Selection for cooperativity causes epistasis predominately between native contacts and enables epistasis-based structure reconstruction

R. Charlotte Eccleston[1], David D. Pollock[2], and Richard A. Goldstein[1]

[1]Division of Infection and Immunity, University College London, London WC1E 6BT, UK; [2]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA

This manuscript was compiled on February 6, 2021

Epistasis and cooperativity of folding both result from networks of energetic interactions in proteins. Epistasis results from energetic interactions among mutants, whereas cooperativity results from energetic interactions during folding that reduce the presence of intermediate states. The two concepts seem intuitively related, but it is unknown how they are related, particularly in terms of selection. To investigate their relationship, we simulated protein evolution under selection for cooperativity and separately under selection for epistasis. Strong selection for cooperativity created strong epistasis between contacts in the native structure, but weakened epistasis between non-native contacts. In contrast, selection for epistasis increased epistasis in both native and non-native contacts, and reduced cooperativity. Because epistasis can be used to predict protein structure only if it preferentially occurs in native contacts, this result indicates that selection for cooperativity may be key for predicting structure using epistasis. To evaluate this inference, we simulated the evolution of Guanine nucleotide-binding protein (GB1) with and without cooperativity. With cooperativity, strong epistatic interactions clearly map out the native GB1 structure, whilst allowing the presence of intermediate states (low cooperativity) obscured the structure. This indicates that using epistasis measurements to reconstruct protein structure may be inappropriate for proteins with stable intermediates.

Protein folding | Protein structure prediction

Two mutations have an epistatic interaction if their combined effect on a trait is not equal to the sum of their independent effects (1). The effect may be on fitness, function, or a physical property such as stability. Epistasis has been demonstrated many times experimentally. It has been found to impact the rate of adaptation (2), to constrain mutational trajectories leading to drug resistance (3, 5), and to impact yeast metabolism (4). It has been observed in the evolution of influenza (6, 7), between beneficial mutations in an evolving population of Escherichia coli (8), during the evolution of RNA viruses (9), and in the evolution of new enzyme activity (10, 11). Epistasis influences the amino acid preferences at different sites (12) and can have a substantial impact on protein evolution by restricting certain evolutionary pathways and by opening up new ones, resulting in sequences and functions that were not previously available (13). It has been suggested that epistasis is highly pervasive, affecting up to 90 per cent of substitutions (14).

Experimentally measured epistasis can be used to predict the 3D native structure of a protein. For example, Olson et al. (2014) (15) measured the epistasis between the majority of possible residue pairs of the Guanine nucleotide-binding protein (GB1) protein, which was used by Rollins et al. (2019) (16) to predict the protein's 3D structure. Such prediction methods assume that the majority of epistatic pairs are in contact in the native state, an assumption supported by experimental evidence (15). In the native state structure, the side chains of residues in contact interact, and so they no longer behave independently. This can result in non-additivity in terms of protein properties such as stability. However, native contacts are not the only interactions that determine protein properties. Mutations in contacts present in intermediate states and unfolded state structures that alter the stability of those states relative to the native state will impact properties such as stability. It is therefore unclear why experimental evidence suggests that mostly native contacts interact epistatically.

Cooperativity in protein folding

Proteins are under evolutionary pressures to fold and unfold cooperatively (17), where breaking a small number of interactions leads to complete unfolding. When proteins fold cooperatively, they move from the unfolded to the folded state, avoiding intermediate state. The disadvantage of stable intermediate states is that they are prone to aggregation and can lead to mis-folding, which is known to play a role in many diseases, including amyloid diseases such as Alzheimer's and Parkinson's (18–20). Many small, single domain proteins, for example, display highly cooperative two-state folding (21, 22), in which only the native and fully unfolded states are occupied, due to

Significance Statement

We investigated the relationship between cooperativity and epistasis and found low cooperativity results in high epistasis between non-native contacts, whilst high cooperatively results in epistasis mainly between native contacts. This provides a mechanistic explanation for why epistasis measurements can be used to reconstruct protein structure. The structure of GB1 protein has been successfully reconstructed using epistasis measurements and we calculated its epistasis distribution for a cooperative and non-cooperative model. The structure of the native state is clearly mapped out in the cooperative model, but becomes obscured in the non-cooperative model, due to the presence of a folding intermediate. We thus conclude that using epistasis measurements to reconstruct the native state of proteins with stable intermediates may not be appropriate.

R.G, D.P and R.C.E designed the research, R.C.E carried out the research and wrote the paper

The authors declare no conflict of interest

[2]To whom correspondence should be addressed. E-mail: r.goldstein@ucl.ac.uk

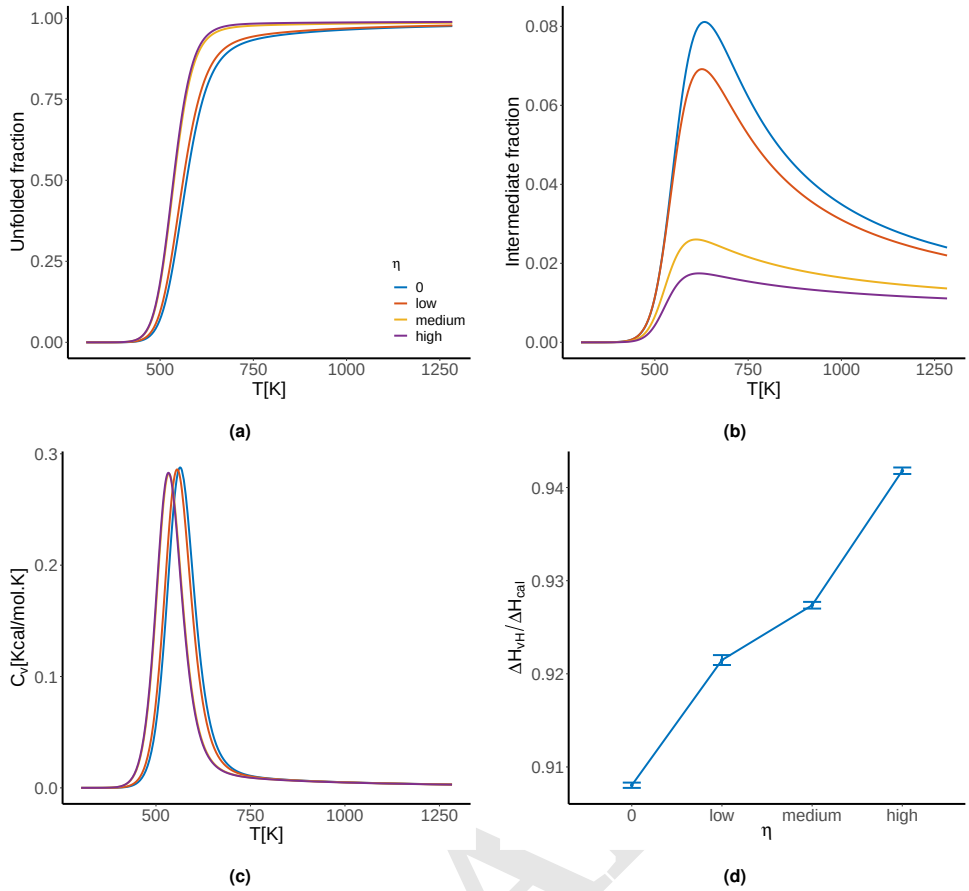


Fig. 1. Investigating the cooperativity of sequences evolved under zero, low, medium and high values of the tuning coefficient η by considering, (a) the fraction of the system found in the fully unfolded state during denaturation under increased temperature, (b) the fraction in the ensemble of intermediate states during the unfolding transition, (c) heat capacity curves during the unfolding transition, the area under which is the enthalpy change associated with the transition, and (d) the van't Hoff ratio of the unfolded transition associated with each value of the cooperativity tuning coefficient η. All values are averaged over the set of 1000 most evolved sequences for each evolutionary simulation.

the instability of any intermediate states. In contrast, larger, multi-domain proteins, often fold stepwise via the formation of partially unfolded forms (PUFs), where each PUF is made up of one or more cooperative structural units known as foldons (19). Cooperativity of folding is also observed in macromolecular complexes, and strong co-evolutionary preferences have been observed between cooperative proteins composing part of a macromolecular complex, where the components display a conserved self-assembly order (23).

Cooperative folding requires the presence of unfavourable destabilizing interactions at structurally important sites in partially folded states, and/or highly favourable interactions that stabilize the native state, whilst not over-stabilizing those intermediate states in which the stabilised native contact is present. This was demonstrated by Yadahaldi and Gosavi when the designed non-cooperative protein Top7 was made to fold cooperatively by introducing stabilising mutations at a set of native contacts and destabilising mutations at residue pairs that were found to stabilise intermediate states (24).

Cooperativity and epistasis thus both involve sometimes strong interactions among adjacent amino acid residues in the native structure. It seems possible that selection for one might drive the other, or visa versa, but how they influence each other is unknown. We chose to investigate this by simulating protein evolution using a mechanistic model based in thermodynamics and statistical mechanics, that has been shown to be able to reproduce many important features of protein evolution such as epistasis and co-evolution (12, 29). We evolved a protein under different levels of selection for cooperativity to explore how and why epistasis differs between cooperative and non-cooperative sequences.

To investigate how selection for cooperativity impacts 3D structure reconstruction using epistasis data, we simulated the evolution of the GB1 protein for a two-state (containing native and unfolded states) and three-state (containing native, unfolded and intermediate states) model and determined the distribution of epistasis between all pairs of residues.

Results

We performed 10 evolutionary simulations for 50,000 generations of a protein sequence based on the structure of a cysteine-free variant of *Escherichia coli* ribonuclease H (RNase H). For these simulations we calculated the fitness based on the probability that a protein would be in its native state at thermal equilibrium. We also included a fitness penalty that reduced the fitness of proteins with folding intermediates, allowing us to tune the impact of this penalty using a cooperativity tuning coefficient, η. The folding pathway of RNase H has been determined at near amino acid resolution (**?**). We generated a series of intermediate partially-folded states based on the

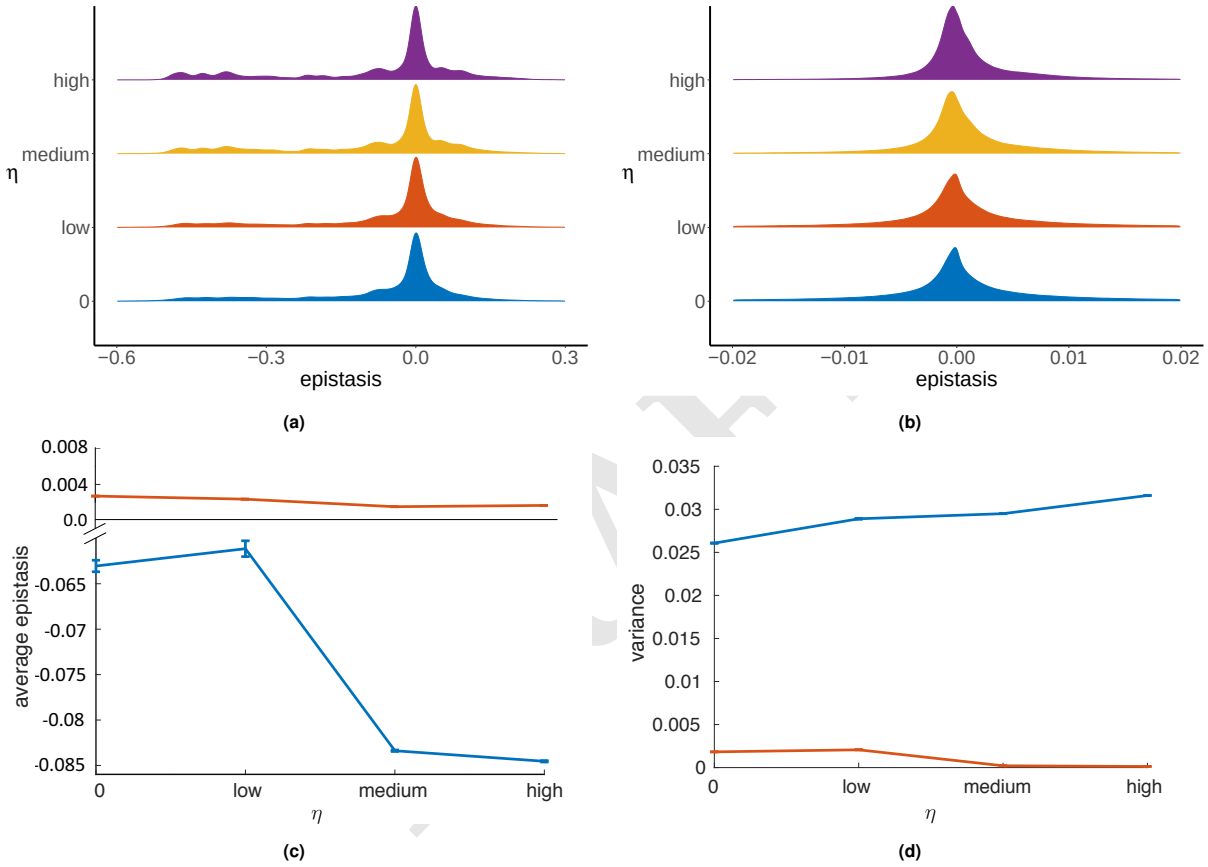


Fig. 2. The distribution of epistasis as the selection coefficient η increases. The normalised distribution of the epistasis in protein stability between a) native contacts and b) non-native contacts when evolving proteins under varying degrees of selection for cooperativity. As selection for cooperativity increases, more native contacts experience higher magnitude (more negative) epistasis, whilst more non-native contacts experience very low levels of epistasis. The area under each curve sums to 1. c) The mean of the epistasis distributions, and d) the variance of the epistasis distribution, of the final 2,000 generations of the 50,000 generations simulated, averaged over all 10 simulations for native contacts (blue) and non-native contacts (red). The error bars represent the variance of these values across the 10 simulations. The average of the epistasis distribution at the native contacts becomes more negative as the value of the selection coefficient η increases and the variance in the distribution increases. The average of the epistasis distribution at the non-native contacts goes to zero as the selection coefficient η increases, and the variance decreases.

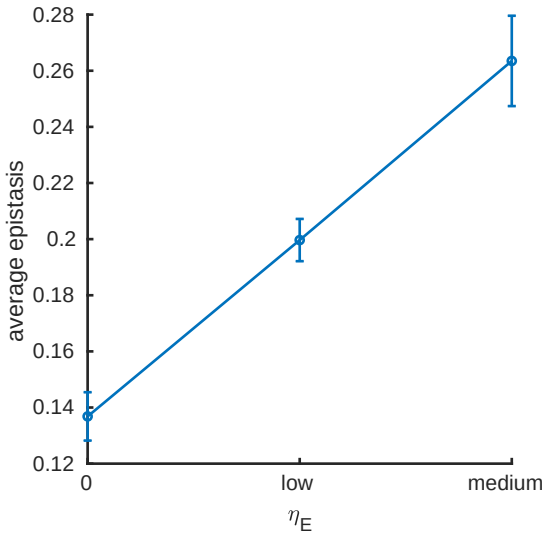


Fig. 3. The mean absolute epistasis in protein stability (y-axis), averaged over all 10 simulations, between each pair of native contacts when evolving proteins under increasing selection (η_E = 0, low, medium) for the absolute epistasis in protein stability at the native contacts (x-axis). The error bars depict the variance in the mean. The average absolute epistasis between native contacts increases as the value of the selection coefficient η_E increases.

step-wise folding pathway, in which the folded regions of the proteins were fixed to their position in the folded structure and the unfolded regions were modelled as a freely joined chain defined by the position of the C_β atoms, with bond lengths between $3-7$Å. We also included an excluded volume term prohibiting C_β atoms from being closer than 3Å (see Supplementary Information for more detail).

We carried out simulations for four different values of the cooperativity tuning coefficient η: no selection for cooperativity ($\eta = 0$), low ($\eta = 5 \times 10^{-7}$), medium ($\eta = 5 \times 10^{-6}$) and high ($\eta = 1 \times 10^{-5}$) (Eq. 7).

Two-state folding generally results in sharp sigmoidal melting curves and a peak in the heat capacity at the melting temperature T_m, although multi-state transitions can also show such behaviour (30, 31). The level of cooperativity is determined experimentally by calculating the ratio κ of the van't Hoff enthalpy change ΔH_{vH} evaluated at T_m to the calorimetric enthalpy change ΔH_{cal} of the entire transition (32, 33). The van't Hoff enthalpy change is calculated purely from the difference in the enthalpy of the native and unfolded states, whilst the calorimetric enthalpy change is the experimentally measured enthalpy change during the unfolding transition. If the system is purely two-state, the calorimetric enthalpy change is equal to the difference between the enthalpies of the native and unfolded state, and so the ratio $\kappa = \Delta H_{vH}/\Delta H_{cal}$ equals 1. Values of $\kappa \approx 1$ are observed for many globular proteins (34–36). For folding simulations where the distribution of the protein states is available, we can directly distinguish two-state folding by examining the underlying populations of intermediate states during the folding transition. In this case lower occupation of intermediates indicates higher levels of cooperativity.

Multiple lines of evidence indicate that our selection for cooperativity is effective in increasing the cooperativity of the folding transition in our simulations. Firstly, the sharpness of the sigmoidal melting curves increases as the value of the cooperativity tuning coefficient increases (Fig. 1a). Secondly, the value of the van't Hoff criterion κ increases with selection for cooperativity from $\kappa \approx 0.91$ in the absence of selection for cooperativity, to $\kappa \approx 0.94$ for high selection (Fig. 1d). Finally, if we consider the total fraction of the population occupying the intermediate states (i.e. the fraction of the population not in either the native or fully unfolded states), which shows that as selection for cooperativity increases, the fraction in the intermediate states decreases (Fig. 1b).

Selection for cooperativity causes epistasis to increase between native contacts but decrease between non-native contact pairs. We then calculated the epistasis in protein stability (Eq. 15) between each possible pair of residues in the protein for the final 2,000 generations of the 50,000 generations simulated, and calculated the mean epistasis between each pair of residues averaged over all simulations, for the different values of selection for cooperativity. We investigated the distribution of epistasis between pairs of residues in contact in the native state (Fig. 2a) and pairs of residues not in contact in the native state (Fig. 2b). The sign convention we adopted for defining stability is in the direction of folding (Eq. 5), and so negative epistasis, for example, occurs when wild-type residues at positions i and j mutually stabilize each other compared to the mutant "non-interacting" residues.

As selection for cooperativity increases, the epistasis distribution between native contacts becomes less peaked around zero and the average of the distribution becomes more negative, whilst the variance of the distribution increases (blue line on Fig. 2c & 2d respectively).

In contrast, for the non-native contacts the average epistasis goes towards zero and the variance decreases. In other words, the more cooperative sequences display higher magnitudes of negative epistasis between pairs of native contacts, but smaller magnitudes of epistasis between the non-native pairs compared with sequences associated with lower cooperativity in protein folding.

Selection for epistasis at native contacts leads to a decrease in cooperativity. If cooperativity increases epistasis at native contacts, is the converse true? As a thought experiment, we investigated this question by directly selecting for epistasis between native contacts, though we do not expect this sort of selection in nature. The coefficient η_E increases selection for sequences with large epistasis at native contacts (see Eq. 8). We performed 10 evolutionary simulations for three values of the tuning coefficient η_E: no selection ($\eta_E = zero$), low ($\eta_E = 1 \times 10^{-7}$), and medium ($\eta_E = 1 \times 10^{-6}$), and determined the average epistasis between each pair of native contacts during the evolutionary process. Selecting for the average epistasis between native contacts was much more computationally expensive than selection for cooperativity, and therefore we chose to simulate evolution for just 5,000 generations. To enable a fair comparison between the epistasis distributions for selection for stability only ($\eta_E = 0$) and the two levels of selection for epistasis ($\eta_E = 1 \times 10^{-7}$ and 1×10^{-6}), we only considered the first 5,000 generations of the $\eta = 0$ simulations presented in the previous section. To determine the epistasis distributions for each value of η_E, we calculated the epistasis in protein stability (Eq. 15) between each possible pair of residues in the protein for the final 2,000

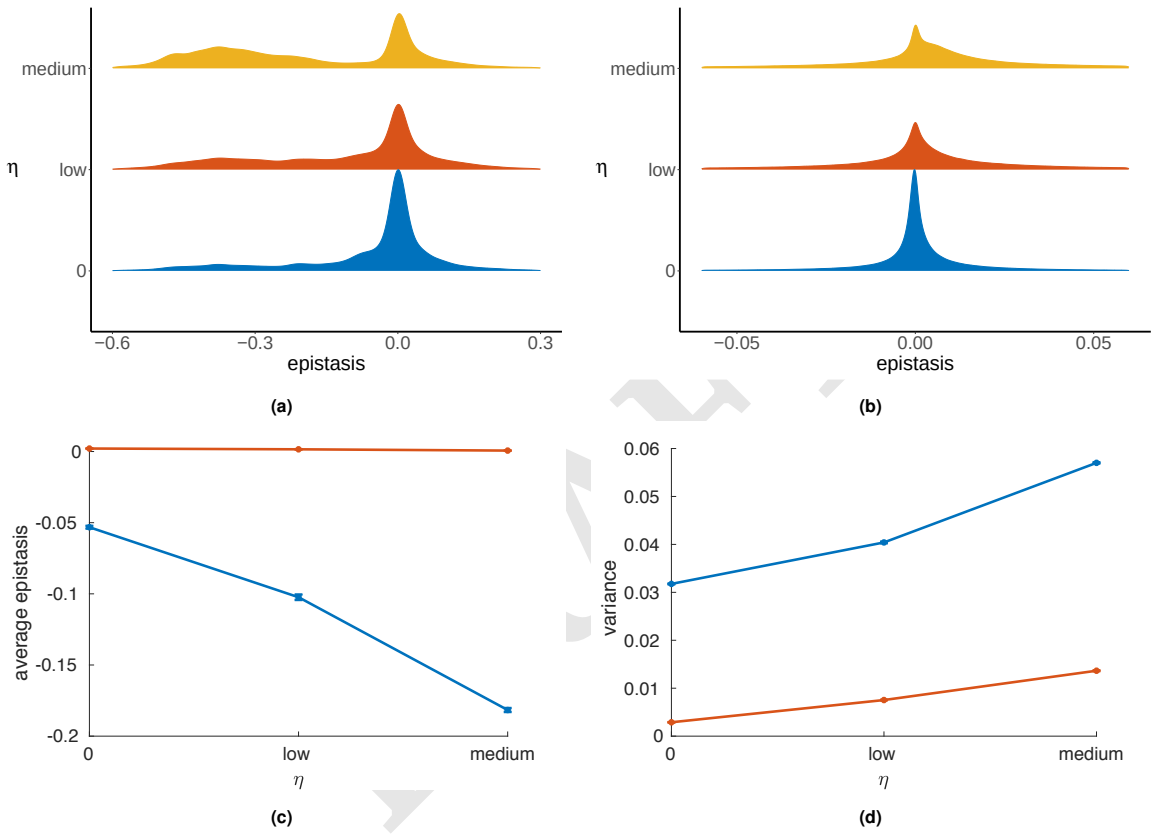
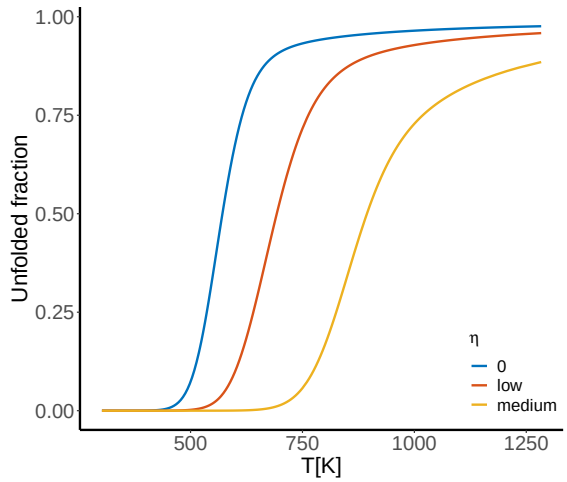
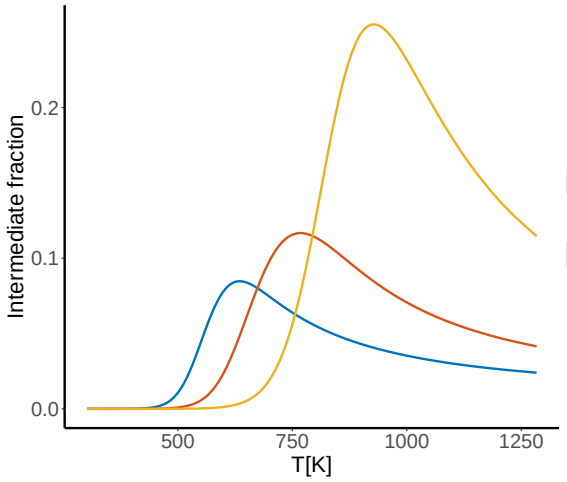


Fig. 4. The distribution of epistasis as the selection coefficient η_E increases. The normalised distributions of epistasis between a) native and b) non-native contacts when evolving sequences under different magnitudes of selection for the average magnitude of epistasis between the native contacts. As the value of the selection coefficient η_E increases, a higher number of native contacts experience greater magnitude negative epistasis, whilst a higher number of non-native contact pairs experience non-zero epistasis. The area under the curves sum to 1. c) The mean of the epistasis distributions, and d) the variance of the epistasis distribution, of the final 2,000 generations of the 5,000 generations simulated, averaged over all 10 simulations for native contacts (blue) and non-native contacts (red). The error bars represent the variance of these values across the 10 simulations. The average of the epistasis distribution at the native contacts becomes more negative as the value of the selection coefficient η_E increases and the variance in the distribution increases. The average of the epistasis distribution at the non-native contacts remains roughly constant as the selection coefficient η_E increases, but the variance in the distribution increases.



(a)



(b)

Fig. 5. Investigating the cooperativity of the unfolding transition with increased selection for the average magnitude of epistasis at the native contacts, by considering a) The fraction of the population in the unfolded state during denaturation and b) the fraction of the population in the ensemble of partially folded states during denaturation. As the value of the selection coefficient η_E increases, the unfolding transition becomes less sharp and the fraction of the population in the intermediate states increases, showing the folding is becoming less cooperative.

generations of the 5,000 generations simulated, and calculated the mean epistasis between each pair of residues, averaged over all simulations. As selection for epistasis increases, the average magnitude of epistasis per native contact per substitution increases (Fig. 3), demonstrating that the selection works as intended.

The effect on the distribution of mean epistasis among native contact pairs is similar to what was observed for cooperativity, but the effect is stronger (Fig. 4a). However, there was also more epistasis at non-native contact pairs, although epistasis between these pairs was not directly selected for (Fig. 4b). The average epistasis at native contacts becomes sharply more negative (blue line Fig. 4c), while for the non-native contacts the average is unchanged (red line Fig. 4c) but the variance, and thus the levels of both positive and negative epistasis increases (red line Fig. 4d).

We investigated the cooperativity of the evolved sequences via the protein's melting curves and the fraction of the system in the intermediate states during unfolding, because this is sufficient to determine cooperativity. Although we observed earlier that selection for cooperativity induces epistasis at native contacts, the inverse is not true. Instead, selection for epistasis at native contacts results in less cooperativity. The melting curve becomes less sharp and shifts to the right (Fig. 5a), indicating the protein passes through more stable intermediate states as it unfolds. The fraction of the ensemble of intermediate states also increases (Fig. 5b). Thus, although selecting for cooperativity induces epistasis at the native contacts, selecting for epistasis at the native contacts does not induce cooperativity, but instead decreases it.

The intermediate and unfolded ensemble approaches the unfolded state distribution for selection for cooperativity. To understand why selection for higher cooperativity increases epistasis between native contacts and decreases epistasis between non-native contacts, we considered how epistasis arises in the model, and how the stability of each state impacts our epistasis calculations. We can re-write Eq. 15, the epistasis between residues i and j, as $\epsilon_{i,j} = \epsilon^{NS}_{i,j} - \epsilon^{K,u}_{i,j}$, where $\epsilon^{NS}_{i,j} = G^{NS}_{ij} + G^{NS}_{WT} - G^{NS}_{i} - G^{NS}_{j}$, is the epistasis in the free energy of the native state, and $\epsilon^{K,u}_{i,j} = G^{K,u}_{ij} + G^{K,u}_{WT} - G^{K,u}_{i} - G^{K,u}_{j}$, is the epistasis in the free energy of the intermediate and unfolded ensemble, $\{K, u\}$, where $K = \{k\}$ denotes the k intermediate states and u denotes the unfolded state. For native contacts, the epistasis is determined by both the epistasis in the native state and the intermediate and unfolded ensemble, and whether epistasis is positive or negative is determined by a trade-off between the two values. For non-native contacts, the epistasis in the free energy of the native state, $\epsilon^{NS}_{i,j}$, is zero. Therefore positive epistasis at non-native contacts arises when $\epsilon^{K,u}_{i,j}$ is negative, and negative epistasis at the non-native contacts arises when $\epsilon^{K,u}_{i,j}$ is positive.

From Eq. 1 we can see that the epistasis between residues i and j in the free energy of a single structure is $\gamma(A_i, A_j)Q_{i,j}$, where $\gamma(A_i, A_j)$ is the contact potential between amino acids at residues i and j, and $Q_{i,j}$ is equal to 1 if residues are in contact and 0 otherwise. Therefore, the epistasis between two residues i and j is equal to the contact potential between the two amino acids if they are in contact in the native state, and zero otherwise..

The free energy of each state in the intermediate and

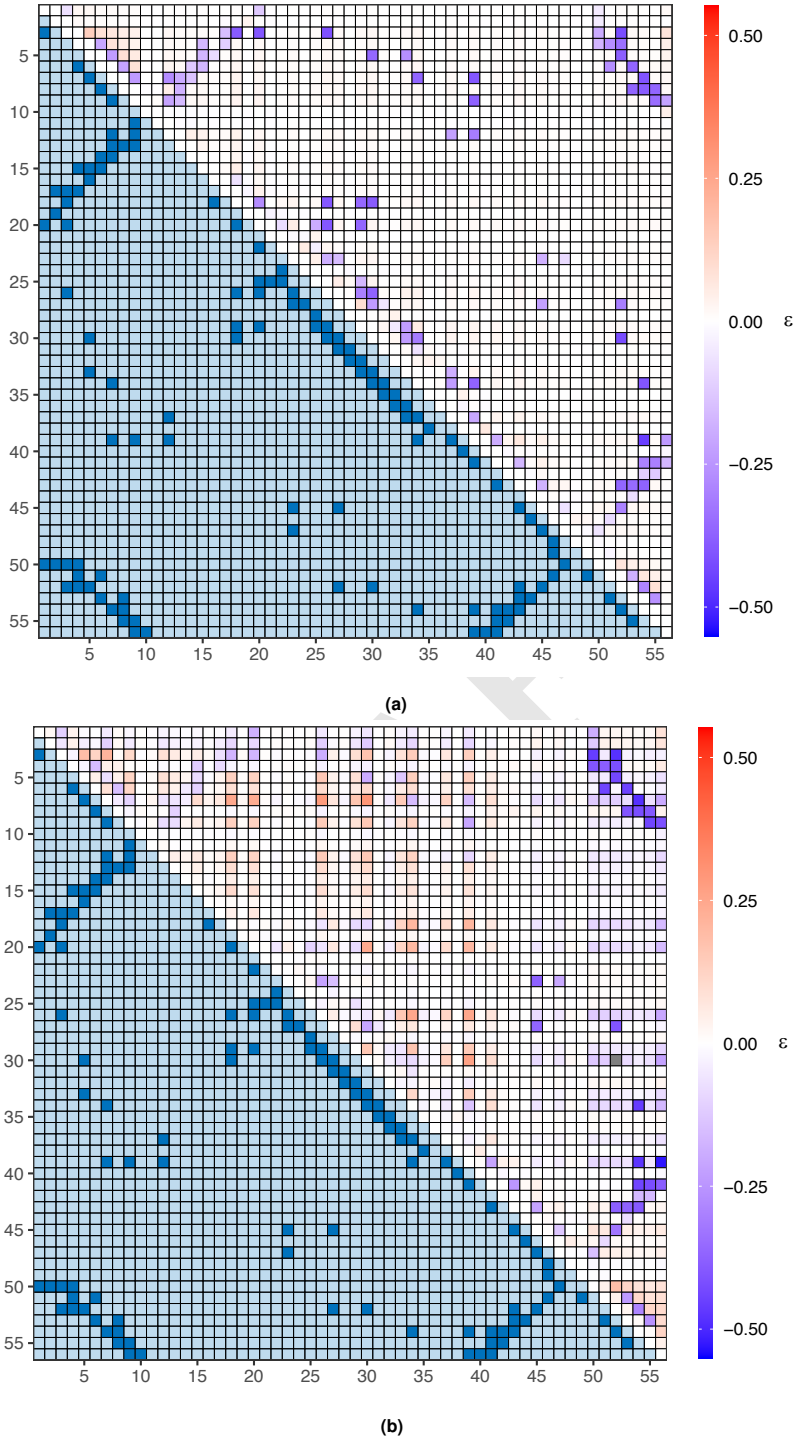


Fig. 6. The average epistasis ϵ between each possible pair of residues for a) a two-state and b) a 3-state system. The upper triangle of each heat map is the calculated epistasis, the lower triangle is the contact map for the GB1 protein, where a value of 1 means the residues are in contact in the native state. The epistasis for the two-state system accurately maps out the structure of the GB1 protein, with the majority of native contacts experiencing high magnitudes of negative epistasis. For the three state system, the native contact map becomes obscured by high levels of positive epistasis at non-native contacts. These results suggest epistatic measurements work well when reconstructing the native state of two-state systems, but are less successful for multi-state systems.

unfolded ensemble was determined using a large number of dummy structures. From Eq. 3 the epistasis between residues i and j in one of the intermediate states k, or the unfolded state u, is $\gamma(A_i, A_j)\langle Q_{i,j}\rangle_{k\vee u} - \langle Q_{i,j}\rangle_{k\vee u}(1 - \langle Q_{i,j}\rangle_{k\vee u})\gamma(A_i, A_j)/2k_BT$, where $\langle Q_{i,j}\rangle_{k\vee u}$ is the average probability of residues i and j being in contact in the ensemble of the chosen intermediate or unfolded state, denoted $k \vee u$.

Epistasis in the free energy of one of these states, between residues i and j, arises when a large fraction of dummy structures contain this contact, and so $\langle Q_{i,j}\rangle_{k\vee u}$ is large, resulting in changes to the average and variance of the free energy of the state in question. If a particular pair has a high probability of contact in several intermediate states, this can lead to epistasis in the free energy of the intermediate and unfolded ensemble.

To understand why epistasis between non-native contacts decreases as selection for cooperativity increases, we consider the distribution of the probability that residues i and j are in contact in the intermediate and unfolded ensemble, $\{K, u\}$ (Eq. 16). For one of the intermediate or unfolded states, the average probability residues i and j are in contact, $\langle Q_{i,j}\rangle_{k\vee u}$, will be a number between 0 and 1, i.e. it is the fraction of structures in the ensemble of state $k \vee u$ that contains the i-j contact. When selection for cooperativity is imposed, the intermediate states are destabilised and as selection increases the probability of being in any of the intermediate states goes to zero. This results in the distribution of contact probabilities becoming more concentrated around lower values (Fig. 7), demonstrating the contact probabilities of the intermediate and unfolded ensemble are becoming more like those of the unfolded state.

Because the probability any pair of residues i and j are in contact in the unfolded state is small, the corresponding epistasis in the intermediate and unfolded ensemble will be small. Therefore, as selection for cooperatively increases, the epistasis in the intermediate and unfolded ensemble decreases. Because the unfolded ensemble contains mostly non-native contacts, there is a decrease in epistasis at non-native contacts as selection for cooperativity increases. Similarly, given the equation for epistasis between residues i and j, $\epsilon_{i,j} = \epsilon_{i,j}^{NS} - \epsilon_{i,j}^{K,u}$, we can see that as $\epsilon_{i,j}^{K,u}$ goes to zero, for native contacts $\epsilon_{i,j} \approx \epsilon_{i,j}^{NS}$, explaining the increase in the magnitude of the epistasis between native contacts as cooperativity increases.

Sequences under selection for the average magnitude of epistasis between native contacts display broad epistasis distributions at both native and non-native contacts (Fig. 4). Under this selection regime, intermediate states are stabilised (Fig. 5b). This happens because selection for epistasis at native contacts selects for pairs of residues with large contact potentials since $\epsilon_{i,j}^{NS} = \gamma(A_i, A_j)Q_{i,j}$, and so those intermediate state ensembles containing native contacts will be stabilised. This results in a decrease in cooperativity and an increase in the variance in the epistasis between both native and non-native contacts.

If we again consider the distribution of contact probabilities in the partially folded and unfolded ensemble, we observe that as selection for epistasis at native contacts increases, the distribution of probabilities spreads out, with some pairs of residues having a contact probability between 0.8 and 1 (Fig. 8). This happens because some of the intermediate states, which are being stabilised relative to the unfolded state, have highly structured areas with contact probabilities of 1 or almost 1. In other words, the distribution of contact probabilities in the intermediate and unfolded ensemble are becoming more like the native state contact probabilities, and less like the unfolded state contact probabilities. As mentioned earlier, epistasis in the intermediate and unfolded ensemble arises when a particular pair has a high probability of contact in this ensemble. Therefore, the larger number of high probability contacts in the intermediate and unfolded ensembles suffices to explain the broader distribution of epistasis between non-native contacts when there is high selection for epistasis at native contacts.

The 3D structure of multi-state proteins cannot be predicted using epistasis. Methods for inferring 3D protein structure using measured epistasis rely on the assumption that the largest magnitude epistasis occurs between native contacts. In the previous section we observed the distribution of epistasis between non-native contact pairs became broader as the protein became less cooperative. Therefore, it is possible that native structure inference methods using epistasis measurements may not be suitable for proteins with stable intermediate states. To examine this hypothesis, we simulated the evolution of the GB1 domain of streptococcal protein G, (PDB ID 1PGA) for a cooperative system and a non-cooperative system. The cooperative system was comprised of the native and fully unfolded state, where the free energy of the unfolded state ensemble was approximated using a large number of dummy structures generated by a random coil model. The non-cooperative system had an additional ensemble of intermediate states in which beta sheets 3 and 4 (residues 40-56) were unstructured. The free energy of the intermediate state ensemble was approximated using the same method as the unfolded state ensemble. The systems were evolved under selection for stability alone, and so the fitness of the protein was determined exclusively by the fraction in the folded state.

We calculated the epistasis in protein stability between all pairs of residues for both the cooperative and non-cooperative system (Fig. 6a and 6b respectively), for 100 sequences over 10 runs and averaged for each pair. For the cooperative system high magnitudes of negative epistasis occurred almost exclusively at native contacts and, when compared with the known GB1 native structure, the epistasis accurately mapped out the structure to a high degree of accuracy. Many of the highly epistatic pairs predicted by the model correspond to the measured highly epistatic pairs used to reconstruct the 3D structure of GB1 by Rollins et al. (2019)(16).

For the non-cooperative system, however, the magnitude of the negative epistasis at the majority of the native contact pairs decreased. Some contacts continued to have large negative epistasis (e.g. 1-10, 40-56 and 50-56), but the overall structure is less evident. Furthermore, more contacts display strong positive epistasis compared to the cooperative system.

Discussion

We observed that selection for cooperativity in protein folding changes the distribution of epistasis in simulated proteins. Proteins with higher cooperativity were associated with more epistasis between native contacts and less epistasis between non-native contacts compared to less cooperative proteins. Conversely, we observed that selection for epistasis at native

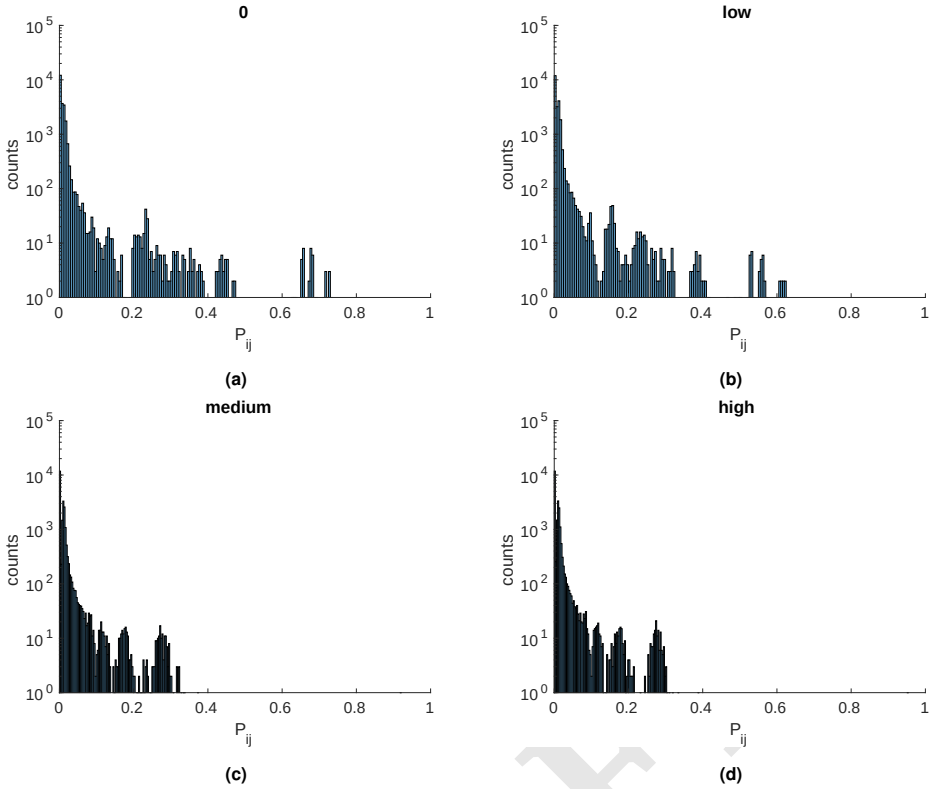


Fig. 7. Histogram of the distribution of contact probabilities $P_{i,j}$ between site i and site j when evolving proteins under selection for cooperativity, when η is set to a) 0, b) low, c) medium and d) high

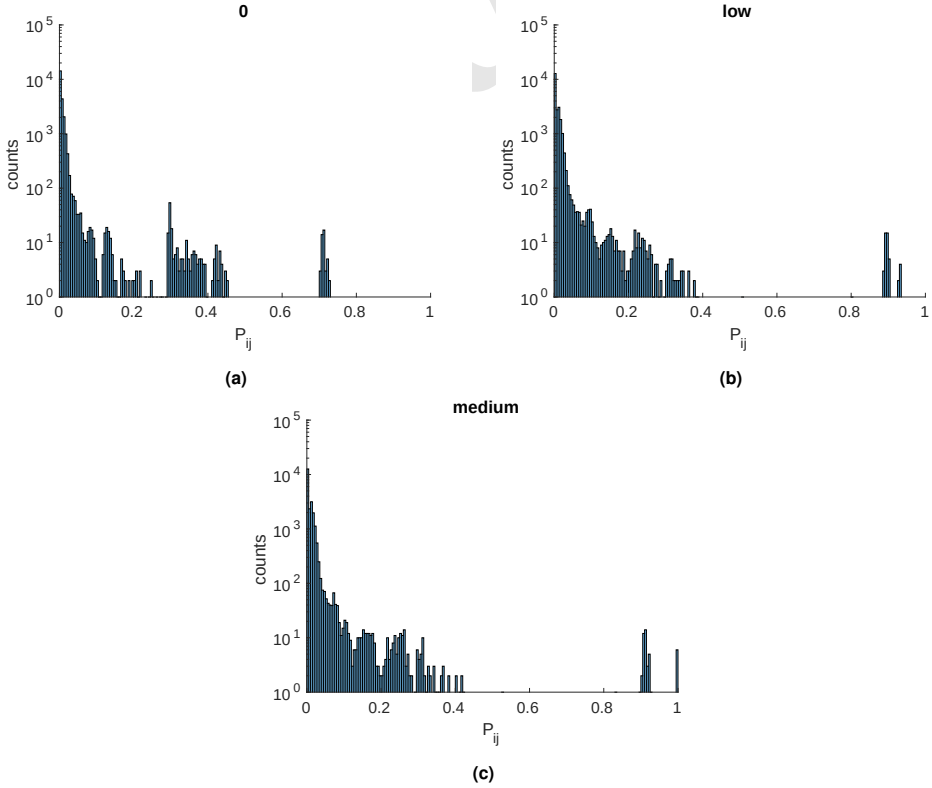


Fig. 8. Histogram of the distribution of contact probabilities $P_{i,j}$ between site i and site j when evolving proteins under selection for epistasis between native contacts, when η is set to a) 0, b) low and c) medium

contacts results in less cooperativity as selection increases.

This leads us to conclude that selection for cooperativity is not equivalent to selection for epistasis at native contacts, and suggests that high levels of epistasis at non-native contacts is detrimental to cooperative folding, and could lead to the aggregation of partially folded states. It is likely therefore that highly cooperative proteins will only display epistasis between native contacts. Because a large number of proteins fold cooperatively, these results provide a possible explanation for experimental observations that have found the majority of epistatic pairs to be native contacts.

We would thus expect that natural proteins with stable intermediates in their unfolding transition would display greater epistasis between non-native contacts than natural proteins that have two-state transitions. This suggests that the use of epistasis measurements to reconstruct the native state of these non-cooperative proteins, under the assumption that epistasis occurs only at native contacts may be problematic.

We gained further support for this theory by simulating the evolution of the GB1 protein for a cooperative and a non-cooperative system. The highest magnitude negative epistasis in the cooperative system occurred between native contact pairs and the pattern of high magnitude negative epistasis traced out the native structure well. The inclusion of an intermediate state in the non-cooperative system, however, reduced the magnitude of the negative epistasis between those native contacts present in the intermediate state, and introduced strong positive epistasis at non-native contacts.

The intermediate state contains the majority of the native state contacts, as only residues 40-56 are unfolded. These native contacts are in contact in 100% of the intermediate dummy structures, and so the probability of them being in contact in the unfolded and intermediate ensemble is high, meaning epistasis in the free energy of this ensemble of states for these native state contacts will be relatively large.

The large epistasis between these native contacts in the free energy of the unfolded and intermediate ensemble acts to partially cancel out the epistasis between these pairs in the native state ($\epsilon_{i,j} = \epsilon_{i,j}^{NS} - \epsilon_{i,j}^{K,u}$), resulting in lower magnitude epistasis for the native contacts contained in the intermediate state. As a result, it may be more difficult to infer the native state structure.

GB1 is a small protein and so it is unlikely to have intermediate states like the artificial one created for the purposes here. Therefore, it is likely that the structure of smaller proteins will be better inferred using measured epistasis than larger proteins that have folding intermediates.

Olson et al. (2014) noted, however, that positive epistasis occurred between a cluster of conformationally correlated residues. Otowinoski (2008) (**?**) sought to explain the epistasis observed by Olson et al (2014), using a two- and three-state model of protein-ligand binding, but neither model could explain the presence of the positive epistasis, and they suggested that a model including additional conformational states might capture this epistasis better. Therefore, even small proteins such as GB1 may have additional states or correlation in residue dynamics that might obscure prediction of the native state structure using measurements of epistasis.

Co-evolution between both native and non-native contact pairs may occur in non-cooperative proteins. For cooperative proteins, however, we expect that co-evolution occurs almost exclusively between pairs that interact in the native structure. It should be noted however, that whilst epistasis is pre-requisite to co-evolution, strong epistasis can prevent either site involved from changing and so there might be no observable co-evolution.

Furthermore, Sailer and Harms (2017) (**?**) investigated the predictability of evolutionary trajectories using a lattice protein model and found the presence of additional conformational ensembles in the model made evolution unpredictable. They observed pairwise epistasis in a two-state model and higher-order epistasis in a three-state model in the evolutionary trajectories of a small 12-amino acid protein. The pairwise epistasis in the two-state model was due to direct contact between residues, whilst higher-order epistasis in the three-state model resulted from the redistribution of the relative probabilities of structures in the ensemble. While we did not consider higher-order epistasis in this work, we did observe that the epistasis associated with non-native contacts was the result of epistasis in the free energies of the non-native ensembles, and that this epistasis was more prevalent in less cooperative proteins. Therefore, it is likely that we would observe prevalent higher-order epistasis in our model under in lower selection for cooperativity and little higher-order epistasis under higher selection for cooperativity.

Sailer and Harms also found that a pairwise model was able to perfectly predict evolutionary trajectories for two-state model but not the three-state model, and that predictions could not be improved even when including higher-order epistasis. Therefore, from their observations, we may hypothesize that it may be easier to use sequence data to predict protein structure for proteins that evolved under selection for cooperativity than for those that are not, due to the large number of intermediate ensembles.

J. Wells (1990) (37) remarked that the simple additive behaviour between many pairs of mutants is surprising given the highly cooperative nature of protein folding, but provides a few examples to the contrary where epistasis arises between contacting residues. We propose that it is *because* protein folding is highly cooperative that few residue pairs exhibit epistasis unless they are in contact in the native state.

Materials and Methods

Protein model. The free energy G of an amino acid sequence $\{A_1, A_2, ..., A_N\}$, where N is the length of the protein, in a specific structure can be calculated using a simple contact potential:

$$G = \sum_{i<j} \gamma(A_i, A_j) Q_{i,j}, \quad (1)$$

where $\gamma(A_i, A_j)$ is the contact potential between amino acids A_i and A_j in positions i and j respectively, determined by Miyazawa and Jernigan (38), and $Q_{i,j}$ is equal to one if residues i and j are in contact, and zero otherwise. Two amino acids are considered to be in contact if their C_β atoms (C_α in the case of glycine) are within 7Å of one another.

The free energy of the native state G_{NS} was calculated using the structure of a cysteine-free variant of *Escherichia coli* ribonuclease H (RNase H), a 155 residue mixed α/β protein (PDB designation 1F21), using Equation 1. The unfolded and intermediate states will each be associated with an ensemble of possible structures, and the free energy of each structure can be calculated using Equation 1. The number of possible structures within each ensemble is incredibly high, therefore an approximate to the distribution

of energies is required. We used a random coil model (39, 40) to produce random structures of sequences 152 amino acids long and obtained thousands of possible structures for each partially folded ensemble, $K = \{k\}$, where k denotes the individual intermediate states, and fully unfolded state u. For each intermediate or unfolded state, $k \vee u$, we used these structures to parameterise a Gaussian distribution with mean $\bar{G}_{k\vee u}$ and variance $\sigma^2_{k\vee u}$ for, to approximate the degeneracy of states $\rho(G)$ (i.e. the number of states (or structures) within the ensemble that have the same energy). An identical procedure was carried out for the GB1 protein (PDB designation 1PGA) to approximate the free energy associated with the unfolded state ensemble for the two-state model, and both the unfolded and intermediate state ensembles in the three state model.

The partition function of each intermediate or unfolded ensemble is given as:

$$\begin{aligned} Z_{k\vee u} &= N_{k\vee u} \int \rho(G) \exp(-G/k_BT) \mathrm{d}G \\ &= \frac{N_{k\vee u}}{\sqrt{2\pi\sigma^2_{k\vee u}}} \int \exp\left(\frac{-G}{k_BT}\right) \exp\left(\frac{-(G-\bar{G}_{k\vee u})^2}{2\sigma^2_{k\vee u}}\right) \mathrm{d}G \\ &= N_{k\vee u} \exp\left(\frac{\sigma^2_{k\vee u}}{2(k_BT)^2} - \frac{\bar{G}_{k\vee u}}{k_BT}\right) \end{aligned} \tag{2}$$

where k_B is the Boltzmann constant, T is the temperature in Kelvins, and $N_{k\vee u}$ is the total number of possible structures in the partially unfolded state k or the unfolded state u. For each state $N_{k\vee u}$ was set to equal $\gamma^{n_{k\vee u}}$, where γ is the number of conformations per residue and $n_{k\vee u}$ is the number of unfolded residues in the state.

The free energy of each intermediate state k or the unfolded state u can be found using the relation $G_{k\vee u} = -k_BTln(Z_{k\vee u})$:

$$G_{k\vee u} = \bar{G}_{k\vee u} - \frac{\sigma^2_{k\vee u}}{2k_BT} - k_BT \ln N_{k\vee u} \tag{3}$$

We can write the partition function of the system containing both the native state and the ensemble of partially folded and unfolded states can be as:

$$\mathrm{Z} = \exp(-G_{NS}/k_BT) + \exp(-G_u/k_BT) + \sum_k \exp(-G_k/k_BT) \tag{4}$$

The stability of the native state is then given by the difference between the native state free energy and the free energy of the intermediate and unfolded ensemble, $\{K, u\}$:

$$\Delta G = G_{NS} + k_BT \ln\left(\exp(-G_u/k_BT) + \sum_k \exp(-G_k/k_BT)\right) \tag{5}$$

The stability is in the direction of folding, and so the more negative the stability the more stable the protein. The fraction of sequences in the native state at equilibrium, F_{fold} was computed using:

$$F_{\mathrm{fold}} = \frac{\exp(-\Delta G/k_BT)}{1+\exp(-\Delta G/k_BT)} \tag{6}$$

Selection for cooperative folding. The fitness of a sequence was set to equal the fraction of sequences in the native state F_{fold} minus a penalty for non-cooperative folding, F_{coop}, which was set to equal the average number of folded residues multiplied by a factor η. The fitness of a sequence was therefore calculated as:

$$\begin{aligned} F &= F_{\mathrm{fold}} - F_{\mathrm{coop}} \\ &= \frac{\exp(-\Delta G/k_BT)}{1+\exp(-\Delta G/k_BT)} \\ &\quad - \eta\left(\frac{\sum_k \exp(-G_k/k_BT)}{\exp(-G_u/k_BT) + \sum_k \exp(-G_k/k_BT)}\right), \end{aligned} \tag{7}$$

where the purpose of η is to tune the level of cooperativity i.e. a larger value of η would require selection for mutations which destabilise the intermediate states k, leading to greater cooperativity in folding.

Selection for epistasis. To select for mutations which are highly epistasic among native contacts, the fitness of a sequence was set to equal the fraction folded F_{fold} minus a penalty for sequences with little epistasis between native contacts, F_{epi}.

$$\begin{aligned} F &= F_{\mathrm{fold}} - F_{\mathrm{epi}} \\ &= \frac{\exp(-\Delta G/k_BT)}{1+\exp(-\Delta G/k_BT)} - \eta_E \frac{1}{E} \end{aligned} \tag{8}$$

Here, E is the average magnitude of the epistasis, $\epsilon_{i,j}$, the between each pair of native contacts, $E = \langle|\epsilon_{i,j}|\rangle$. Therefore, the larger the value of E, the lower the fitness penalty. $\epsilon_{i,j}$ is calculated using Eq. 15.

Quantifying cooperativity. Cooperativity in the protein folding transition is determined experimentally using the van't Hoff criterion, defined as the ratio of the van't Hoff enthalpy, ΔH_{vH}, evaluated at T_m, to the calorimetric enthalpy ΔH_{cal} of the entire transition.

The calorimetric enthalpy, ΔH_{cal}, is the enthalpy change during the observed unfolded transition, and can be calculated from the area under the heat capacity curve, with a baseline correction (**? ? ?**), between the temperature at which the majority if the system is in the native state T_N and the temperature at which the majority of the system is in the unfolded state T_U:

$$\Delta H_{cal} = \int_{T_N}^{T_U} [C_v(T) - f_N(T)C_{v,N}(T) - f_U(T)C_{v,U}(T)] \mathrm{d}T \tag{9}$$

where $C_v(T)$ is the heat capactiy of the system, $f_N(T)$ and $f_U(T)$ are the fraction of the system in the native and fully unfolded state, respectively, and $C_{v,N}(T)$ and $C_{v,U}(T)$ are the hypothetical heat capacities of the pure native and pure fully unfolded states respectively.

The heat capacity C_v was calculated as the differential with respect to temperature of the average enthalpy of the system, $H(T)$. The average enthalpy, $H(T)$, of the system at temperature T was calculated as the differential of system partition function (Eqn. 4) with respect to temperature, $H(T) = -\partial \ln Z/\partial\beta$:

$$H(T) = \frac{\sum_i (\bar{G}_i - \frac{\sigma_i^2}{2k_BT}) \exp(-G_i/k_BT)}{Z}, \tag{10}$$

where $i = \{\mathrm{NS}, \mathrm{u}, \mathrm{K}\}$ denoting a sum over all states of the system. The van't Hoff enthalpy is found from the effective equilibrium constant K_{eff}, which is the ratio of the fraction of the population in the unfolded state, f_u to the fraction in the remaining states, $K_{eff} = f_u/(1-f_u)$. The van't Hoff enthalpy can then be calculated using the van't Hoff equation:

$$\Delta H_{vH} = k_BT^2 \frac{\mathrm{d}\ln K_{eff}}{\mathrm{d}T} \tag{11}$$

The van't Hoff criterion can then be found as:

$$\kappa = \frac{\Delta H_{vH}}{\Delta H_{cal}} \tag{12}$$

If the value of $\kappa \approx 1$ then the transition can be considered to be 2-state, whereas for multistate processes $\kappa < 1$.

Evolutionary simulations. We simulated the evolution of a 155 amino acid protein, where the initial nucleic acid sequence was constructed by choosing a set of codons at random, and the fitness of the sequence was equal to Equation 7. Mutations in the nucleic acid would be made following K80 mutation model with equal nucleotide frequencies and a ratio of transition to transversion probabilities of 2.0, where mutations resulting in stop codons were rejected. When a mutation is introduced, the probability of fixation of this mutation depends upon its impact on protein fitness, where we can calculate the selective advantage s of a mutant using:

$$s = \frac{F' - F}{F} \tag{13}$$

where F is the fitness of the premutation wild-type and F' is the fitness of the mutated sequence. The selective advantage s can either be zero, positive or negative indicating the mutation to be either synonymous, advantageous or deleterious.

At each generation we consider all possible mutations to the nucleic acid sequence and calculate the probability of fixation of each mutation using Kimura's expression for diploid organisms:

$$P_{\text{fix}} = \frac{1 - \exp(-2s)}{1 - \exp(-4N_{\text{eff}}s)} \tag{14}$$

where N_{eff} is the effective population size which due to mating behaviour and population structure is in general smaller than the true population size, and here was set to equal 10^6. We then chose a mutation to accept with a probability proportional to the probability of fixation given in Equation 14.

Quantifying epistasis. Epistasis occurs between two mutations when the sum of their independent effects on a trait, $(\Delta\Delta G_i + \Delta\Delta G_j)$, is larger or smaller than their combined effect on the trait, $\Delta\Delta G_{i,j}$. To determine the epistasis between the amino acids at sequence positions i and j, for a given wild-type sequence S_{WT} with stability ΔG_{WT}, we determine the stability ΔG_i of the structure if we substitute a non-interacting amino acid $A_\emptyset$ at residue i. Similarly, we substitute a non-interacting amino acid $A_\emptyset$ in to the wild-type sequence at residue j to determine the stability ΔG_j. For the double mutation i, j, we substitute a non-interacting amino acid at both positions i and j simultaneously. We then calculate epistasis for stability between two sites i and j within the protein as:

$$\epsilon_{i,j} = \Delta\Delta G_{i,j} - (\Delta\Delta G_i + \Delta\Delta G_j). \tag{15}$$

where for each pair or single mutation $\Delta\Delta G_x = \Delta G_x - \Delta G_{WT}$, where ΔG_x is the stability following the mutation(s) x. The epistasis between a pair of residues can be either positive or negative. Positive epistasis occurs when the combined impact of two mutations at residues i and j on protein stability $\Delta\Delta G_{i,j}$ is greater than the sum of their individual impact $\Delta\Delta G_i + \Delta\Delta G_j$. Negative epistasis occurs when $\Delta\Delta G_{i,j}$ is less than $\Delta\Delta G_i + \Delta\Delta G_j$.

Calculating the probability a pair of resides *i* and *j* are in contact in the ensemble of partially folded and fully unfolded states. For any pair of residues i and j, we can calculate the contact probability $P_{i,j}$ in the ensemble of partially folded and fully unfolded states as:

$$P_{i,j} = \sum_k P_k \langle Q_{i,j} \rangle_k + (1 - \sum_k P_k) \langle Q_{i,j} \rangle_u \tag{16}$$

P_k is the probability of being in intermediate state k, $\langle Q_{i,j} \rangle_k$ is the average probability residues i and j are in contact in intermediate state k and $\langle Q_{i,j} \rangle_u$ is the average probability they are in contact in the unfolded state.

ACKNOWLEDGMENTS. R.G. and R.C.E. are funded by UK Biotechnology and Biological Sciences Research Council BB/P007562/1, D.D.P. is funded by National Institutes of Health GM083127

1. Starr TN, Thornton JW (2016) Epistasis in protein evolution. *Protein Science* 25(7):1204–1218.
2. Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ (2011) Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science* 332(6034):1190 LP – 1192.
3. Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* 312(5770):111 LP – 114.
4. Segrè D, DeLuna A, Church GM, Kishony R (2004) Modular epistasis in yeast metabolism. *Nature Genetics* 37:77.
5. Salverda MLM, et al. (2011) Initial Mutations Direct Alternative Pathways of Protein Evolution. *PLOS Genetics* 7(3):e1001321.
6. Gong LI, Suchard MA, Bloom JD (2013) Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* 2:e00631.
7. Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB (2011) Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. *PLOS Genetics* 7(2):e1001301.
8. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF (2011) Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. *Science* 332(6034):1193 LP – 1196.
9. Sanjuán R, Cuevas JM, Moya A, Elena SF (2005) Epistasis and the Adaptability of an RNA Virus. *Genetics* 170(3):1001 LP – 1008.
10. Wang X, Minasov G, Shoichet BK (2002) Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. *Journal of Molecular Biology* 320(1):85–95.
11. Miton CM, Tokuriki N (2016) How mutational epistasis impairs predictability in protein evolution and design. *Protein Science* 25(7):1260–1272.
12. Pollock DD, Thiltgen G, Goldstein RA (2012) Amino acid coevolution induces an evolutionary Stokes shift. *Proceedings of the National Academy of Sciences* 109(21):E1352–E1359.
13. Weinreich DM, Watson RA, Chao L, Harrison R (2005) PERSPECTIVE:SIGN EPISTASIS AND GENETIC CONSTRAINT ON EVOLUTIONARY TRAJECTORIES. *Evolution* 59(6):1165–1174.
14. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. *Nature* 490:535.
15. Olson C, Wu N, Sun R (2014) A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology* 24:2643–51.
16. Rollins NJ, et al. (2019) 3D protein structure from genetic epistasis experiments. *Nature Genetics* 51:1170–1176.
17. Watters AL, et al. (2007) The Highly Cooperative Folding of Small Naturally Occurring Proteins Is Likely the Result of Natural Selection. *Cell* 128(3):613–624.
18. Dobson CM (2001) The structural basis of protein folding and its links with human disease. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356(1406):133 LP – 145.
19. Dobson CM (2003) Protein folding and misfolding. *Nature* 426:884.
20. Thomas PJ, Qu BH, Pedersen PL (1995) Defective protein folding as a basis of human disease. *Trends in Biochemical Sciences* 20(11):456–459.
21. Jackson S (1991) Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* 30:10428–10435.
22. Zwanzig R (1997) Two-state models of protein folding kinetics. *Proceedings of the National Academy of Sciences* 94:148–150.
23. Mallik S, Akashi H, Kundu S (2015) Assembly constraints drive co-evolution among ribosomal constituents. *Nucleic Acids Research* 43(11):5352–5363.
24. Yadahalli S, Gosavi S (2014) Designing cooperatively Ito the designed protein Top7. *Proteins* 82:364–374.
25. Faisca P (2006) Cooperativity and the origins of rapid, single-exponential kinetics in protein folding. *Protein Science* 15:1608–1618.
26. Chan J, Stiles W (2001) Energetics of side chain packing staphylococcal nuclease assessed by systematic double mutant cycles. *Biochemistry* 40:14004–14011.
27. Akerman EJ, Ang ETH, Kanter JR, Tsigelny I, Palmer T (1998) Identification of Pairwise Interactions in the alpha-Neurotoxin-Nicotinic Acetylcholine Receptor Complex through Double Mutant Cycles. *Journal of Biological Chemistry* 273:10958.
28. Horovitz A (1996) Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Folding and Design* 1:R121–R126.
29. Pollock DD, Pollard ST, Shortt JA, Goldstein RA (2017) Mechanistic Models of Protein Evolution BT - Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts, ed. Pontarotti P. (Springer International Publishing, Cham), pp. 277–296.
30. Chan H, Bromberg S, Dill K (1995) Models of cooperativity in protein folding. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 348(1323):61 LP – 70.
31. Tsong TY, Baldwin RL, McPhie P, Elson EL (1972) A sequential model of nucleation-dependent protein folding: Kinetic studies of ribonuclease A. *Journal of Molecular Biology* 63(3):453–469.
32. Savo L (1979) *Physiochemical aspects of protein denaturation.* (John Wiley and Sons).
33. Privalo P (1979) Stability of proteins: small globular proteins. *Advances in Protein Chemistry* 33:167–241.
34. Saboury A, Moosavi Movahedi A (1994) Clarification of calorimetric and van't Hoff enthalpies for evaluation of protein transition states. *Biochemistry and Molecular Biology Education* 22:210–211.
35. Privalo P (1982) Stability of proteins: proteins which do not present a single cooperative system. *Advances in Protein Chemistry* 35:1–104.
36. Chan H (2000) Modelling Protein Density of States: additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. *Proteins* 40:543–571.
37. Wells JA (1990) Additivity of Mutational Effects in Proteins. *Biochemistry* 29:8509–8517.
38. Miyazawa S, Jernigan R (1999) An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins* 36:357–69.
39. Flory PJ (1953) *Principles of Polymer Chemistry.* (Cornell University Press).
40. Flory PJ (1989) *Statistical Mechanics of Chain Molecules.* (Hanser Gardner Publications).

Figure Captions

Figure 1 Investigating the cooperativity of sequences evolved under zero, low, medium and high values of the tuning coefficient η by considering, (a) the fraction of the system found in the fully unfolded state during denaturation under increased temperature, (b) the fraction in the ensemble of intermediate states during the unfolding transition, (c) heat capacity curves during the unfolding transition, the area under which is the enthalpy change associated with the transition, and (d) the van't Hoff ratio of the unfolded transition associated with each value of the cooperativity tuning coefficient η. All values are averaged over the set of 1000 most evolved sequences for each evolutionary simulation.

Figure 2 The distribution of epistasis as the selection coefficient η increases. The normalised distribution of the epistasis in protein stability between a) native contacts and b) non-native contacts when evolving proteins under varying degrees of selection for cooperativity. As selection for cooperativity increases, more native contacts experience higher magnitude (more negative) epistasis, whilst more non-native contacts experience very low levels of epistasis. The area

under each curve sums to 1. c) The mean of the epistasis distributions, and d) the variance of the epistasis distribution, of the final 2,000 generations of the 50,000 generations simulated, averaged over all 10 simulations for native contacts (blue) and non-native contacts (red). The error bars represent the variance of these values across the 10 simulations. The average of the epistasis distribution at the native contacts becomes more negative as the value of the selection coefficient η increases and the variance in the distribution increases. The average of the epistasis distribution at the non-native contacts goes to zero as the selection coefficient η increases, and the variance decreases.

Figure 3 The mean absolute epistasis in protein stability (y-axis), averaged over all 10 simulations, between each pair of native contacts when evolving proteins under increasing selection (η_E = 0, low, medium) for the absolute epistasis in protein stability at the native contacts (x-axis). The error bars depict the variance in the mean. The average absolute epistasis between native contacts increases as the value of the selection coefficient η_E increases.

Figure 4 The distribution of epistasis as the selection coefficient η_E increases. The normalised distributions of epistasis between a) native and b) non-native contacts when evolving sequences under different magnitudes of selection for the average magnitude of epistasis between the native contacts. As the value of the selection coefficient η_E increases, a higher number of native contacts experience greater magnitude negative epistasis, whilst a higher number of non-native contact pairs experience non-zero epistasis. The area under the curves sum to 1. c) The mean of the epistasis distributions, and d) the variance of the epistasis distribution, of the final 2,000 generations of the 5,000 generations simulated, averaged over all 10 simulations for native contacts (blue) and non-native contacts (red). The error bars represent the variance of these values across the 10 simulations. The average of the epistasis distribution at the native contacts becomes more negative as the value of the selection coefficient η_E increases and the variance in the distribution increases. The average of the epistasis distribution at the non-native contacts remains roughly constant as the selection coefficient η_E increases, but the variance in the distribution increases.

Figure 5 Investigating the cooperativity of the unfolding transition with increased selection for the average magnitude of epistasis at the native contacts, by considering a) The fraction of the population in the unfolded state during denaturation and b) the fraction of the population in the ensemble of partially folded states during denaturation. As the value of the selection coefficient η_E increases, the unfolding transition becomes less sharp and the fraction of the population in the intermediate states increases, showing the sequences are becoming less cooperative.

Figure 6 The average epistasis ϵ between each possible pair of residues for a) a two-state and b) a 3-state system. The upper triangle of each heat map is the calculated epistasis, the lower triangle is the contact map for the GB1 protein, where a value of 1 means the residues are in contact in the native state. The epistasis for the two-state system accurately maps out the structure of the GB1 protein, with the majority of native contacts experiencing high magnitudes of negative epistasis. For the three state system, the native contact map becomes obscured by high levels of positive epistasis at non-native contacts. These results suggest epistatic measurements work when reconstructing the native state of two-state systems, but are less successful for multi-state systems.

Figure 7 Histogram of the distribution of contact probabilities $P_{i,j}$ between site i and site j when evolving proteins under selection for cooperativity, when η is set to a) 0, b) low, c) medium and d) high.

Figure 8 Histogram of the distribution of contact probabilities $P_{i,j}$ between site i and site j when evolving proteins under selection for epistasis between native contacts, when η is set to a) 0, b) low and c) medium.