Physical Biology

OPEN ACCESS

RECEIVED
15 November 2023

REVISED
16 December 2023

ACCEPTED FOR PUBLICATION
18 January 2024

PUBLISHED
31 January 2024

PAPER

Structure of the space of folding protein sequences defined by large language models

A Zambon[1], R Zecchina[2,*] and G Tiana[1,3,*]

[1] Department of Physics and Center for Complexity and Biosystems, Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
[2] Bocconi University, via Roentgen 1, 20136 Milano, Italy
[3] INFN, Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
[*] Authors to whom any correspondence should be addressed.

E-mail: riccardo.zecchina@unibocconi.it and guido.tiana@unimi.it

Keywords: energy landscape, protein evolution, canonical-ensemble sampling, machine learning

Supplementary material for this article is available online

Abstract

Proteins populate a manifold in the high-dimensional sequence space whose geometrical structure guides their natural evolution. Leveraging recently-developed structure prediction tools based on transformer models, we first examine the protein sequence landscape as defined by an effective energy that is a proxy of sequence foldability. This landscape shares characteristics with optimization challenges encountered in machine learning and constraint satisfaction problems. Our analysis reveals that natural proteins predominantly reside in wide, flat minima within this energy landscape. To investigate further, we employ statistical mechanics algorithms specifically designed to explore regions with high local entropy in relatively flat landscapes. Our findings indicate that these specialized algorithms can identify valleys with higher entropy compared to those found using traditional methods such as Monte Carlo Markov Chains. In a proof-of-concept case, we find that these highly entropic minima exhibit significant similarities to natural sequences, especially in critical key sites and local entropy. Additionally, evaluations through Molecular Dynamics suggests that the stability of these sequences closely resembles that of natural proteins. Our tool combines advancements in machine learning and statistical physics, providing new insights into the exploration of sequence landscapes where wide, flat minima coexist alongside a majority of narrower minima.

1. Introduction

Protein evolution can be described as a stochastic process in the space of sequences. Although it is not possible to predict exactly the course of this process [1], evolution is strongly constrained by functional requirements. One of them is that of foldability, namely that most proteins must display a unique and well-defined native conformation to be functional. This is quite a robust requirement that filters out the vast majority of protein sequences [2].

The space of folding sequences is then a subset of the space of all sequences, whose properties affect the evolution of the protein. Proteins displaying a given function tend to conserve their structure (within an RMSD of 2.5 Å) even among very distant homologous [3]. Consequently, it is reasonable to assume that conformational similarity to the wild-type protein is a feature that contributes to the functionality of a mutant, and thus to its evolutionary fitness. Such conformational similarity can be quantified by a distance from the structure of a reference wild-type protein, thus defining a landscape in which evolution is expected to take place through low conformational distance trajectories.

This landscape in sequence space is analogous to the energy landscape of other complex systems studied in physics. In the case of disordered systems, like spin glasses, the energy landscape is rugged [4] and its minima are separated by high barriers that prevent diffusion across their conformational space [5]. Although the excluded volume of the amino acids is like geometric frustration in glasses and in jamming problems, proteins are quite different from prototypical models of disordered systems. Proteins are small, the hydrophobic amino acids superpose a ferromagnetic-like interaction to the other disordered interactions, and their backbone makes their physical properties quite peculiar.

The space of sequences of proteins that fold to a stable native conformation was studied using minimal protein models that, although not realistic from the biochemical point of view and thus not predictive, display some of the complexity of natural proteins [6]. The properties of a sequence in these models are only determined by its native energy because the rest of the conformational spectrum is self-averaging [7]; the thermodynamic properties of a sequence are thus determined essentially by the energy E_N of the native conformation. Monte Carlo techniques that control E_N are then a suitable tool for sampling the space of sequences. In this way, it was shown that stable proteins display a complex hierarchical organization with regions not connected by single-point mutations and conserving few mutually interacting residues [8]. Nonetheless, it was shown that folding sequences connected by neutral paths can visit vast regions of the space [9].

The recent development of machine-learning algorithms [10, 11] able to predict the native structure of an input sequence paves the way to studying the space of folding sequences in the context of a realistic, predictive model. Using these algorithms, one can bypass the need of using effective energies, which are not always reliable, to characterize the foldability of a sequence.

In the present work, we chose a reference protein structure of interest and used structure predictors to define an effective energy for each sequence. The goal was to characterize the low energy manifolds of the associated landscape. For this purpose, we cast the exploration problem in the physical framework of the canonical ensemble, where several efficient sampling algorithms are available [12].

We use a large language model for structure prediction and we combine it with different exploration algorithms. Our method is designed to navigate efficiently through regions of sequence space that have high local entropy (neutral regions). We have put this method to the test on a well studied protein structure, generating predictions that are validated against existing data or through molecular dynamics simulations. The objective of this study is to demonstrate how various innovative approaches, such as language models and algorithms driven by local entropy, can be effectively merged.

The paper is organized as follows: first we describe the methods used to define the effective energy and sample the associated space within the framework of the canonical ensemble. Then, we present the results obtained varying the selective temperature of the system. After selecting a realistic value of the selective temperature, we describe the structure of the energy minima in sequence space, focusing particularly on the width of the corresponding basins. Inspired by the techniques used in connection with artificial intelligence, we finally test an algorithm that can identify large energy minima. We discuss the relevance of these results for protein evolution.

2. Methods

2.1. Sampling the effective energy of a sequence

In order to sample the space of sequences folding to a given reference conformation $\boldsymbol{r}_0$, we employed a canonical ensemble formalism where each sequence is characterized by an effective energy defined as the fraction of contacts that its native conformation has in common with $\boldsymbol{r}_0$,

$$E(\boldsymbol{\sigma}) = \frac{\sum_{ij} |\Delta_{ij}(\boldsymbol{r}(\boldsymbol{\sigma})) - \Delta_{ij}(\boldsymbol{r}_0)|}{\sum_{ij} \left[\Delta_{ij}(\boldsymbol{r}(\boldsymbol{\sigma})) + \Delta_{ij}(\boldsymbol{r}_0)\right]}, \tag{1}$$

where $\Delta_{ij}(\boldsymbol{r})$ $(\Delta_{ij}(\boldsymbol{r}_0))$ is the contact map of the native conformation $\boldsymbol{r}$ $(\boldsymbol{r}_0)$, whose elements are 1 if any heavy atom of amino acid i is within 4 Å from any heavy atom of amino acid j and 0 otherwise, with $|j-i| > 1$ in order to eliminate the contribution of trivial contacts. Thus, the energy ranges between 0, when all contacts of a sequence are the same as in $\boldsymbol{r}_0$, and 1 if all contacts are different.

The native conformation associated to a generic sequence of amino acids $\boldsymbol{\sigma}$ was predicted by ESMFold [11], a transformer protein language model defined by approximately 15 billion parameters trained over 65 million protein sequences. The same model was also employed to predict the structure $\boldsymbol{r}_0 = \boldsymbol{r}(\boldsymbol{\sigma}_0)$ of the reference sequence $\boldsymbol{\sigma}_0$.

The sampling was carried out with a Metropolis algorithm [13] at different temperatures T_s (expressed in energy units, cf figure 1(a), that here have the meaning of evolutionary bias towards good (i.e. low-energy) folding sequences. Throughout the simulation, at each step, a random single-site mutation was proposed and the newly generated mutant was accepted or rejected based on its energy, that is the Metropolis rate is here $w(\boldsymbol{\sigma}'|\boldsymbol{\sigma}) = p_{ap}(\boldsymbol{\sigma}'|\boldsymbol{\sigma}) \cdot \min[1, \exp(-[E(\boldsymbol{\sigma}') - E(\boldsymbol{\sigma})]/T_s)]$, where the *a priori* probability is uniform for pairs of sequences with only one different site.

Summing up, at each step a random single-point mutation is proposed, the conformation associated with the mutated sequence is predicted by ESM-Fold, its structural difference quantified by the effective energy of equation (1) and the mutation is accepted or rejected according to the Metropolis criterion. Here, the temperature T_s controls the acceptance rate, in the sense that the lower the temperature, the more unlikely are accepted mutations that modify the contact map of the protein. This procedure is iterated $10^5 - 10^6$ times.

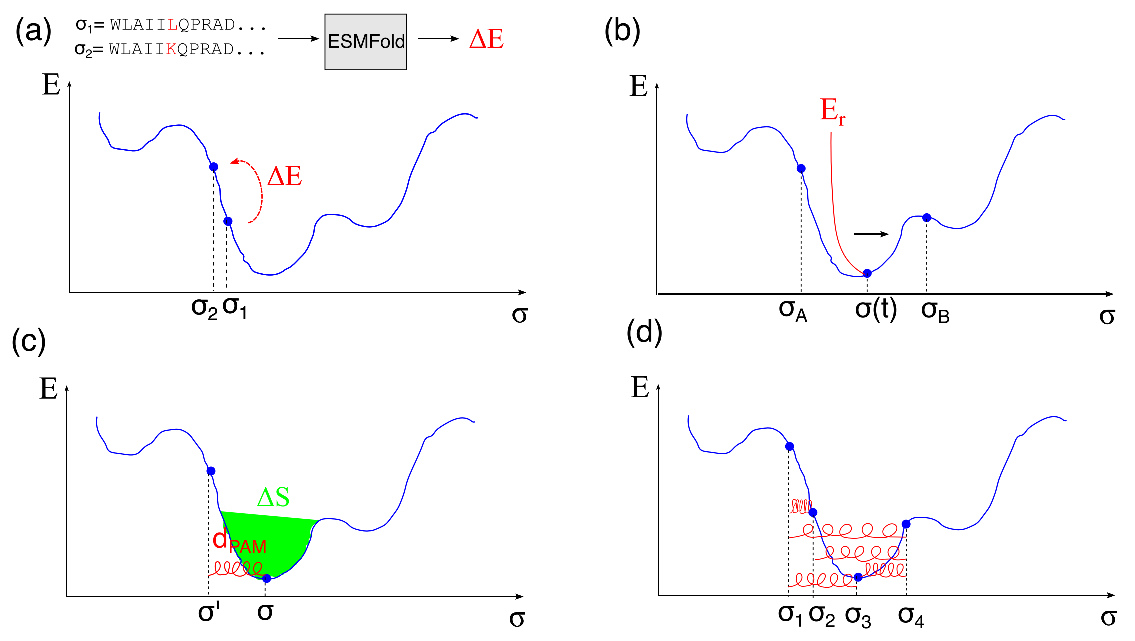


Figure 1. A sketch of the algorithms used in the present work. (a) We used a Monte Carlo scheme to sample the sequence ($\boldsymbol{\sigma}$) landscape, in which the probability of accepting a mutation depends on the change ΔE of effective energy. (b) In the ratchet simulations used to generate trajectories from a sequence $\boldsymbol{\sigma}_A$ to $\boldsymbol{\sigma}_B$, a time-dependent energy E_r damps the fluctuations in the direction opposite to $\boldsymbol{\sigma}_B$. (c) To estimate the local entropy $\Delta S(\boldsymbol{\sigma})$, we used thermodynamic integration on simulations restrained by a pulling force to $\boldsymbol{\sigma}$. (d) Wide energies basins are sampled by an algorithm in which different replicas are mutually restrained by pulling forces.

2.2. Ratcheted sampling

In order to estimate the energy barriers along trajectories from a sequence $\boldsymbol{\sigma}_A$ to a sequence $\boldsymbol{\sigma}_B$, we employed a Metropolis algorithm which starts from $\boldsymbol{\sigma}_A$ and damps the fluctuations in the direction opposite to $\boldsymbol{\sigma}_B$ (cf figure 1(b). This is based on the principle of the ratchet and the paw and it was used to generate trajectories in the space of protein conformations that resemble physical trajectories [14].

The Metropolis algorithm is applied with an energy that is given by equation (1) summed to

$$E_r(\boldsymbol{\sigma}(t)) = \begin{cases} \frac{k}{2}\left[d(\boldsymbol{\sigma}(t),\boldsymbol{\sigma}_B) - d_m(t)\right]^2 & \text{if } d(\boldsymbol{\sigma}(t),\boldsymbol{\sigma}_B) > d_m(t) \\ 0 & \text{if } d(\boldsymbol{\sigma}(t),\boldsymbol{\sigma}_B) \leqslant d_m(t) \end{cases} \qquad (2)$$

where $d_m(t) \equiv \min_{t'<t} d(\boldsymbol{\sigma}(t'),\boldsymbol{\sigma}_B)$ is the minimum Hamming distance to $\boldsymbol{\sigma}_B$ encountered along the trajectory. This time-dependent energy favors the moves towards $\boldsymbol{\sigma}_B$, without exerting work to push the system. In this way, the system crosses the lowest energy barriers as in the unbiased trajectories [15].

2.3. Local Entropy

The local entropy has been introduced as a tool for analyzing complex energy landscapes in which flat regions coexist with rugged ones, in the context of non-convex neural networks [16]. The local entropy of a discrete system is defined as

$$S_{T_s,\gamma}(\boldsymbol{\sigma}) = \log\left[\sum_{\boldsymbol{\sigma}'} e^{-E(\boldsymbol{\sigma}')/T_s - \gamma d_{\mathrm{PAM}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')}\right] \qquad (3)$$

and is meant to quantify the width of the energy basin around a sequence $\boldsymbol{\sigma}$. Here γ is a Lagrange multiplier that controls the average distance from $\boldsymbol{\sigma}$.

To define the neighborhood of a sequence, we define a distance

$$d_{\mathrm{PAM}}(\boldsymbol{\sigma},\boldsymbol{\sigma}') = N^{-1}\sum_i \left[1 - \frac{P(\sigma_i,\sigma_i') + P(\sigma_i',\sigma_i)}{2}\right] \qquad (4)$$

that keeps into account the chemical similarity between amino acids, where

$$P(\alpha,\beta) = \frac{\mathrm{PAM1}[\alpha,\beta]}{\sum_{\gamma\neq\beta}\mathrm{PAM1}[\gamma,\beta]}$$

is the transition rate from the β to the α amino acid as defined by the PAM1 matrix, whose elements are the empirical probabilities multiplied by 1000 of single-point mutations in proteins [17], setting the diagonal elements $P(\alpha,\alpha) = 1$. A further advantage of d_{PAM} with respect to the Hamming distance d is that it varies essentially as a real variable.

From the identity

$$\frac{\partial S_{T_s,\gamma}(\boldsymbol{\sigma})}{\partial\gamma} = -\frac{\sum_{\boldsymbol{\sigma}'} d_{\mathrm{PAM}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')\, e^{-E(\boldsymbol{\sigma}')/T_s - \gamma d_{\mathrm{PAM}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')}}{\sum_{\boldsymbol{\sigma}'} e^{-E(\boldsymbol{\sigma}')/T_s - \gamma d_{\mathrm{PAM}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')}} = -\langle d_{\mathrm{PAM}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')\rangle_{T_s,\gamma} \qquad (5)$$

and keeping in mind that

$$\lim_{\gamma\to\infty} S_{T_s,\gamma}(\boldsymbol{\sigma}) = -E(\boldsymbol{\sigma})/T_s \tag{6}$$

one can derive the local entropy difference $\Delta S_{T_s,\gamma}(\boldsymbol{\sigma}) \equiv S_{T_s,\gamma}(\boldsymbol{\sigma}) - S_{T_s,\infty}(\boldsymbol{\sigma})$ with respect to the single sequence $\boldsymbol{\sigma}$. The subtraction of the reference entropy $S_{T_s,\infty}(\boldsymbol{\sigma})$ allows us to compare directly the values of $\Delta S_{T_s,\gamma}(\boldsymbol{\sigma})$ associated with different sequences $\boldsymbol{\sigma}$. This is found by calculating the integral

$$\Delta S_{T_s,\gamma}(\boldsymbol{\sigma}) = \int_{\gamma}^{\infty} \langle d_{\mathrm{PAM}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')\rangle_{T_s,\gamma'}\, d\gamma', \tag{7}$$

that can be estimated numerically from simulations performed at different values of γ (cf figure 1(c).

2.4. Replica simulations

It was shown in [18] that it is possible to bias the sampling of wide minima, that is states of the system with high local entropy, with a replicated Monte Carlo algorithm; here, one does not sample the standard Boltzmann distribution, but a distribution based on the local entropy. We then consider Monte Carlo algorithms in which different replicas of the system are coupled together by an interaction potential depending on the distance between their mutual sequences (cf figure 1(d). Each replica evolved by a Metropolis algorithm based on the coupling potential

$$E_{\mathrm{rep}}\left(\{\boldsymbol{\sigma}_i\}_{i=1}^{y}\right) = \sum_{i=1}^{y} E(\boldsymbol{\sigma}_i) + \gamma^* \sum_{i=1}^{y}\sum_{j\neq i}^{y} d_{\mathrm{PAM}} \times \left(\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_j\right) \tag{8}$$

where $E(\boldsymbol{\sigma}_i)$ is the effective energy of the ith replica (see equation (1)) and $\gamma^* = \gamma T_s$, with γ being the Lagrange multiplier indicated in equation (3).

In each simulation, we increased slowly that value of γ, until the y replicas collapsed on a single high-entropy sequence. Eventually, we obtained a single sequence from each simulation. We repeated the whole procedure to collect more sequences.

3. Results

3.1. Thermodynamics of the space of sequences

We performed samplings of the sequence space at different temperatures T_s for protein G, a widely-studied small protein [19] made of an alpha helix and two beta hairpins. Each simulation lasted for at least $\sim 3\times 10^5$ steps (see some examples in figures S1 and S2 in the supplementary material); we calculated from them the average energy and the specific heat using a multiple-histogram algorithm [20]. The system displays a marked transition at temperature $T_s^c \approx 1.1\times 10^{-2}$ between sequences whose native structure has more than 80% common contacts with the reference structure to sequences with less than 50% of predicted contacts (upper panel in figure 2). The specific heat also displays a broad shoulder centered at $T_s^n \approx 2.4\times 10^{-3}$, at which the average similarity between the contact maps is approximately 95%. It should be noted that the typical relative error in the prediction of the contacts of experimentally known structures is approximately 0.1 (cf figure S3 in the supplementary material), so in the low-temperature phase (i.e. $T_s \leqslant 3.2\times 10^{-3}$), native conformations are indistinguishable from the experimental one.

It is worth mentioning that, although the space of sequences is combinatorially large, the quantities of interest seem to have reached convergence in the simulation time. To check this, we removed the first steps of the simulation at which the auto-correlation of the Hamming distance from the reference sequence was above 0.1, we then divided the rest of the simulation into non-overlapping time blocks and we calculated the average and the standard deviation in each block, showing that they reach stationary values (cf figures S1 and S2 in the supplementary material).

ESMFold quantifies the degree of confidence in the predicted position of each atom with the pLDDT parameter [10]. We calculated the average pLDDT over all the C_α atoms of each sequence sampled at a defined temperature. The average pLDDT (lower panel of figure 2) of sequences sampled at low temperatures is comparable with that of extant sequences obtained from the pdb, which is 80.1 ± 12.6 (cf figure S3 in the supplementary material), suggesting that the algorithm is confident that the predicted structures correspond to the unique native state of the protein. The pLDDT is roughly constant at the value of approximately 85 for $T_s < T_s^n$, and then it starts decreasing. However, it remains above 70, which is commonly regarded as the threshold for a good prediction, up to T_s^c. At higher temperatures, it drops to 40, which is considered a mark of disorder [21]. This suggests that at high temperatures not only sequences are not folding to structures different from the reference one, but they are not folding at all.

Interestingly, at all temperatures, the sequence can step away from the initial, reference sequence (cf figure S4 in the supplementary material). Even at the lowest simulated temperature $T_s = 8\times 10^{-4}$ at which the effective energy is essentially zero, the average similarity from the initial reference sequence is $\overline{q(\boldsymbol{\sigma},\boldsymbol{\sigma}_0)} = 0.37$. This result agrees with the experimental observation that real proteins can change up to ∼75% of their sequence while still maintaining their function [3].

The main assumption we made in sampling the sequence space, as described above, is that ESMFold inference for mutant sequences is as good as that for native ones. Although it is difficult to validate this hypothesis, and we cannot rule out the presence of artifacts when sequences are distant from natural

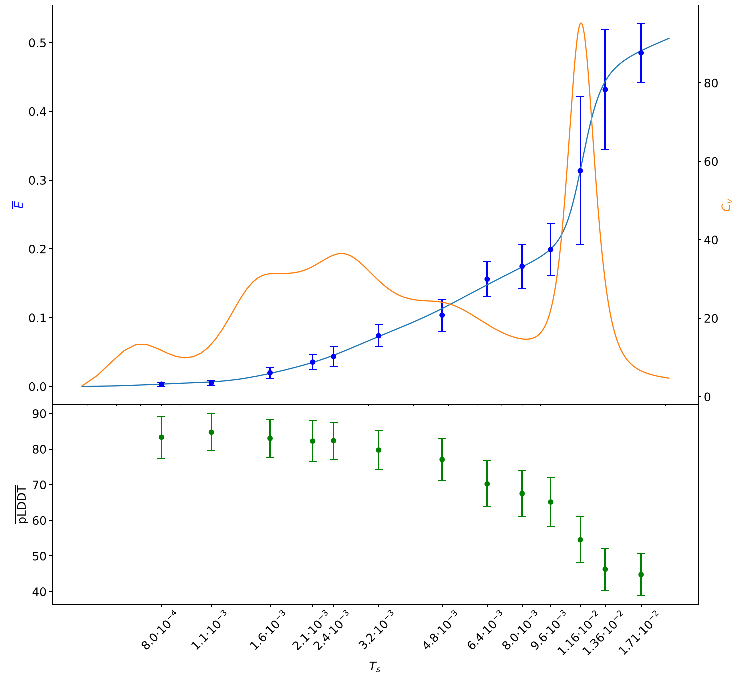


Figure 2. Upper panel: the measured average effective energy $\overline{E}$ (blue circles) calculated from the samplings at different temperatures T_s. The solid lines indicate $\overline{E}$ and C_v as estimated from a multiple-histogram algorithm. Lower panel: the values of pLDDT (green circles) representing the average degree of confidence associated to a sequence for its predicted native structure at the different temperatures T_s.

ones, we did perform some checks on the generated sequences.

First, we performed some molecular dynamics simulations of three sequences generated by ESMFold (table 2) with a protein model that is regarded as realistically predictive [22], starting from the putative native conformation for 200 ns at $T = 310$ K. The three sequences display an average RMSD to the initial conformation of 0.18 ± 0.04 nm, 0.22 ± 0.06 nm and 0.25 ± 0.09 nm, respectively (cf figure S5 in the supplementary material). These values are the typical mutual similarities of homologous proteins [3], and they are lower than the value 0.36 ± 0.10 nm found for a selected high-energy sequence.

Moreover, we verified that sampled sequences display concentrations of the twenty kinds of amino acids comparable to those of natural proteins (figure S6 in the supplementary material). In particular, at $T_s = 3.2 \times 10^{-3}$ the overall concentration of hydrophobic and negatively-charged residues is similar in natural (0.30 and 0.12, respectively) and designed (0.29 and 0.11, respectively) proteins. Only the fraction of positively-charged residues is slightly lower in designed proteins (0.08) than that in natural proteins (0.11). These results suggest that the proteins sampled at low temperature are native-like.

3.2. Structure of the space of sequences

A standard tool used to study the energy landscape of complex systems is the distribution of similarity q between the sampled states [4]. In the present case, the value of q between two sequences is defined as the fraction of sites that host the same kind of amino acids. The sampled distribution $p(q)$ displays a unimodal shape in all simulations, whose maximum q_{EA} increases at lower temperatures (figure 3).

Table 1. The results of the fit of $p(q)$ with the model of equation (9). The lowest temperature cannot be fitted with a binomial.

T	N_{f}	n_{aa}
1.7×10^{-2}	0	20
6.4×10^{-3}	0	11
3.2×10^{-3}	0	6
1.6×10^{-3}	6	5
8.0×10^{-4}	1	3

A preliminary interpretation of these curves can be obtained using a very simple model in which the amino acids of the protein of length $N = 56$ can vary with uniform probability, except for a number N_{f} of them that are fixed and identical in all sequences. This gives a binomial distribution

$$p(q) = \binom{N - N_{\text{f}}}{qN - N_{\text{f}}} \frac{1}{n_{\text{aa}}^{qN-N_{\text{f}}}} \left(1 - \frac{1}{n_{\text{aa}}}\right)^{N(1-q)}, \quad (9)$$

where n_{aa} is the number of different types of amino acids.

At high temperature ($T = 1.71 \times 10^{-2}$) we find $N_{\text{f}} \approx 0$ and $n_{\text{aa}} \approx 20$ (see table 1), with the distribution peak centered at $q_{\text{EA}} \approx 1/20$. This is compatible with a state in which amino acids vary essentially at random. Thus, we conclude that for $T_s > T_s^c$ the system displays a single disordered phase.

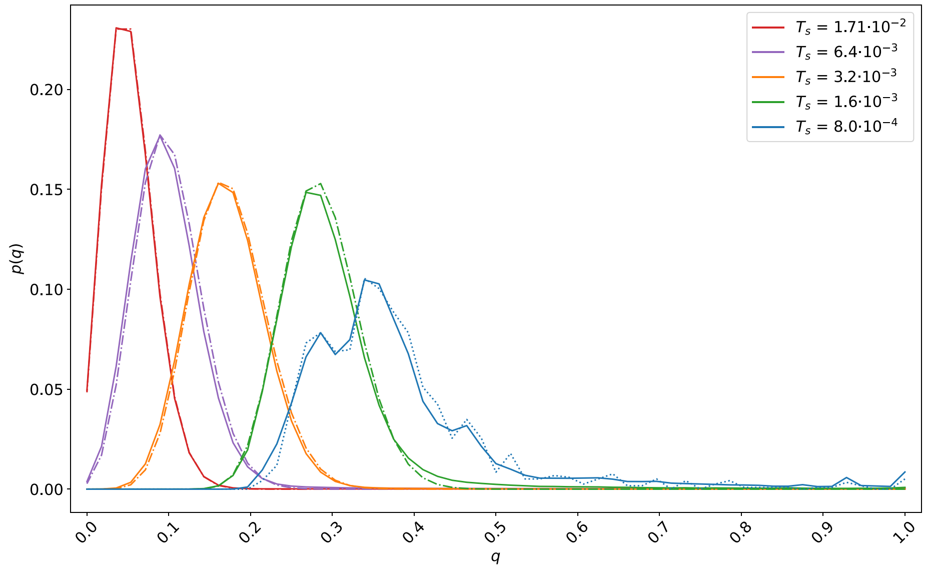


Figure 3. The distribution of sequence similarity q between the pairs of sequences sampled at different temperatures T_s. The dot-dashed line is the binomial fit of equation (9). The dotted line at $T_s = 8 \times 10^{-4}$ is obtained by decimation of the data, taking one sequence every 10^4 steps.

At lower temperatures (e.g. $T_s = 6.4 \times 10^{-3}$, $T_s = 3.2 \times 10^{-3}$ and $T_s = 1.6 \times 10^{-3}$, figure 3), the $p(q)$ is still compatible with a binomial distribution. The first two of them (where $T_s^n < T_s < T_s^c$) are fitted with the parameters $N_\mathrm{f} = 0$ and $n_\mathrm{aa} < 20$ indicating that, according to the minimal model, all residues of the chain can still change, but there is a selection on the type of amino acids they can host.

At the lowest temperature $T_s = 8 \times 10^{-4}$ ($T_s < T_s^n$, at which the predicted structure is essentially identical to the reference one), the shape of $p(q)$ is more irregular, with a tail reaching as maximum similarity $q_M = 1$. This suggests that the explored manifold is more complex than at larger temperatures, with energy minima at any mutual distance. The fact that $q_M = 1$ indicates that the number of minima is small enough that the probability that the system returns to the same sequences is not negligible. To rule out the possibility that these results are artifacts due to long correlation times in the sampling, we have down-sampled the data, obtaining a distribution almost identical to the original one (dotted line in figure 3).

To challenge the minimal binomial model, we have then analyzed the degree of conservation of the sites of the protein as a function of the temperature. The site entropy $\mathcal{S}(i) \equiv -\sum_\alpha p_i(\alpha) \log p_i(\alpha)$, where $p_i(\alpha)$ is the probability of observing the amino acid of kind α at site i, is zero if the site is perfectly conserved and $\log 20 \approx 3$ if it displays a uniform probability of hosting the 20 amino acids.

At high temperatures ($T_s > T_s^c$), the distribution of amino acids is uniform in all sites (figure 4). At the lowest temperature ($T_s < T_s^n$), there are 7 sites that are never mutated and another 9 that are highly conserved, their entropy being lower than 1. Interestingly, approximately one-third of sites display an entropy larger than 2, comparable to that of high-temperature sequences. At the intermediate temperatures ($T_s^n < T_s < T_s^c$) there is still a (variable) number of low-entropy, highly conserved sites, and a majority of sites whose entropy is comparable to that of high-temperature sequences.

The picture that emerges is that, at all temperatures $T_s < T_s^c$, there is a clear partitioning between highly and poorly conserved sites, and that the main effect of temperature is to define the ratio between the two. As a consequence, in the case of protein sequences the distribution $p(q)$ (figure 3) may not contain all the relevant information and it could be misleading, erroneously suggesting that even at low temperatures the system is in a highly disordered state.

In particular, there are some sites, i.e. 5, 14, 26, 30, 41, 43 and 54, that are completely conserved at the lowest temperature $T_s = 8 \times 10^{-4}$ in figure 4 and are remarkably conserved at all $T_s < T_s^c$. We call them K-sites and their conservation is not apparent in the plot of $p(q)$ (but can nonetheless be accounted by another, tailor-made similarity measure; cf figure S7 in the supplementary material). At $T_s < T_s^n$ visited combinations of the K-sites amino acids are closely distributed in the sequences space, contrary to the information carried by the standard Hamming similarity distribution (cf figure 3). Above T_s^n, the K-sites start mutating more freely. Here, the number of relevant sites for the structural properties of the sequence diminishes as the temperature rises, since the structural constraint on the amino acid sequences becomes less and less rigid.

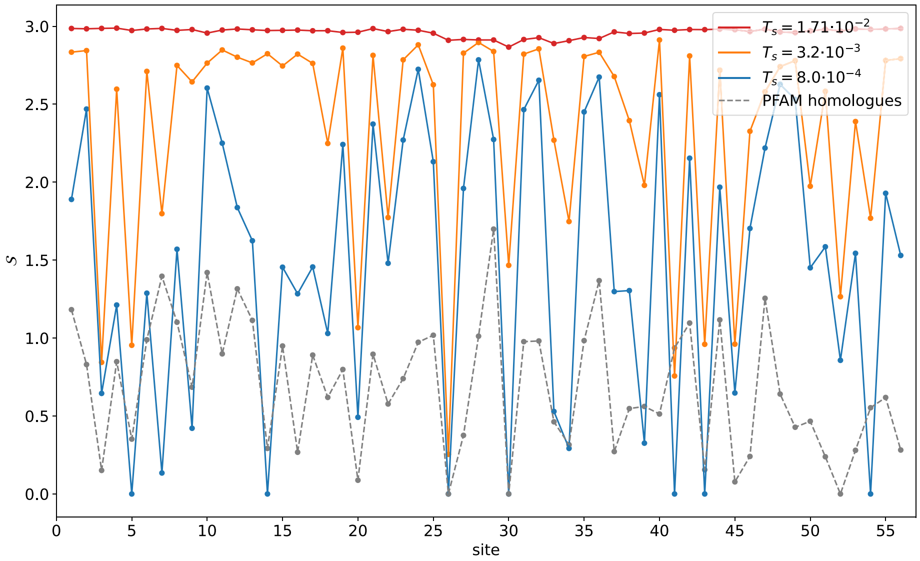


Figure 4. The degree of conservation $\mathcal{S}$ of the sites quantified by the associated site entropy for the sequences sampled at different temperatures T_s (colored lines) and for the experimental homologs of protein G (dashed line).

Table 2. Some of the sequences used in the calculations. 1PGB, E1NUT2, K9EXL1A and K9EXL1B are natural sequences labeled by their pdb code. S1, S2 and S3 are artificial sequences obtained by quenching the temperatures in Monte Carlo samplings. In bold, the K-sites amino acids for each sequence.

id	E	sequence
1PGB	0	MTYK**L**ILNGKTLK**G**ETTTEAVDAAT**A**EKV**F**KQYANDNGVD**G**E**W**TYDDATKTFT**V**TE
E1NUT2	0.17	TVYH**F**QYDKKGTS**I**RQDFAAVNKEI**A**EMH**F**KEYATESGLD**A**H**F**AYNEANQTFV**Y**KD
K9EXL1A	0.31	EVYT**F**YYRTQNQN**G**ATTVKASSPRE**A**LEY**F**QNFLSERGLD**F**N**W**HYESEDRVFT**A**SE
K9EXL1B	0.14	AVYT**F**VYNTKGKN**G**ATTVKASSPEE**A**LEY**F**QNWAKENDLE**L**D**W**SYDEDTKTFT**G**RE
S1	0.04	PTYR**M**EVMSTHFE**A**VVGIEAPNYPA**A**LHG**F**VLFCHCLGVL**A**Q**F**TYCATHNFFK**V**WQ
S2	0.07	HWYR**F**VHHGPNHE**C**MGVARVPHVHW**L**MNA**V**EKATKAANIK**C**K**Y**RWSARHRTLW**C**YT
S3	0.06	HEYS**C**MLISPLRT**A**TQVFEATNRAM**A**HWF**F**EDMALWLGYI**K**K**W**TYNERFHMYT**V**TF

3.3. Comparison with experimental data

Sequences produced by the natural evolution of protein G can be obtained from the PFAM database [23]. The main statistical observable that can be calculated from these data and compared with the simulations is the site entropy (dashed curve in figure 4).

There are at least two important reasons that make the comparison difficult. First, PFAM sequences are not an unbiased ensemble that reflects the evolution of organisms but they are affected by the choices of researchers to study specific homologs. Moreover, simulations only require that a sequence folds to the correct native state but does not add any functional requirement. This simplification is likely to increase the entropy of sites that lie on the surface of the protein and that are involved in interactions with the cellular environment.

For this reasons, there is no value of T_s at which the entropy of the simulated sequences matches that of the experimental data. Natural sequences conserve non-K-sites much more than any simulation. On the other hand, K-sites are partially conserved similar to what simulations do in the intermediate temperature range.

Studying the Pearson correlation coefficient between the experimental and the simulated entropy per site (cf figure S8 in the supplementary material), it is clear that there is not a significant difference in correlation for temperatures $T_s < T_s^c \approx 1.1 \times 10^{-2}$.

In what follows, we shall focus our attention on temperature $T_s = 3.2 \times 10^{-3}$, which belongs to the intermediate regime as experimental data seem to do; at the same time, it is high enough that simulations are computationally fast.

3.4. Ruggedness of the landscape is mainly determined by changes in K-sites

An interesting feature of the energy landscape of sequences is its ruggedness. To investigate this point, we generated some artificial low-energy sequences, starting from infinite T_s, that is from random sequences and quenching the temperature to 3.2×10^{-3} (see table 2), studying the energy landscape along the trajectories that link them to the protein G sequence (1PGB in table 2).

We used a ratchet algorithm (see Methods, section 2.2) to generate trajectories at $T_s = 3.2 \times 10^{-3}$ between pairs of sequences. This algorithm

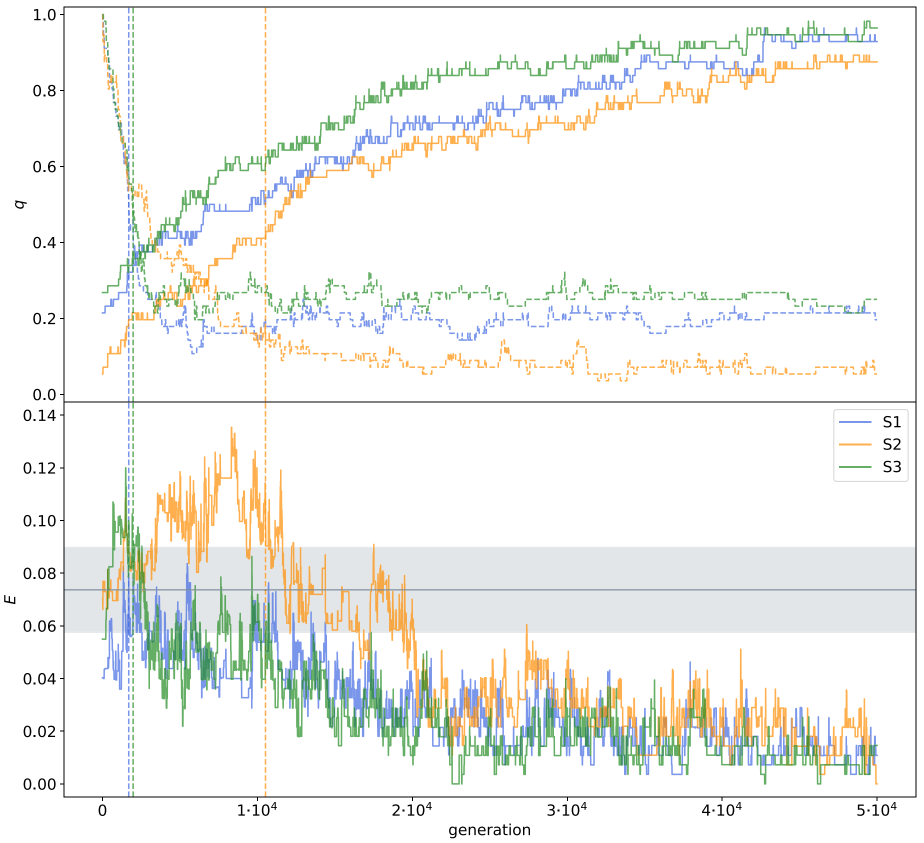


Figure 5. The Hamming similarity parameter q (upper panel) and the effective energy E (lower panel) along trajectories generated with a ratchet algorithm between the system and the 1PGB sequence at $T_s = 3.2 \times 10^{-3}$. The Hamming similarity parameter q is calculated from the initial (dashed curve) and from the target 1PGB (solid curve) sequence along each trajectory. The grey line and the shaded area in the lower panel are the mean energy and the associated standard deviation, respectively, at $T_s = 3.2 \times 10^{-3}$. The vertical dotted lines mark the time when the K-sites approached the target combination of the 1 PGB sequence. In the legend, the starting sequences for each trajectory (cf table 2).

does not push the sequence toward its target but only dumps fluctuations in the opposite direction. Consequently, we expect that it will not force the system to cross barriers higher than those that would cross spontaneously by thermal fluctuations [15].

Trajectories can leave the initial sequence in a few thousand mutations and reach the target sequence in less than 10^5 mutations (upper panel in figure 5).

The maximum energy reached by the simulation is in the range between 0.09 and 0.14 (lower panel in figure 5), which is larger than the spontaneous fluctuations that the system displays at this temperature, at which the mean effective energy is $\overline{E} = 0.074 \pm 0.016$ (cf figure 2). This fact indicates that the system can encounter relevant energy barriers along its motion. The peak in the energy is close to the time when the K-sites approach that of the target sequence (dashed vertical lines) (cf figure S9 in the supplementary material). The peak is largest for sequence S2, which displays the most different K-sites combination from that of the protein G (cf table 2).

After the K-sites are changed, all sequences take several tens of thousands generations to reach the target sequence. However, mutations of amino acids in sites that are not K-sites do not generate energy barriers comparable to thermal fluctuations.

Summing up, trajectories between low-energy sequences are neutral except for changes in K-sites, which generate barriers that are anyway surmountable by thermal fluctuations.

3.5. Local entropy of the basins

Proteins are expected to tolerate random mutations in order to be evolutionary fit [24]. Such tolerance can be characterized by the width of the neighborhood of a protein sequence $\boldsymbol{\sigma}$, quantified by its local entropy difference $\Delta S_{T_s,\gamma}(\boldsymbol{\sigma})$ (equation (7)), as done in the energy landscape of other kinds of complex systems [18]. Note that the definition of this entropy has nothing to do with the entropy $\mathcal{S}$ defined above to describe the conservation of amino acids in protein sites.

We calculated the local entropy of the basins defined by natural sequences and by low-energy sequences sampled by the Monte Carlo algorithm but not present in nature (cf table 2). For each natural sequence (see table 2), we first ran $\sim 10^4$ steps at $T_s = 3.2 \times 10^{-3}$ in order to obtain, for each basin, typical sequences for that temperature which still maintain the same K-sites of the starting natural ones. The sequences produced by this equilibration process, which is necessary to compare correctly the

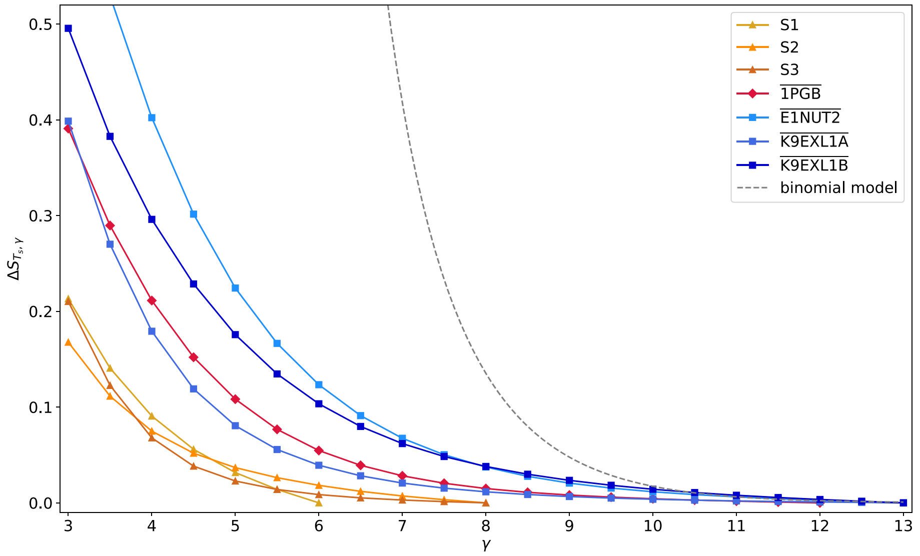


Figure 6. The local entropy difference $\Delta S_{T_s,\gamma}(\boldsymbol{\sigma})$ plotted as a function of the parameter γ for (equilibrated) natural sequences and artificial ones. The grey dotted line represents the theoretical local entropy difference for the binomial model at $T_s = 3.2\times10^{-3}$ (see table 1).

width of the basins within the framework of the canonical ensemble, are labeled with an overbar (cf figure 6 and table S1). We then proceeded to calculate the local entropy difference $\Delta S_{T_s,\gamma}$ as a function of the Lagrange multiplier γ that controls the average distance from the representative sequence, using equation (7).

Interestingly, at any value of γ the local entropy of the basins defined by natural sequences is significantly larger than those of artificial sequences (see figure 6). So, at any length scale around each $\boldsymbol{\sigma}$, natural sequences display a wider energy basin than artificial ones, while the associated energies are similar (cf table S1).

The entropy $\Delta S_{T_s,\gamma}$ is for all folding sequences markedly lower than that of the binomial model (dashed curve in figure 6). This is not unexpected, as the manifold sampled at realistic T_s is not convex, as would be if the binomial approximation were correct.

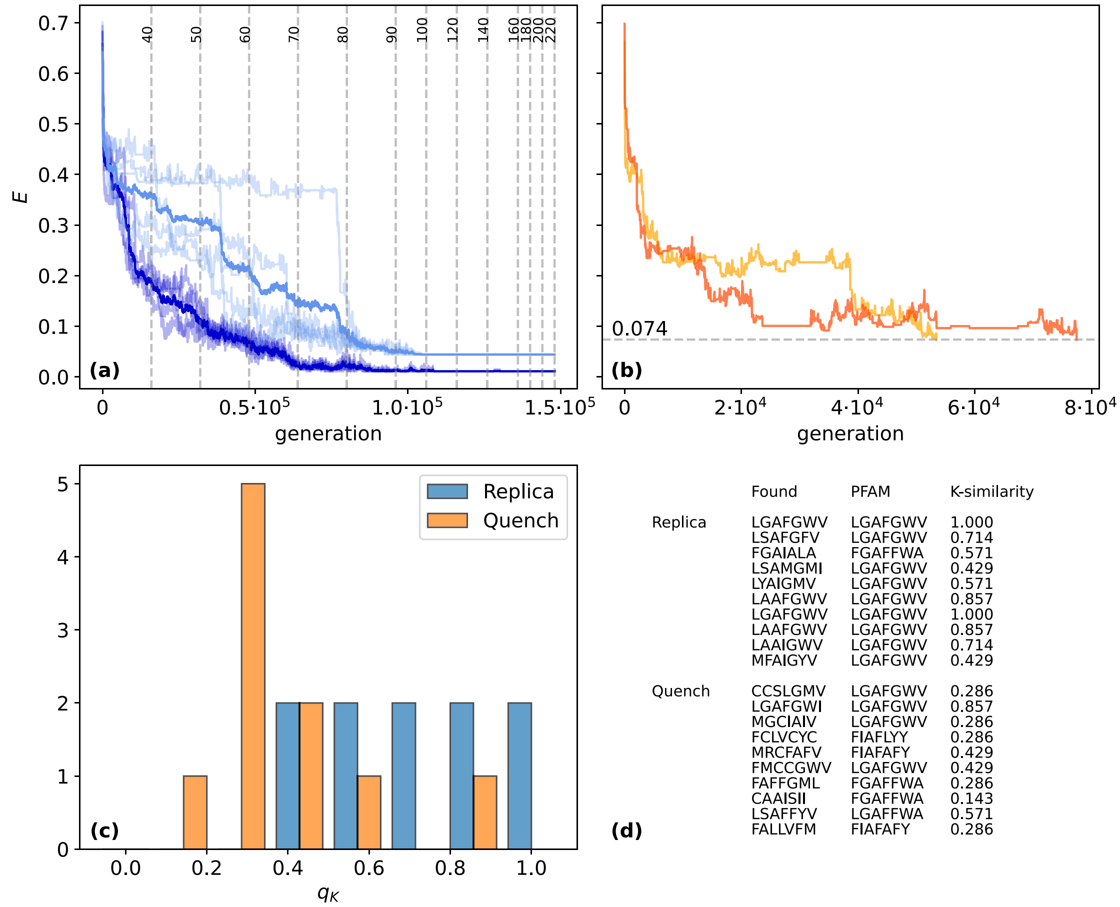
Matching the dependence of $\Delta S_{T_s,\gamma}$ on γ with that of $\overline{q}$ on γ, one can infer the dependence of $\Delta S_{T_s,\gamma}$ as a function of $\overline{q}$ (see figures S10 and s11 in the supplementary material). This curve displays a linear growth, indicating that the number of low-energy sequences in the neighborhood of each $\boldsymbol{\sigma}$ grows exponentially with the distance from it.

3.6. Searching for high-local entropy sequences

A relevant question is then whether there is a way to find efficiently sequences in wide energy basins, avoiding those that lie in narrow minima. In the field of artificial neural network, this goal was achieved sampling the space of the network parameters with replicas whose mutual distances are coupled together by the Lagrange multiplier γ and varying (annealing) slowly γ until the system converges to a unique set of parameters [18]. We have applied the same strategy to the space of protein sequences, as described in section 2.4.

Starting from random sequences, the system can converge to a unique sequence of low energy with annealings of the order of $\sim10^5$ steps (figure 7(a). We compared the K-sites of these sequences with those of sequences generated with quenches from infinite temperature to $T_s = 3.2\times10^{-3}$ (figure 7(b), recording a sequence when its energy reaches the average value at this temperature (cf figure 2). The quenching procedure is meant to obtain sequences in the neighborhood of the initial, random one but with an energy typical for the final temperature of the quench; not being allowed to explore massively the sequence space, we expect that these sequences are distributed evenly in sequence space. In this way, we build negative examples of sequences with the same average energy as those produced by the replica algorithm but that are not selected based on the local entropy of the energy basins where they lie.

In particular, we compared the K-sites of ten sequences obtained from the replica simulations with sequences obtained from ten temperature quenches. We defined q_K as the maximum Hamming similarity between the K-sites of a simulated sequence with those of any natural sequence taken from the PFAM database. In this way, $q_K(\boldsymbol{\sigma}) = 1$ if there is at least a natural sequence displaying the same K-sites of the simulated sequence $\boldsymbol{\sigma}$. The average $\overline{q_K}$ of the sequences obtained from the replica simulations is 0.71 ± 0.21, which is significantly larger than the value 0.39 ± 0.20 obtained from the quenches (cf figures 7(c) and (d). Furthermore, the p-value (obtained from a t-test) for the two distributions is 2.4×10^{-3}. Thus, sequences in large basins are more



	Found	PFAM	K-similarity
Replica	LGAFGWV	LGAFGWV	1.000
	LSAFGFV	LGAFGWV	0.714
	FGAIALA	FGAFFWA	0.571
	LSAMGMI	LGAFGWV	0.429
	LYAIGMV	LGAFGWV	0.571
	LAAFGWV	LGAFGWV	0.857
	LGAFGWV	LGAFGWV	1.000
	LAAFGWV	LGAFGWV	0.857
	LAAIGWV	LGAFGWV	0.714
	MFAIGYV	LGAFGWV	0.429
Quench	CCSLGMV	LGAFGWV	0.286
	LGAFGWI	LGAFGWV	0.857
	MGCIAIV	LGAFGWV	0.286
	FCLVCYC	FIAFLYY	0.286
	MRCFAFV	FIAFAFY	0.429
	FMCCGWV	LGAFGWV	0.429
	FAFFGML	FGAFFWA	0.286
	CAAISII	FGAFFWA	0.143
	LSAFFYV	LGAFFWA	0.571
	FALLVFM	FIAFAFY	0.286

Figure 7. (a) The effective energy of replicated simulations with $y = 5$. Transparent curves are the effective energy of each replica sequence, while the solid lines represent their mean. The vertical grey dotted lines indicate the value of γ along the annealing. (b) The effective energy of quenched simulations. The horizontal grey dotted line is the mean effective energy at $T_s = 3.2 \times 10^{-3}$ (reported in the bottom left corner). (c) The non-normalized distributions of q_K for the replicated (in blue) and quenched (in orange) simulations. (d) The K-sites found from the replicated and quenched simulations, compared to the most similar amino acids from the PFAM database and the associated q_K.

similar from the point of view of K-sites combinations to natural sequences than other ones selected at random (by the temperature quench) with comparable energy.

Molecular-dynamics simulations of a sequence found by the replica algorithm (cf figure S5 in the supplementary material) display a RMSD to the putative native structure of 0.14 ± 0.02 nm, which is the more similar to the wild-type sequence 1pgb than those obtained from a temperature quench.

It is also worth mentioning that if sequences are let evolve for $\approx 3 \times 10^5$ generations after the quench, they can reach the widest basins where natural sequences lie (cf figure S12 in the supplementary material). This makes the case of protein sequences different from other complex systems for which the replica algorithm was originally developed, where the system is unable to reach wide basins with simple Monte Carlo moves [18]. The main reason for this difference seems to be that the relevant dimensions of the sequence space are just the seven ones that define the K-sites. Thus, the effective dimension of this system is much smaller and it can therefore be sampled much more easily, than that of other complex systems.

We stress that the advantage of the replica algorithm to find wide basins is not much that of computational efficiency, that is marginal for a protein as small as protein G, but that of guaranteeing to ignore narrow basins, allowing us to distinguish the properties of wide minima (in the present case, the composition of K-sites) from those of minima at large. Our proof of concept however shows that the algorithm can be easily extended to the case of more complex, larger proteins.

4. Discussion

Characterizing the space of sequences folding to a well-defined native structure is useful to define the constraints that bind evolutionary trajectories of proteins. Recently developed machine-learning models like ESMFold are an efficient tool to define the landscape of foldable sequences.

A concern intrinsically associated with this approach is the reliability of the predictions of the machine-learning model for sequences that are far from natural ones. A feature of ESM-Fold that makes it particularly suitable for our goal is that, at variance with other predictors, it does not use information from alignments of homologous sequences. Its prediction does not stem from what amino acids are observed in the very same sites in evolutionary-related proteins, but from the overall information coming from all available protein structures, which gives good predictions even for test sets not containing sequences homologous to those used for the training (cf appendix B in [25]). Consequently, its performance is not expected to drop as the sampling departs from the set of naturally-observed sequences. In fact, the molecular dynamics simulations we did starting from the predicted native conformation of sequences very far from the natural ones proved very stable. On the contrary, AlphaFold [10] did not produce any reasonable structure given the same sequences, since no homologs are found. Moreover, an advantage of ESM-Fold with respect to Alphafold is that the former is remarkably faster (cf figure S13 in the supplementary material).

Assessing the ability of ESM-Fold to predict the folding ability of sequences that are distant from those in available families is a difficult task. A massive experimental study was performed with 228 proteins generated by a simulated annealing of the similarity to different target structures [26]. The 67% of these sequences, displaying no or weak similarity to known sequences, fold to soluble and monomeric structures.

Some indirect studies are also available in the literature. Analysis of *de novo* proteins, that is weakly-structured proteins that do not have known homologs, has shown that ESM-Fold displays the same variability of specific disorder-predictors in assessing the degree of folding of proteins far from the training set [27]. Using back-propagation to design novel sequences shows that while Alphafold generates sequences with unnatural profiles, ESM-Fold does not undergo this limitation [28].

The picture that emerges from our sampling of the sequence space of protein G is that foldable sequences form a wide basin that contains all natural homologs and a constellation of smaller basins that display similar effective energy as the main one but that are narrower, displaying lower entropy. The different basins are characterized by different combinations of amino acids in a limited number of specific sites, here termed K-sites. The other amino acids, not belonging to K-sites, are rather free to mutate, thus generating a connected set of well-folding sequences up to very large Hamming distance from the wild-type. Only mutations in the K-sites seems to generate energy barriers corresponding to poorly-folding sequences.

The presence in the sequence landscape of different basins characterized by specific choices of amino acids in few key sites of the protein was already found in minimal models [8] and in simplified models with knowledge-based potentials [29]. This fact suggests that it is not a consequence of the particular energy function used here, but it is a general feature of this kind of systems. Differently from what suggested in the case of simplified protein models, the key sites we found are not those critical for folding kinetics [19].

An important result of this study is that natural sequences folding to the structure of protein G belong to a wide basin, which maximizes the local entropy. One could hypothesize that being able to accumulate several mutations while maintaining the same native structure is an evolutionary advantage for a protein.

Wide energy basins can be found very efficiently with algorithmic schemes borrowed from the theory of complex systems. These do not seem to mimic in any way the evolutionary dynamics of proteins but are indeed a fast computational tool. For a small protein like protein G, we saw that it is possible to find the widest basin simply with a Monte Carlo algorithm in a manageable computational time, even without resorting to local entropy minimization. However, it seems unlikely that the same can be done for larger system, in which the dimension of the sequence space is larger. On the other hand, the algorithm for entropy-driven search can be made more efficient than in the present proof of concept in a number of ways [30].

The computational scheme described here can be extended in many ways, for example by modifying the effective energy to include other evolutionary constraints beyond that of foldability, such as thermodynamic stability or affinity for other molecules. In addition, the *a priori* move of the Monte Carlo algorithm can be modified to allow for insertions and deletions as well as mutations, making the length of the protein variable.

The structure of the energy landscape we found for protein sequences seems quite different from the typical landscapes of systems with complex interactions like spin glasses and constraint minimization problems, which are much more rugged and display a much larger number of well-separated basins. As a matter of fact, the ruling role that K-sites have on the effective energy of sequences makes their physical properties simpler than those of other complex systems.

Of course, foldability is just one of the constraints that the evolution of a protein must satisfy. Large language models can anyway be employed to define effective energies that encode other properties of sequences, like their thermodynamic stability or their binding properties to specific targets. The strategy developed here is agnostic of the physical meaning of the effective energy.

5. Conclusions

We defined an effective energy based on currently-available large language model and explored the energy landscape associated with the sequences folding to the native conformation of a small protein. This problem can be conveniently cast into the framework of the canonical ensemble of statistical physics, using the tools developed in this field. We found that folding sequences populate few basins of similar energy; one of them is much wider than the others and contain naturally-evolved sequences. Different basins are characterized by specific arrangement of the amino acids in a small subset of the sites of the protein. We showed that a computational algorithm based on replicated searchers can identify very efficiently the widest basins.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://dataverse.unimi.it/dataset.xhtml?persistentId = doi:10.13130/RD_UNIMI/WQ8QUA [31].

Conflict of interest

The authors declare no conflict of interest.

ORCID iD

G Tiana https://orcid.org/0000-0001-9868-1809

References

[1] Lässig M, Mustonen V and Walczak A M 2017 Predicting evolution *Nat. Ecol. Evol.* **1** 1–9

[2] Shakhnovich E I and Gutin A 1990 Implications of thermodynamics of protein folding for evolution of primary sequences *Nature* **346** 773–5

[3] Sander C and Schneider R 1991 Database of homology-derived protein structures and the structural meaning of sequence alignment *Proteins* **9** 56–68

[4] Mézard M, Parisi G, Sourlas N, Toulouse G and Virasoro M 1984 Nature of the spin-glass phase *Phys. Rev. Lett.* **52** 1156–59

[5] Mackenzie N D and Young A P 1982 Lack of ergodicity in the infinite-range ising spin-glass *Phys. Rev. Lett.* **49** 301–4

[6] Govindarajan fand G R A 1997 Evolution of model proteins on a foldability landscape *Proteins* **29** 461–6

[7] Shakhnovich E I and Gutin A M 1993 Engineering of stable and fast-folding sequences of model proteins *Proc. Natl Acad. Sci. USA* **90** 7195–9

[8] Tiana G, Broglia R A and Shakhnovich E I 2000 Hiking in the energy landscape in sequence space: a bumpy road to good folders *Proteins* **39** 244–51

[9] Govindarajan S and Goldstein R A 1997 Evolution of model proteins on a foldability landscape *Proteins* **29** 461–6

[10] Jumper J *et al* 2021 Highly accurate protein structure prediction with AlphaFold *Nature* **596** 583–9

[11] Lin Z *et al* 2022 Evolutionary-scale prediction of atomic level protein structure with a language model *bioRxiv Preprint* (https://doi.org/10.1101/2022.07.20.500902)

[12] Shakhnovich E I and Gutin A 1993 A new approach to the design of stable proteins *Prot. Eng.* **6** 793–800

[13] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 Equation of state calculations by fast computing machines *J. Chem. Phys.* **21** 1087–92

[14] Camilloni C, Broglia R A and Tiana G 2011 Hierarchy of folding and unfolding events of protein G, CI2 and ACBP from explicit-solvent simulations *J. Chem. Phys.* **134** 45105

[15] Tiana G and Camilloni C 2012 Ratcheted molecular-dynamics simulations identify efficiently the transition state of protein folding *J. Chem. Phys.* **137** 235101

[16] Baldassi C, Ingrosso A, Lucibello C, Saglietti L and Zecchina R 2015 Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses *Phys. Rev. Lett.* **115** 128101

[17] Dayhoff M O, Schwartz R M and Orcutt B C 1978 A model of evolutionary change *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation) ch 22, pp 345–52

[18] Baldassi C, Ingrosso A, Lucibello C, Saglietti L and Zecchina R 2016 Local entropy as a measure for sampling solutions in constraint satisfaction problems *J. Stat. Mech.* 023301

[19] McCallister E L, Alm E and Baker D 2000 Critical role of beta-hairpin formation in protein G folding *Nat. Struct. Biol.* **7** 669–73

[20] Ferrenberg A M and Swendsen R H 1989 Optimized Monte Carlo data analysis *Phys. Rev. Lett.* **63** 1195–8

[21] Tunyasuvunakool K *et al* 2021 Highly accurate protein structure prediction for the human proteome *Nature* **596** 590–6

[22] Robustelli P, Piana S and Shaw D E 2018 Developing a molecular dynamics force field for both folded and disordered protein states *Proc. Natl Acad. Sci. USA* **115** E4758–66

[23] Punta M *et al* 2012 The Pfam protein families database *Nucl. Acid Res.* **40** 290–301

[24] Guo H H, Choe J and Loeb L A 2004 Protein tolerance to random amino acid change *Proc. Natl Acad. Sci. USA* **101** 9205–10

[25] Hsu C *et al* 2022 Learning inverse folding from millions of predicted structures *Proc. 39th Int. Conf. on Machine Learning* pp 8946–70

[26] Verkuil R *et al* 2022 *bioRxiv Preprint* (https://doi.org/10.1101/2022.12.21.521521)

[27] Aubel M, Eicholt L and Bornberg–Bauer E 2023 Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning *F1000Research* (https://doi.org/10.12688/f1000research.130443.1)

[28] Jeliazkov J L, Dell'Alamo D and Karpiak J D 2023 ESMFold hallucinates nativelike protein sequences *bioRxiv Preprint* (https://doi.org/10.1101/2023.05.23.541774)

[29] Tiana G and Broglia R A 2009 The molecular evolution of HIV-1 protease simulated at atomic detail *Proteins* **76** 895–910

[30] Pittorino F, Lucibello C, Feinauer C, Perugini G, Baldassi C, Demyanenko E and Zecchina R 2021 Entropic gradient descent algorithms and wide flat minima *J. Stat. Mech.* 124015

[31] Tiana G 2023 Replication data for Structure of the space of folding protein sequences defined by large language models UNIMI Dataverse V2 (https://doi.org/10.13130/RD_UNIMI/WQ8QUA)