutations being significantly larger or smaller than the sum of their effects considered independently. Interestingly, epistatic sites are not necessarily near each other in the folded structure of a protein, and may even be located on opposite sides of a molecule. However, the mechanistic reasons for long–range epistasis remain obscure. Here, we study long–range epistasis in proteins using a previously developed model in which off–lattice polymers are evolved under ligand binding constraints. Epistatic effects in the model are qualitatively similar to those recently reported for small proteins, and many are long–range. We find that a major reason for long–range epistasis is conformational change—a recurrent theme in both positive and negative epistasis being the transfer, or exchange of material between the ordered nucleus, which supports the binding site, and the liquid–like surface of a folded molecule. These local transitions in phase and folded structure are largely responsible for long–range epistasis in our model.

Introduction

Epistasis in protein biophysics refers to the non–additive effects of amino acid mutations on protein folding and function [1]. An epistatic interaction is said to occur between two mutations when their combined effect on a trait is either larger or smaller than the sum of their effects considered independently. For example, a mutation that, by itself, damages the biochemical function of a protein may have a neutral or beneficial effect when considered in the presence of another mutation—a form of positive epistasis. Conversely, mutations that have a neutral, or nearly neutral effect on function individually, may produce a damaging effect in combination—a form of negative epistasis. Recent analyses of amino acid mutations in proteins reveal that mutations at many pairs of sites are epistatic, and suggest that epistasis plays a significant role in protein evolution [1–13]. Interestingly, epistatic sites are not necessarily in contact with each other in the folded state of a protein, and may even be located on opposite sides of a molecule. However, because the partially folded and mis–folded ensembles resulting from epistatic mutations are difficult to study, the mechanistic reasons for long–range epistasis remain somewhat obscure.

In this work, we investigate long–range epistasis using a previously developed model in which polymers were evolved to imitate the behavior of small ligand binding proteins [14]. The model re–capitulates basic properties of evolved proteins, such as folding to an ordered, soluble native structure, maintenance of amino acid sequence complexity [15], linear rates of amino acid change as a function of solvent exposure [16], packing density [17], and distance from the binding site [18], and linear rates of structure divergence as a function of the number of accepted mutations [19]. Below, we sample epistatic effects in evolved polymers by random selection of pair mutations, and we study the folded and mis–folded ensembles of single and double mutants in instances of significant epistasis—epistasis being measured in terms of the probability of folding a structure in which the binding site is correctly formed. Epistatic effects in the model resemble those reported by Olson et. al [2] for the small IgG–binding domain of protein G (protein GB1), and many are long–range. We find that a major reason for long–range epistasis is conformational change: In positive epistasis, either or both mutations disrupt the binding site, however, the double mutant folds to a locally re–configured native state in which the binding site is maintained. In negative epistasis, a single mutation leads to a neutral, or slightly deleterious change in native structure; This change conflicts with a second (formerly neutral, or slightly deleterious) mutation, leading to more frequent mis–folding of the binding site. A recurrent theme in both positive and negative epistasis is the transfer, or exchange of material between the ordered nucleus, which supports the binding site, and the disordered surface of a molecule, reminiscent of the theory of allostery in protein domains [20]. These local transitions in phase and structure are largely responsible for long–range epistasis in our model. Alternatively, neutral, or slightly deleterious mutations that preserve the native structure can conspire to frustrate formation of the binding site during folding. In this case, a mutation that increases the energy of the native fold relative to mis–folded states, or provides for greater conformational freedom during folding, amplifies the negative effect of a second mutation, presumably as a result of proximity to a thermodynamic phase transition [1, 2]. We find, however, that this process does not necessarily result in a complete change of phase.

In the following, we describe the distribution of epistatic effects in our model, and then explore the mechanisms of positive and negative epistasis in specific examples. First, we briefly outline our model; A more detailed description can be found in the Methods section.

Model

The polymer model is a chain of point monomers that interact as low resolution amino acids via spherically symmetric pair potentials. Polymers evolve kinetically by Langevin dynamics. The fitness of a sequence is determined by folding $\mathcal{N} = 127$ polymer replicas on a parallel computer and analyzing the resulting ensemble of structures: Each replica is initiated from a random coil state below the folding temperature of a typical viable sequence. The amount of time allowed for folding is determined by the number of amino acids in a sequence, N, according to an estimate provided by Lin and Zewail [21]. The temperature is then reduced substantially, the replicas are equilibrated for a short period, and a final ensemble of folded and quenched structures, Γ, is recovered.

To obtain the sequences studied in this work, polymers were first evolved to recover an ordered (but otherwise un–restricted) folding domain, as determined by the Lindemann melting criterion [22]. To apply the Lindemann criterion, we select a structure $\mathbf{x}^{\star}$ and an ensemble $\Delta\Gamma^{\star}$ of $3\mathcal{N}/4$ replicas to represent the dominant energy basin recovered by the folding procedure. Here, $\mathbf{x}^{\star}$ plays a role analogous to the equilibrium (lattice) positions of atoms in a crystal.

The structure $\mathbf{x}^{\star}$ and the ensemble $\Delta\Gamma^{\star}$ are selected to minimize the structurally aligned distance between $\mathbf{x}^{\star}$ and replicas $\mathbf{x}^{\mu}$ included in $\Delta\Gamma^{\star}$. To determine distance, the structures $\mathbf{x}^{\mu}$ are aligned to $\mathbf{x}^{\star}$ by rotation, translation, and reflection through the closest $2N/3$ pairs of monomers using methods described in reference [14]. Finally, sequences are selected to recover at least 15 ordered monomers, where order is measured by fluctuations of monomer positions $\mathbf{x}_{j}^{\mu}$ in structures $\mathbf{x}^{\mu} \in \Delta\Gamma^{\star}$ against their positions $\mathbf{x}_{j}^{\star}$ in the reference structure $\mathbf{x}^{\star}$.

Under this condition, polymers spontaneously evolve ordered surface cavities (putative binding sites), resembling the active sites of small enzymes [23]. These "enzyme–like" sequences are subsequently evolved to recover an ordered binding site compatible with a model ligand, with the requirement of folding an ordered domain relaxed. To enforce the binding condition, the ligand is optimally docked onto the binding sites of replicas folded by a given sequence (see Methods). If the distances between amino acids in the binding site of a replica (including the ligand) are each within 1 Angstrom of their corresponding distances in a pre–defined target state, the replica is considered "active". The fitness of a sequence is then defined by the fraction of active replicas, $\mathtt{P} = \mathcal{N}^{\star}/\mathcal{N}$, recovered by the folding and docking procedures.

In earlier work, we found that ligand binding, as defined above is, by itself, sufficient to maintain an ordered folding domain during evolution [14]: In order to maintain a given binding site structure against thermal fluctuations, it is necessary to maintain an ordered nucleus as a scaffolding to support the binding site. As a result, both binding affinity and thermodynamic stability are implicated in the fitness parameter $\mathtt{P}$.

Results

Below, we select two evolved sequences from these simulations to explore epistasis in the model (S1 and S2 Figs). In each sequence, pairs of sites are selected at random and subjected to random mutations roughly consistent with the genetic code (see S3 Fig and ref. [24]). For each pair of mutations, we compute the change in fitness, $\Delta\mathtt{P}_{\nu} = \mathtt{P}_{\nu} - \mathtt{P}_{0}$ where $\mathtt{P}_{\nu}$ is the fitness of a (single or double) mutant sequence, and $\mathtt{P}_{0}$ is the fitness of the initial evolved sequence. Epistasis is measured as

$$\epsilon = \Delta\mathrm{P}_{12} - \Delta\mathrm{P}_{1} - \Delta\mathrm{P}_{2} \tag{1}$$

in obvious notation. Fig 1 provides a correlation plot of $\Delta\mathtt{P}_{12}$ versus $\Delta\mathtt{P}_{1} + \Delta\mathtt{P}_{2}$ for the initial sequence in S1 Fig. The dashed line is a plot of the equation $\epsilon = 0$.

To estimate the uncertainty, or error in a single fitness measurement (i.e., against the value that would be obtained in the limit $\mathcal{N} \to \infty$), we computed the width, $\sigma(\mathtt{P})$, of the fitness distribution, $\omega(\mathtt{P})$, for a number of different sequences. The distribution of fitness values for an evolved sequence is shown in Fig 2. From these considerations, we find that the typical error, $\delta\mathtt{P}$, in a measurement of $\mathtt{P}$ (i.e., the typical width of a distribution, $\langle\, \sigma(\mathtt{P})\rangle$) is about $\delta\mathtt{P} \simeq 0.037$. The value of $\mathtt{P}_{0}$ used in Fig 1 is based on 10^{3} measurements, and consequently, the errors in $\Delta\mathtt{P}_{\nu}$ are essentially the same as the errors in $\mathtt{P}_{\nu}$. If errors in $\mathtt{P}_{\nu}$ are considered independent, the error in $\Delta\mathtt{P}_{1} + \Delta\mathtt{P}_{2}$ can be estimated as $\sqrt{2}\,\delta\mathrm{P}$, and the error in ϵ can be estimated as $\delta\epsilon \simeq \sqrt{3}\,\delta\mathrm{P}$.

In Fig 3, we plot the distribution of ϵ values, $\omega(\epsilon)$, for points in Fig 1 that satisfy either $\Delta\mathtt{P}_{12} \geq \lambda$, or $\Delta\mathtt{P}_{1} \geq \lambda$ and $\Delta\mathtt{P}_{2} \geq \lambda$, with $\lambda = -0.2$ (i.e., for positive epistasis, the double mutant is neutral, or slightly deleterious, and likewise for single mutants in negative epistasis. This cut through the distribution is somewhat arbitrary, and is simply intended to include mutants that have a chance of becoming fixed in evolution). The width of the distribution, $\sigma(\epsilon)$, is about 3 times the error in a measurement of ϵ. This result is maintained for

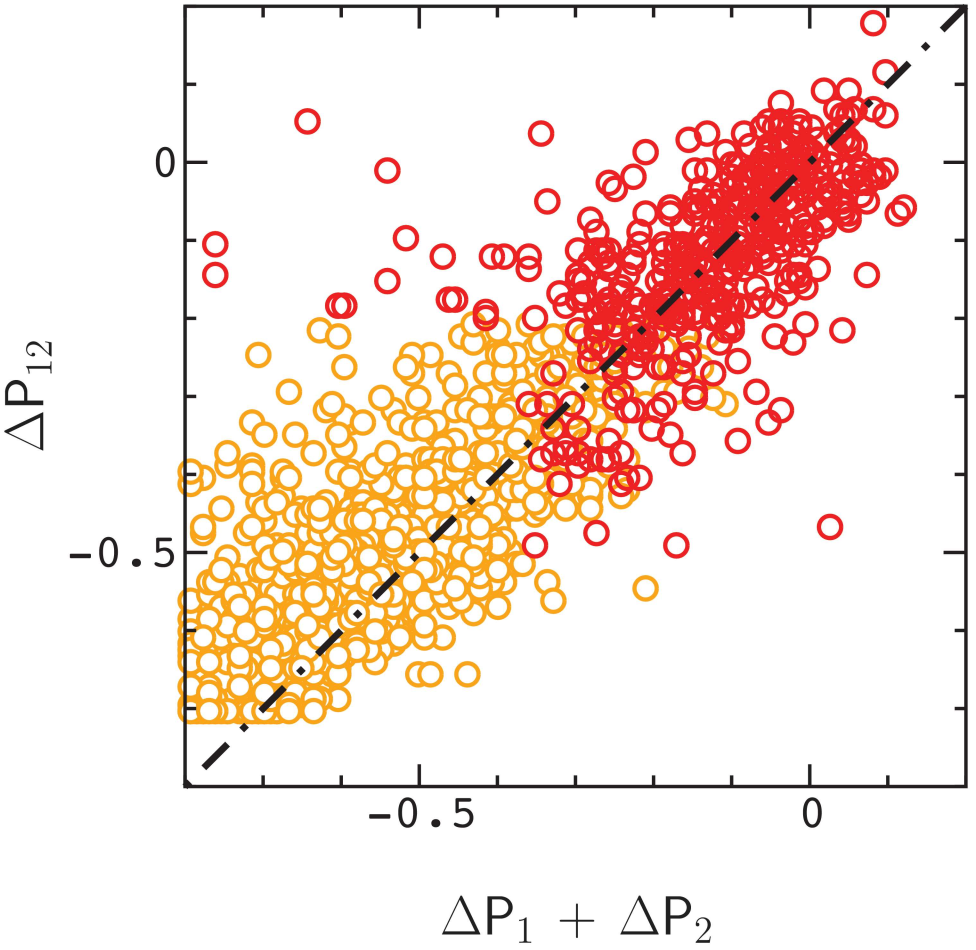

Fig 1. Correlation between ΔP_{12} and $\Delta P_1 + \Delta P_1$ for $\simeq 1.5 \times 10^3$ pair mutations randomly sampled from an evolved sequence folding to the ensemble shown in S1 Fig. The dashed line is a plot of the equation $\epsilon = 0$. Data points that satisfy either $\Delta P_{12} \geq \lambda$, or $\Delta P_1 \geq \lambda$ and $\Delta P_2 \geq \lambda$, with $\lambda = -0.2$ are colored red (see Text). The remaining data points are colored orange.

doi:10.1371/journal.pone.0166739.g001

essentially all values of λ (Fig 4). As a result, about 30 percent of the data points exhibit statistically significant epistasis ($|\epsilon| \tilde{>} 3\delta\epsilon$) in rough agreement with the results of Olson et. al [2] for protein GB1 (see also ref. [1]).

Finally, in Fig 5 we plot the distribution of ϵ values for points in Fig 1 as a function of the distance, R, between mutations in the initial fold. Amino acids begin to interact directly when $R < 1.5l$, where $l = 3.8$ Angstroms is the length of a polymer bond (see Methods). As is evident by inspection of Fig 5, many of the samples exhibit pronounced long–range epistasis, with $R > 1.5l$ and $\epsilon \gg \delta\epsilon$. The variation in ϵ values decreases with R, similar to the results for protein GB1, however, positive epistasis is more prevalent than negative epistasis in the model. Similar results to those above are obtained for the sequence in S2 Fig.

To conclude our analysis, we examined the folded ensembles of 12 samples exhibiting various levels of positive and negative epistasis. For each sample, fitness values, P_ν, were re–computed by averaging 10^2 measurements to minimize the error in ΔP_ν. Samples were selected essentially at random to include a range of values of ϵ.

In all 6 instances of positive epistasis (in particular, even when epistasis is weak) the double mutant is found to adopt a locally re–configured native structure that preserves the binding

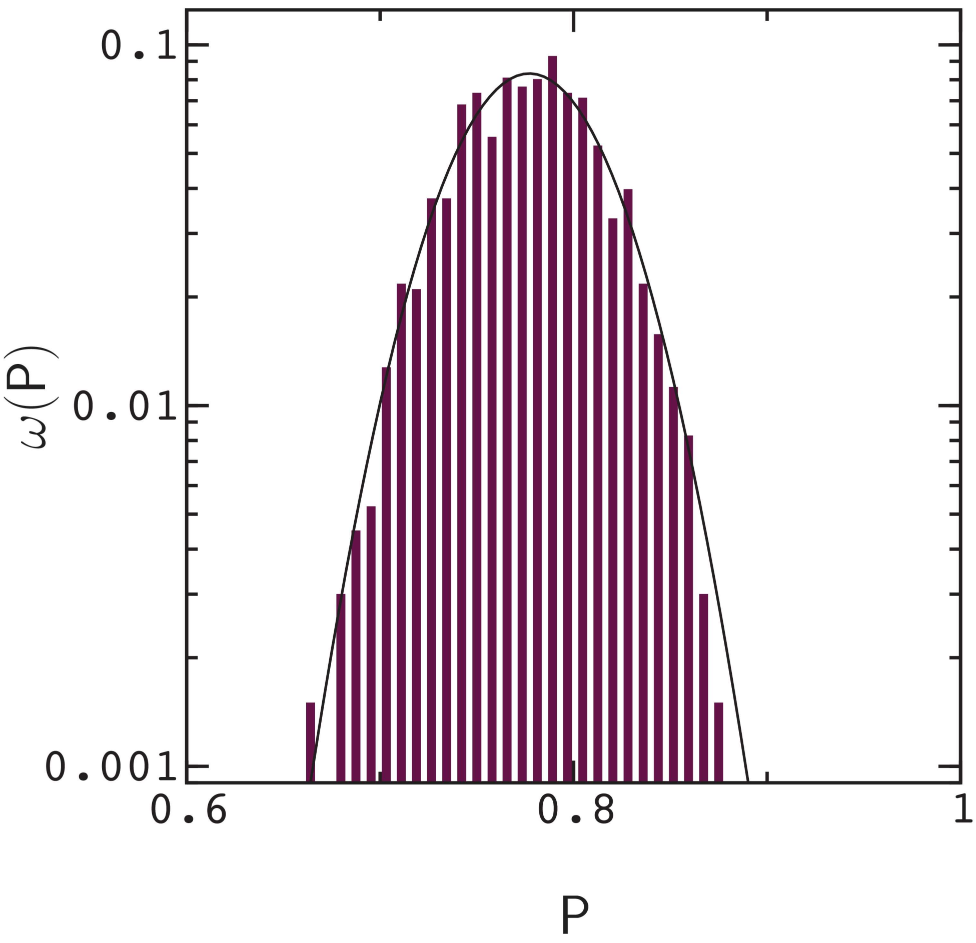

Fig 2. Distribution of fitness values, $\omega(\mathrm{P})$, obtained by the folding and docking procedure for an evolved sequence. The distribution is based on 10^3 measurements. The solid line is a fit to a Gaussian distribution with width $\sigma(P) \simeq 0.037$. Similar results are obtained for mutated sequences with lower mean fitness.

doi:10.1371/journal.pone.0166739.g002

complex. Fig 6 shows a typical example of long–range positive epistasis; Panels in this figure describe (A) the folded structure of the initial, un–mutated sequence, and (B) the folded structure of the double mutant (mutated sites are indicated by dotted spheres). Amino acids (monomers) are colored blue, light blue, blue–green, green, yellow, orange, and red, in order of increasing affinity to solvent. Binding site monomers are colored black. Charged amino acids R (orange), K, D and E (red) interact strongly, and play a similar role to hydrophobic amino acid types such as W, V, L, and I (blue) in stabilizing the binding sites of evolved sequences. To aid interpretation of Fig 6, the folded ensembles of the initial and mutated sequences are provided in S4 Fig. This scheme is repeated for the remaining examples studied below.

In this example, both single mutants recover disordered ensembles in which the binding site is disrupted: The mutation R1 → G1 destabilizes the charged part of the ordered scaffolding (red and orange) that supports the binding site, removing a charged amino acid; The mutation T12 → I12, while adding a hydrophobic amino acid, leads to competing nuclear structures. In the double mutant, the destabilizing effect of the first mutation is compensated by the second through local changes in the structure of chain segments containing the mutated

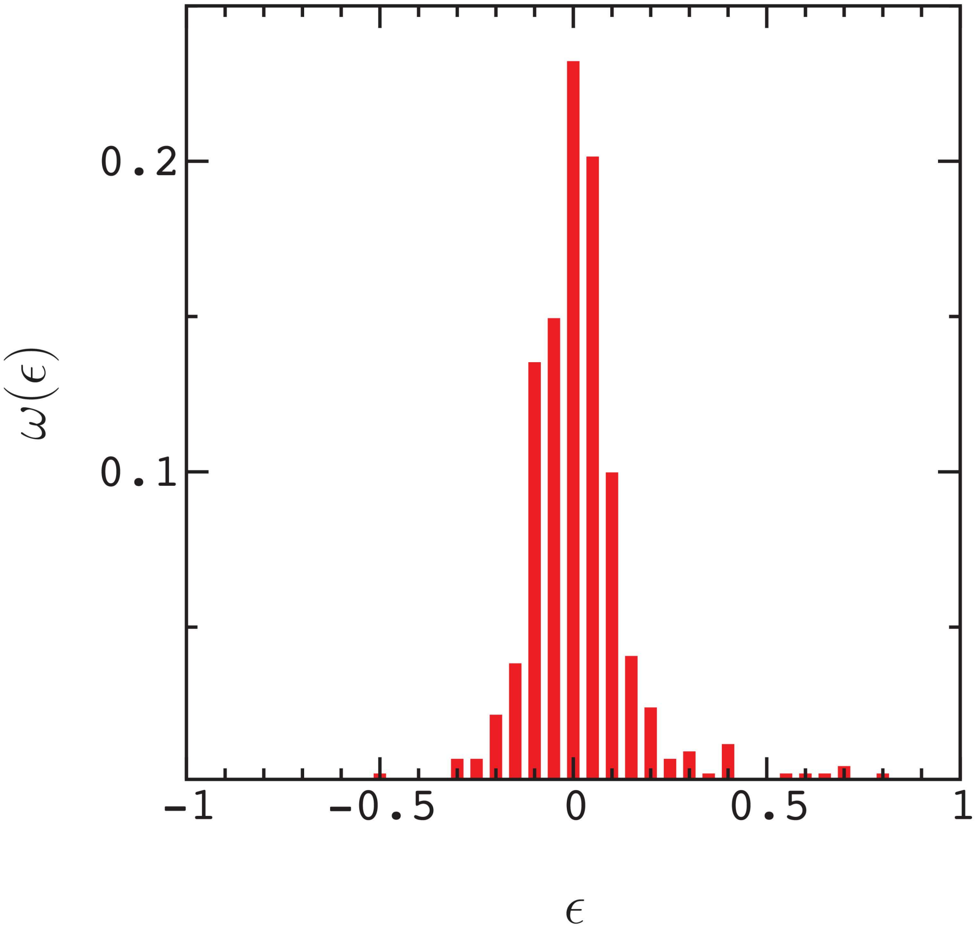

Fig 3. Distribution of epistasis vales, $\omega(\epsilon)$, for (red) data points satisfying the condition defined in Fig 1. The width of the distribution is $\sigma(\epsilon) \simeq 0.14$.

doi:10.1371/journal.pone.0166739.g003

sites, leading to an exchange of material between the ordered nucleus and less ordered (surface) regions of the folded ensembles (S4 Fig).

A similar example is shown in Fig 7. Here, the mutation W4 → S4 removes a hydrophobic monomer from the nuclear scaffolding leading to large fluctuations in the binding site; The mutation G10 → V10, while adding a hydrophobic amino acid to the nucleus, leads to competing nuclear structures during folding (originally, G10 is disordered). In the double mutant, V10 replaces W4 in the nucleus, and the mutated amino acid S4 is disordered. As in Fig 6, the double mutation leads to the re–configuration of loops containing the mutated sites, exchanging material between ordered and disordered phase regions of the folded ensembles. A slightly different mechanism of long–range positive epistasis is shown in Fig 8. Here, the mutation A23 → E23 has a stabilizing effect once the charged amino acid E6 is removed from the nuclear scaffolding by the mutation E6 → G6. Both A23 and E23 are ordered in their respective ensembles. In the double mutant, material is transferred from ordered to disordered phase regions of the folded ensembles by the mutation E6 → G6.

To study negative epistasis, we selected samples in which single mutations are neutral or slightly deleterious. Fig 9 provides an example of long–range negative epistasis in which the structural change caused by one mutation is frustrated by a second, formerly neutral mutation. Panel (A) describes the folded structure of the initial sequence, while panels (B) and (C)

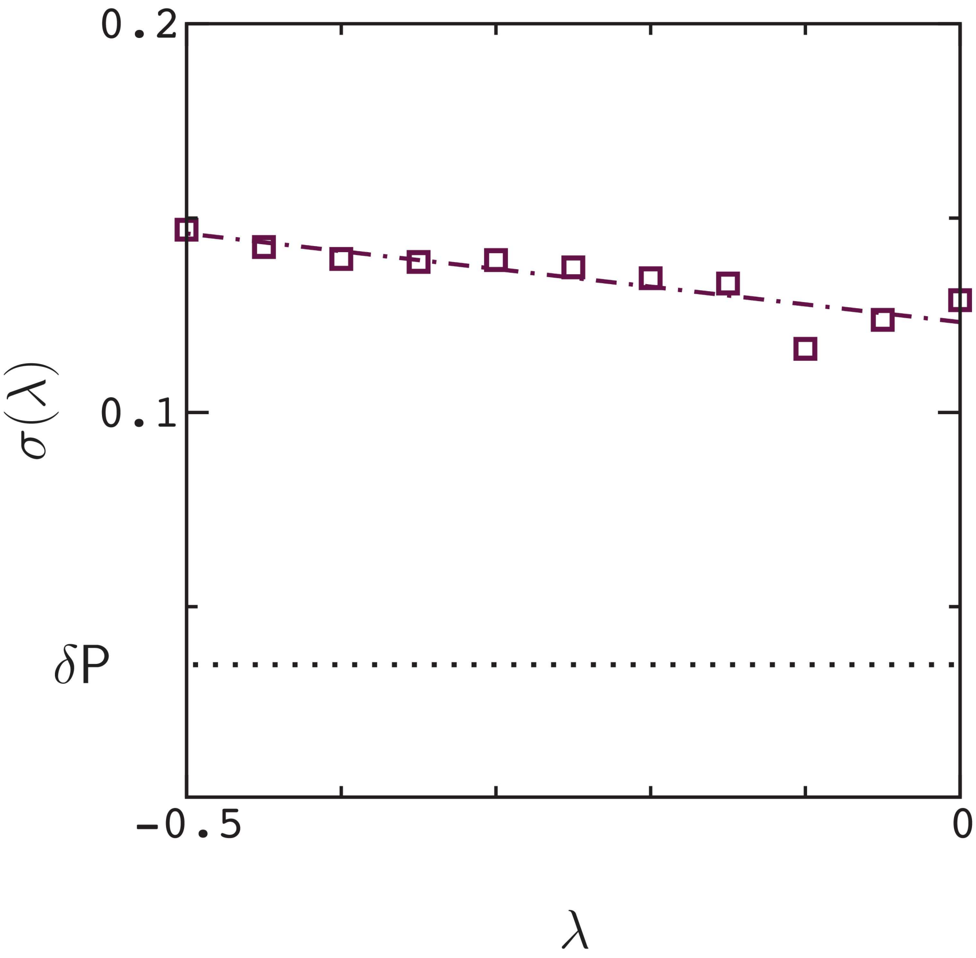

Fig 4. Width $\sigma(\epsilon)$ of the distribution $\omega(\epsilon)$ as a function of λ.

doi:10.1371/journal.pone.0166739.g004

describe the folded structures of the single mutants T15 → R15 and W4 → C4 respectively. The mutation T15 → R15 adds a charged amino acid to the nuclear scaffolding, which brings the mutant amino acid R15 in contact with W4; The interaction between R15 and W4 is attractive, and the amino acid T15 is disordered in the initial ensemble. The second mutation, W4 → C4, is neutral, and leaves the native structure essentially intact. However, the interaction between C4 and R15 is repulsive, which frustrates the transfer of R15 to the nucleus in the double mutant, disrupting the binding site. This pattern is repeated in 2 of the examples we studied.

Fig 10 describes an example of weak, but very long–range negative epistasis in which an exchange of material between ordered and disordered phases is frustrated. Again, panel (A) describes the folded structure of the initial sequence, while panels (B) and (C) describe the folded structures of the single mutants Q25 → E25 and E6 → G6 respectively. The mutation E6 → G6 removes a charged amino acid from the nuclear scaffolding. In the mutant ensemble, the amino acid G6 is disordered, however, this has no effect on fitness. The mutation Q25 → E25 adds an ordered charged monomer; However, in the mutant ensemble, the interactions between E25 and its neighbors are almost all repulsive. In the double mutant, the removal of E6 allows E25 to form favorable contacts with other charged monomers during folding, which leads to more frequent mis–folding of the binding site.

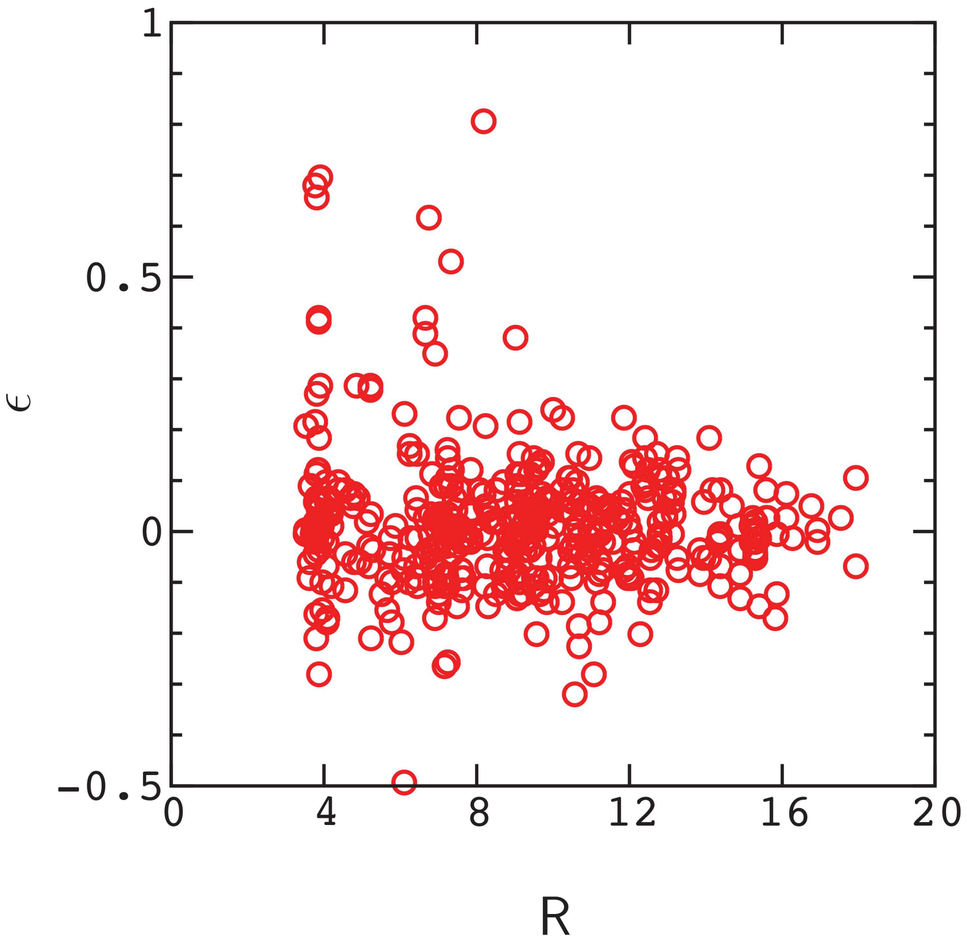

Fig 5. Distribution of epistasis values as a function of the distance, R, between mutations in the reference fold in S1 Fig. Data points satisfy the condition defined in Fig 1. Peaks in the density of points at $\mathrm{R} \sim l$, etc. reflect typical distances between pairs of monomers in the reference fold.

doi:10.1371/journal.pone.0166739.g005

The remaining samples of negative epistasis conform to the process described earlier above, in which a neutral, or slightly deleterious mutation that preserves the native structure amplifies the negative effect of another. An example of this effect exhibiting strong epistasis is provided in S9 Fig. In this sample, the mutation, D19 → Y19, reduces the specificity of interactions with its neighbors, allowing for greater conformational freedom of the binding site, making it more susceptible to the negative effect of the mutation, N13 → D13; The mutated amino acids Y19 and D13 form favorable contacts in mis–folded states of the replica ensemble.

Discussion

To summarize, many instances of long–range epistasis in our study involve local structural changes that transfer or exchange material between ordered and disordered phase regions of the initial and double mutant ensembles. In positive epistasis, the loss of an amino acid on the ordered scaffolding which supports the binding site creates a “vacancy” that is compensated by a second amino acid in the double mutant. The second amino acid either occupies a site close to the vacancy (Fig 6) or fills the vacancy with a similar amino acid (Fig 7). In negative epistasis, the filling of an existing vacancy on the ordered scaffolding is frustrated by a second mutation (Fig 9).

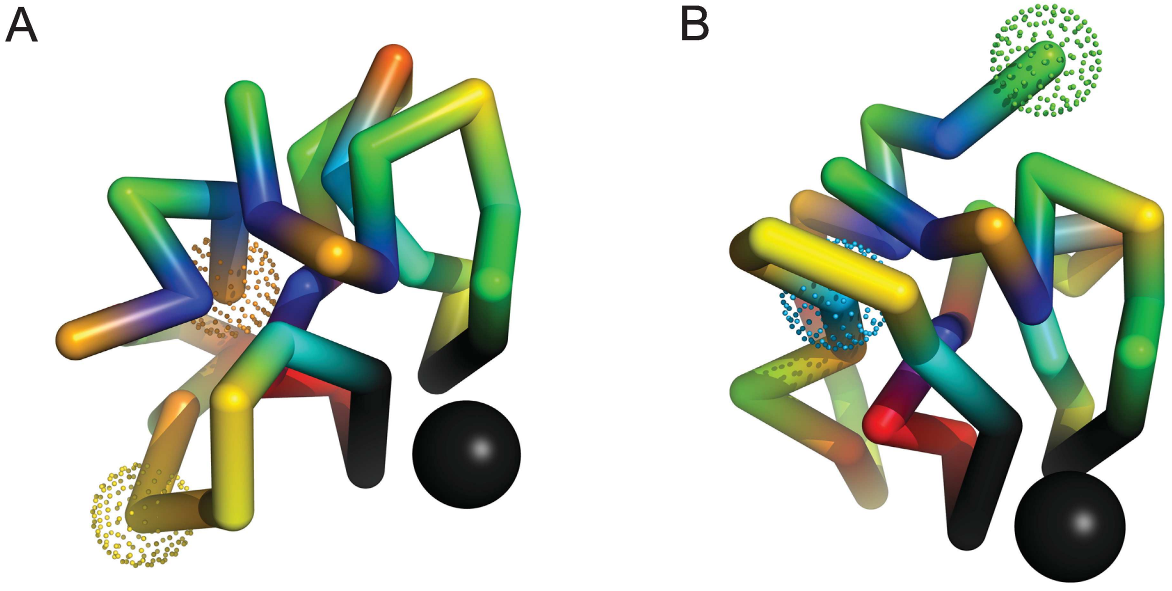

Fig 6. An example of long–range positive epistasis with ϵ = 0.77. Panel (A) describes the reference fold obtained for the initial sequence; Panel (B) describes the reference fold obtained for the double mutant. Dotted spheres indicate the positions of mutated amino acids. Individual mutations, R1 (orange) → G1 (green) and T12 (yellow) → I12 (light blue) are strongly deleterious, with $\Delta P_1 \simeq -0.61$ and $\Delta P_2 \simeq -0.29$ respectively ($\Delta P_{12} \simeq -0.13$). R1 is ordered (T12 is disordered) in the initial ensemble, and I12 is ordered (G1 is disordered) in the double mutant ensemble. The distance between mutated positions in panel (A) is $\mathrm{R} \simeq 8$ Angstroms. Folded ensembles for the initial and mutated sequences are shown in S4 Fig.

doi:10.1371/journal.pone.0166739.g006

In these examples, epistatic mutations connect local environments in a folded molecule that are dis–connected in the initial evolved structure. This may explain why long–range epistatic interactions are difficult to predict [25]: If the environments of mutated residues change significantly in epistasis, the information needed to predict epistatic interactions will depend on the

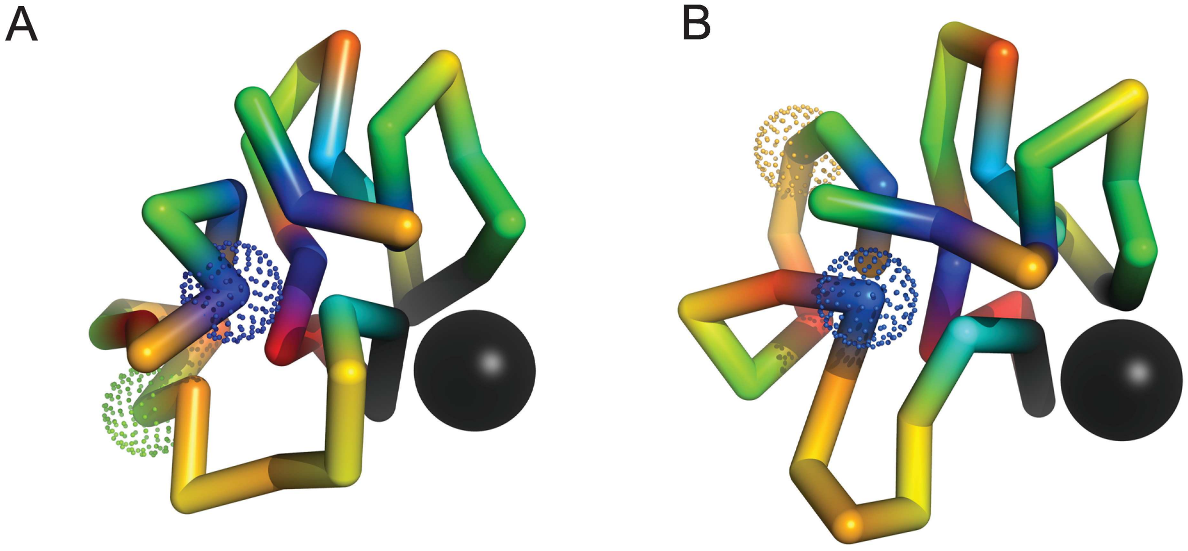

Fig 7. An example of long–range positive epistasis with ϵ = 0.25. Individual mutations, W4 (blue) → S4 (yellow) and G10 (green) → V10 (blue) are slightly deleterious, with $\Delta P_1 \simeq -0.17$ and $\Delta P_2 \simeq -0.11$ respectively ($\Delta P_{12} \simeq -0.04$). W4 is ordered (G10 is disordered) in the initial ensemble, and V10 is ordered (S4 is disordered) in the double mutant ensemble. In the double mutant, V10 occupies the nuclear site originally occupied by W4. The distance between mutated positions in panel (A) is $\mathrm{R} \simeq 9$ Angstroms. Folded ensembles for the initial and mutated sequences are shown in S5 Fig.

doi:10.1371/journal.pone.0166739.g007

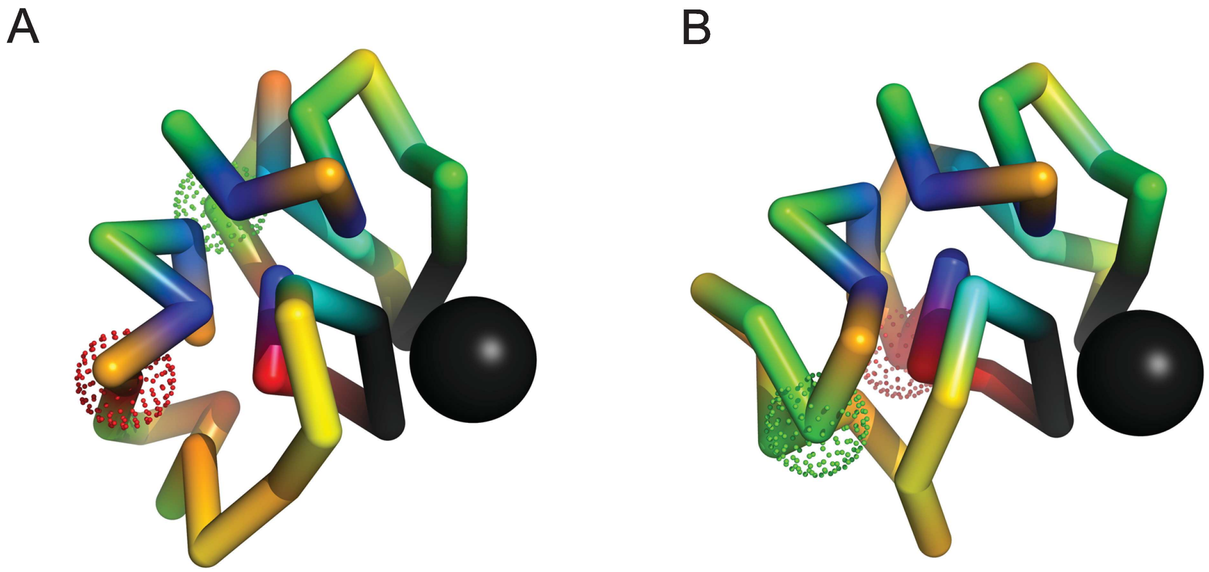

Fig 8. An example of long–range positive epistasis with $\epsilon \simeq 0.2$. Individual mutations E6 (red) → G6 (green) and A23 (green) → E23 (red) are neutral and deleterious, with $\Delta P_1 \simeq 0$ and $\Delta P_2 \simeq -0.24$ respectively ($\Delta P_{12} \simeq -0.04$). E6 and A23 are ordered in the initial ensemble, while E23 is ordered and G6 is disordered in the double mutant ensemble. In the double mutant, material is transferred from ordered to disordered phase regions of the ensembles by the mutation E6 → G6, while the mutation A23 → E23 enhances the stability of the nuclear scaffolding. The distance between mutated positions in panel (A) is $\mathrm{R} \simeq 8$ Angstroms. Folded ensembles for the initial and mutated sequences are shown in S6 Fig.

doi:10.1371/journal.pone.0166739.g008

environments of residues formed in the mutant structures, which will often be excluded from the evolutionary record by purifying selection. In addition, because computer modeling techniques are not sufficiently advanced to predict these changes, it is actually necessary to solve the folded structures of the mutant proteins to reveal the causes of epistasis. As a result, these effects, if they are detected, are very difficult to interpret, and may be more prevalent than expected.

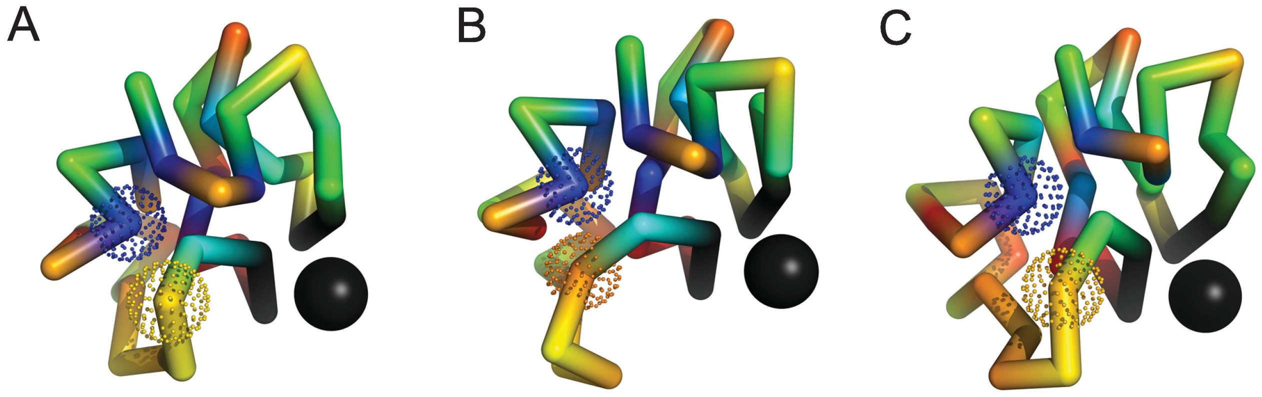

Fig 9. An example of long–range negative epistasis with $\epsilon \simeq -0.32$. Panel (A) describes the reference fold obtained for the initial sequence, while panels (B) and (C) describe the reference folds obtained for the single mutants; Individual mutations T15 (yellow) → R15 (orange) and W4 (blue) → C4 (blue) in panels (B) and (C) are slightly deleterious and neutral, with $\Delta P_1 \simeq -0.17$ and $\Delta P_2 \simeq 0.02$ respectively ($\Delta P_{12} \simeq -0.47$). W4 is ordered (T15 is disordered) in the initial ensemble, while R15 and C4 are ordered in the single mutant ensembles. The mutation T15 → R15 brings the mutant amino acid R15 in contact with W4. However, this structure (B) conflicts with the mutation W4 → C4 (the interaction between R and C is repulsive). The distance between mutated positions in panel (A) is $\mathrm{R} \simeq 6$ Angstroms. Folded ensembles for the initial and mutated sequences are shown in S7 Fig.

doi:10.1371/journal.pone.0166739.g009

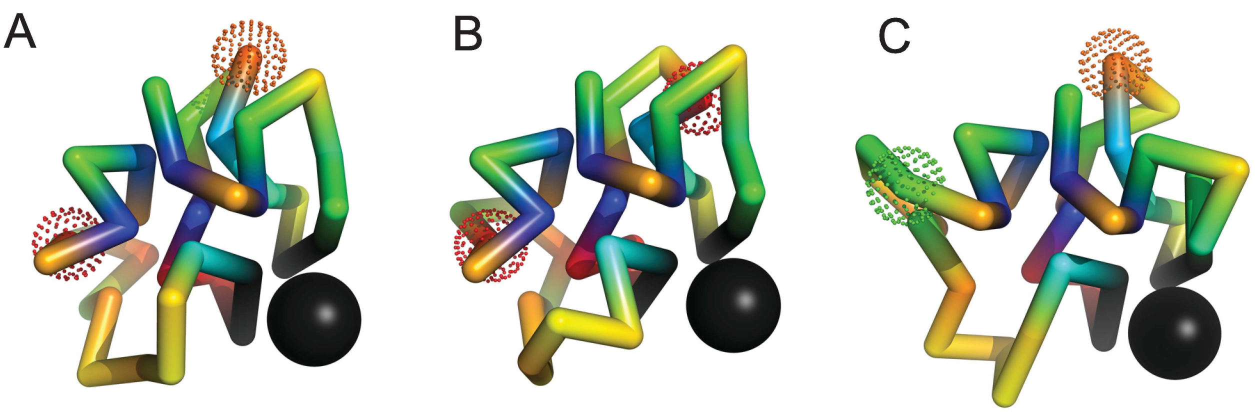

Fig 10. An example of weak but very long–range negative epistasis with $\epsilon \simeq -0.11$. Individual mutations Q25 (orange) → E25 (red) and E6 (red) → G6 (green) in panels (B) and (C) are slightly deleterious and neutral, with $\Delta P_1 \simeq -0.11$ and $\Delta P_2 \simeq 0$ respectively ($\Delta P_{12} \simeq -0.22$). E6 is ordered (Q25 is disordered) in the initial ensemble, and E25 is ordered (G6 is disordered) in the single mutant ensembles. In the double mutant, the removal of E6 allows E25 to form favorable contacts with other charged monomers during folding, which leads to more frequent mis–folding of the binding site, frustrating the exchange of material between ordered and disordered phases. The distance between mutated positions in panel (A) is $\mathrm{R} \simeq 12.5$ Angstroms. Folded ensembles for the initial and mutated sequences are shown in S8 Fig.

doi:10.1371/journal.pone.0166739.g010

Of course, our model is rather small, and neglects many of the constraints present in folded proteins, such as secondary structure. While this is not an obstacle for modeling short proteins (which are commonly devoid of significant secondary structure), secondary structure provides additional stiffness in larger proteins, which can lead to longer range collective effects such as allostery [26], and more exotic mechanisms for epistasis. Accordingly, the structural changes exhibited by our model are probably limited to more flexible regions in larger proteins.

Methods

The polymer model is a chain of point monomers that interact as low resolution amino acids via spherically symmetric potentials. Interactions along the chain are described by potentials of the form,

$$U^{\kappa}(r) \;=\; \frac{\kappa}{2}\,(r-l)^2 \tag{2}$$

where r is the distance between monomers, l is the equilibrium length of a link, and κ is a constant (see below).

Interactions between non–adjacent monomers along the chain are constructed from the unit Morse potential,

$$\mu(r) \;=\; \exp\left(-2\alpha(r-l)\right) \;-\; 2\exp\left(-\alpha(r-l)\right) \tag{3}$$

The attractive minimum of the Morse potential occurs at $r = l$. Let

$$\mu^{r\leq l}(r) \;=\; \vartheta(l-r)\,\mu(r) \tag{4}$$

and

$$\mu^{r\geq l}(r) \;=\; \vartheta(r-l)\,\mu(r) \tag{5}$$

denote the components of the Morse potential in either side of the minimum, where ϑ is the unit step function. The potentials for attractive and repulsive amino acid interactions are

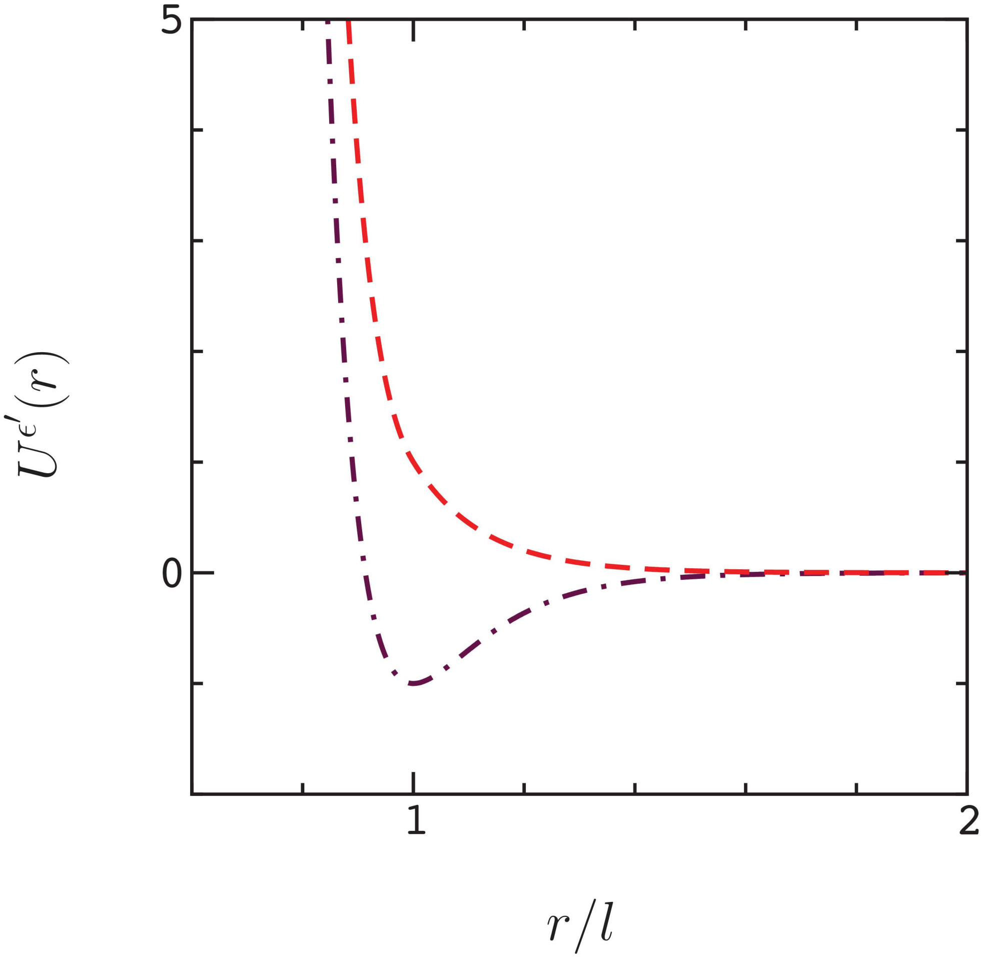

Fig 11. Potential functions, $U^{\epsilon'}(r)$, for cross-chain interactions at unit core strength, $\epsilon = 1$, unit attraction, $\epsilon' = -1$ (dot–dashed line) and unit repulsion, $\epsilon' = 1$ (dashed line).

doi:10.1371/journal.pone.0166739.g011

constructed as,

$$U^{\epsilon' \leq 0}(r) = \epsilon\, \mu^{r \leq l}(r) + (\epsilon + \epsilon')\, \vartheta(l - r) - \epsilon'\, \mu^{r \geq l}(r) \tag{6}$$

and

$$U^{\epsilon' \geq 0}(r) = \epsilon\, \mu^{r \leq l}(r) + \epsilon\, \vartheta(l - r) + \epsilon' \exp\left(-\alpha(r - l)\right) \tag{7}$$

respectively (Fig 11).

Each potential consists of an excluded volume part, $\epsilon\mu^{r \leq l}(r) + \epsilon\vartheta(l - r)$, modulated by the parameter ϵ, and a sequence dependent part, modulated by the parameter ϵ'; The parameter ϵ' takes on different values,

$$\epsilon' = \epsilon\, E_{\mu\nu}/E_{\rm o} \tag{8}$$

depending on the amino acid types involved in an interaction, where $E_{\mu\nu}$ is the energy of a contact between amino acids μ and ν defined by the empirical parameters in reference [27], and $E_{\rm o} = \langle |E_{\mu\nu \geq \mu}| \rangle$ is the average strength of an interaction (the empirical parameters are obtained by re–scaling the Miyazawa–Jernigan parameters [28] using threonine as a reference solvent [27]). The potentials for unit strength attractive and repulsive interactions are plotted in Fig 11.

To describe polymer kinetics, we integrate the Langevin equation using the method of van Gunsteren and Berendsen [29], with monomer mass $m = 1.66{\cdot}10^{-22}$ g, friction coefficient $\gamma = 10$ ps^{-1}, and integration time step, $\Delta t = 0.01$ ps. The parameters used to define the potentials are $l = 3.8$ Angstroms, $\kappa = 11\ k_B\ T_0$, $\alpha = 2.1$ Angstroms^{-1}, and $\epsilon = 2\ k_B\ T_0$, where k_B is Boltzmann's constant and $T_0 = 302.15$ Kelvin.

Folding is initiated from a random coil state below the folding transition temperature of a typical evolved sequence, which we estimate as $T_f \sim 1.25T_0$ from specific heat data. The time allowed for folding is determined by the length of the polymer according to the estimate of Lin and Zewail [21],

$$t_f = N\left(\frac{3}{e}\right)^N \Delta t_f \tag{9}$$

where $\Delta t_f = 10$ ps roughly describes the timescale for positional exchanges among monomers on the surfaces of polymer nuclei. Following this step, the replicas are equilibrated for a short time $t_q = t_f/3$ at temperatures $T_1 = 218.2$ Kelvin and $T_2 = 134.3$ Kelvin.

To determine the number of active replicas in the folded ensemble, it is necessary to dock the model ligand onto the binding site structures recovered by replicas in the folding procedure. To accomplish this, the folded structure of a replica is enclosed in a spherical shell consisting of $\sim 10^4$ evenly distributed points [30]. We then measure, and record the energy of the target ligand (in this work, a single monomer) at each point on this shell. In this procedure, interactions with monomers in the binding site group are considered attractive, and are described by unit Morse potential, $\mu(r)$, while interactions with monomers not included in the binding site group are described by the repulsive core of the Morse potential, $\mu^{r \leq l}(r)$. The radius of the shell is reduced, and the energies are re–computed at each point, iteratively, until the shell lies inside the folded structure of the replica. The structure of the binding site complex (i.e. binding site plus ligand) is determined from this sweep as the configuration with minimal energy. A docked replica is considered active when the distances between amino acids in the binding complex are each within 1 Angstrom of their corresponding distances in a pre–defined target state, determined by averaging the states of replicas with properly formed binding sites recovered by a selected initial (evolved) sequence. The structures used to represent the dominant energy basin recovered by a sequence are obtained using the structural alignment procedure described in ref. [14].

Supporting Information

S1 Fig. (A) Folded structure $\mathbf{x}^\star$, and (B) sample of the folded ensemble $\Delta\Gamma^\star$ recovered by a selected sequence evolved under ligand binding conditions. Amino acids (monomers) are colored blue, light blue, blue–green, green, yellow, orange, and red, in order of increasing affinity to solvent. The binding site monomers and the target ligand (here, a single monomer) are colored black. The ensemble $\Delta\Gamma^\star$ is aligned to $\mathbf{x}^\star$ using methods described in ref. [14].
(TIFF)

S2 Fig. (A) Folded structure, and (B) sample of the folded ensemble recovered by a second evolved sequence. Panels of the figure are described as in S1 Fig.
(TIFF)

S3 Fig. Amino acid exchange probabilities, $p(\mu, \nu) = A_{\mu\nu} / \Sigma_\nu A_{\mu\nu}$, for evolved sequences in ref. [14], where $A_{\mu\nu}$ is the number of transitions recorded between amino acids μ and ν.

Model values are indicated by filled red circles. Empirical values obtained from the data of Dayhoff et. al [24] are indicated by open blue circles. The value of $p(\mu, \nu)$ is indicated by the radius of the corresponding circle. In random sampling of pair mutations, amino acid transitions are allowed when $p(\mu, \nu)>0$.
(TIFF)

S4 Fig. Folded ensembles of initial and mutated sequences corresponding to Fig 6. Panel (U) corresponds to the initial, un–mutated sequence. Panels (1) and (2) correspond to the single mutants R1 (orange) → G1 (green) and T12 (yellow) → I12 (light blue), respectively. Panel (12) corresponds to the double mutant. Dotted spheres indicate the positions of mutated amino acids. Each ensemble $\Delta\Gamma^{\star}$ is aligned to its corresponding reference fold, $\mathbf{x}^{\star}$, using methods described in ref. [14] For clarity, each figure panel includes the 30 closest structures to $\mathbf{x}^{\star}$. Ensembles are rotated to reveal the positions of mutated monomers.
(TIFF)

S5 Fig. Folded ensembles of initial and mutated sequences corresponding to Fig 7. Panel (U) corresponds to the initial, un–mutated sequence. Panels (1) and (2) correspond to the single mutants W4 (blue) → S4 (yellow) and G10 (green) → V10 (blue), respectively. Panel (12) corresponds to the double mutant. Ensembles are arranged as described in S5 Fig.
(TIFF)

S6 Fig. Folded ensembles of initial and mutated sequences corresponding to Fig 8. Panel (U) corresponds to the initial, un–mutated sequence. Panels (1) and (2) correspond to the single mutants E6 (red) → G6 (green) and A23 (green) → E23 (red), respectively. Panel (12) corresponds to the double mutant. Ensembles are arranged as described in S5 Fig.
(TIFF)

S7 Fig. Folded ensembles of initial and mutated sequences corresponding to Fig 9. Panel (U) corresponds to the initial, un–mutated sequence. Panels (1) and (2) correspond to the single mutants T15 (yellow) → R15 (orange) and W4 (blue) → C4 (blue), respectively. Panel (12) corresponds to the double mutant. Ensembles are arranged as described in S5 Fig.
(TIFF)

S8 Fig. Folded ensembles of initial and mutated sequences corresponding to Fig 10. Panel (U) corresponds to the initial, un–mutated sequence. Panels (1) and (2) correspond to the single mutants Q25 (orange) → E25 (red) and E6 (red) → G6 (green), respectively. Panel (12) corresponds to the double mutant. Ensembles are arranged as described in S5 Fig.
(TIFF)

S9 Fig. An example of long–range negative epistasis with $\epsilon \simeq -0.26$. Individual mutations N13 (yellow) → D13 (red) and D19 (red) → Y19 (blue–green) in panels (B) and (C) are nearly neutral, with $\Delta P_1 \simeq -0.07$ and $\Delta P_2 \simeq -0.1$ respectively ($\Delta P_{12} \simeq -0.43$). Both N13 and D19 are ordered in the initial ensemble, and both D13 and Y19 are ordered in the single mutant ensembles. Single mutations do not significantly alter the reference structure of the initial sequence. In the double mutant, the mutation, D19 → Y19, reduces the specificity of interactions with its neighbors, allowing for greater conformational freedom of the binding site, making it more susceptible to the negative effect of the mutation, N13 → D13; The mutated amino acids Y19 and D13 form favorable contacts during folding, which disrupts the binding site in the quenched ensemble. The distance between mutated positions in panel (A) is $\mathrm{R} \simeq 7.3$ Angstroms.
(TIFF)

S1 File. Text files of the data points and structures shown in Figs 1–10 and S1–S9 Figs. Structural ensembles are provided in Protein Data Bank (PDB) format. In each ensemble file, structures (models) are aligned to the reference fold (first model in a given file), and are arranged in order of decreasing alignment quality. Structures are aligned through the closest 2*N*/3 monomers as described in the Text.
(GZ)

Author Contributions

Conceptualization: EN NG.

Data curation: EN.

Formal analysis: EN NG.

Funding acquisition: NG.

Investigation: EN NG.

Methodology: EN.

Project administration: NG.

Resources: NG.

Software: EN.

Supervision: NG.

Validation: EN NG.

Visualization: EN.

Writing – original draft: EN.

Writing – review & editing: EN NG.

References

1. Starr TN, Thornton JW. Epistasis in protein evolution. Prot Sci. 2016; 25:1204–1218. doi: 10.1002/pro.2897
2. Olson CA, Wu NC, Sun R. A Comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr Biol. 2014; 24:2643–2651. doi: 10.1016/j.cub.2014.09.072 PMID: 25455030
3. L M, Golding GB, Dean AM. Pervasive Cryptic Epistasis in Molecular Evolution. PLoS Genet. 2010; 6: e1001162. doi: 10.1371/journal.pgen.1001162 PMID: 20975933
4. Poon AF, Chao L. The rate of compensatory mutation in the DNA bacteriophage *φ* X174. Genetics. 2005; 170:989–999. doi: 10.1534/genetics.104.039438 PMID: 15911582
5. Poon AF, Chao L. Functional origins of fitness effect–sizes of compensatory mutations in the DNA bacteriophage *φ* X174. Evolution. 2006; 60:2032–2043. PMID: 17133860
6. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. Crystal structure of an ancient protein: Evolution by conformational epistasis. Science. 2007; 317:1544–1548. doi: 10.1126/science.1142819 PMID: 17702911
7. Bridgham JT, Ortlund EA, Thornton JW. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature. 2009; 461:515–519. doi: 10.1038/nature08249 PMID: 19779450
8. Breen MS, Kremena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. Nature. 2012; 490:535–538. doi: 10.1038/nature11510 PMID: 23064225
9. McCandlish DM, Rajon E, Shah P, Ding Y, Plotkin JB. The role of epistasis in protein evolution. Nature. 2013; 497:E1–E2. doi: 10.1038/nature12219 PMID: 23719465

10. Breen MS, Kremena C, Vlasov PK, Notredame C, Kondrashov FA. Breen et al. reply. Nature. 2013; 497:E2–E3.

11. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci USA. 2012; 109:E1352–E1359. doi: 10.1073/pnas.1120084109 PMID: 22547823

12. Ashenberg O, Gong LI, Bloom JD. Mutational effects on stability are largely conserved during protein evolution. Proc Natl Acad Sci USA. 2013; 110:21071–21076. doi: 10.1073/pnas.1314781111 PMID: 24324165

13. Pollock DD, Goldstein RA. Strong evidence for protein epistasis, weak evidence against it. Proc Natl Acad Sci USA. 2014; 111:E1450. doi: 10.1073/pnas.1401112111 PMID: 24706894

14. Nelson ED, Grishin NV. Evolution of off–lattice model proteins under ligand binding constraints. Phys Rev E. 2016; 94:022410. doi: 10.1103/PhysRevE.94.022410 PMID: 27627338

15. Hormoz S. Amino acid composition of proteins reduces deleterious impact of mutations. Sci Rep. 2013; 3:2919. doi: 10.1038/srep02919 PMID: 24108121

16. Ramsey DC, Scherrer MP, Zhou T, Wilke CO. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics. 2011; 188:479–488. doi: 10.1534/genetics.111.128025 PMID: 21467571

17. Huang T, Marcos M, Hwang J, Echave J. A mechanistic stress model of protein evolution accounts for site–specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol Biol. 2014; 14:78. doi: 10.1186/1471-2148-14-78 PMID: 24716445

18. Jack BR, Meyer AG, Echave J, Wilke CO. Functional Sites Induce Long–Range Evolutionary Constraints in Enzymes. PLoS Biol. 2016; 14:e1002452. doi: 10.1371/journal.pbio.1002452 PMID: 27138088

19. Illergard K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. Proteins. 2009; 77:499–508. doi: 10.1002/prot.22458 PMID: 19507241

20. England JL. Allostery in protein domains reflects a balance of steric and hydrophobic effects. Structure. 2011; 19:967–975. doi: 10.1016/j.str.2011.04.009 PMID: 21742263

21. Lin MM, Zewail AH. Hydrophobic forces and the length limit of foldable protein domains. Proc Natl Acad Sci USA. 2012; 109:9851–9856. doi: 10.1073/pnas.1207382109 PMID: 22665780

22. Zhou Y, Vitkup D, Karplus M. Native proteins are surface–molten solids: application of the Lindemann criterion for the solid versus liquid state. J Mol Biol. 1999; 285:1371–1375. doi: 10.1006/jmbi.1998.2374 PMID: 9917381

23. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol. 2002; 324:105–121. PMID: 12421562

24. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Atlas of protein sequence and structure. Washington, DC: Nat. Biomed. Res. Found.; 1972. p. 345–352.

25. Fodor AA, Aldrich RW. On evolutionary conservation of thermodynamic coupling in proteins. J Biol Chem. 2004; 279:19046–19050. doi: 10.1074/jbc.M402560200 PMID: 15023994

26. Reynolds KA, McLaughlin RN, Ranganathan R. Hot spots for allosteric regulation on protein surfaces. Cell. 2011; 147:1564–1575. doi: 10.1016/j.cell.2011.10.049 PMID: 22196731

27. Betancourt MR, Thirumalai D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Prot Sci. 1999; 8:361–369. doi: 10.1110/ps.8.2.361 PMID: 10048329

28. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol. 1996; 256:623–644. doi: 10.1006/jmbi.1996.0114 PMID: 8604144

29. van Gunsteren WF, Berendsen HJC. Algorithms for brownian dynamics. Mol Phys. 1982; 45:637–647. doi: 10.1080/00268978200100491

30. Tegamark M. An icosahedron–based method for pixelizing the celestial sphere. ApJ Lett. 1996; 470: L81.