Mutational paths in protein-sequence landscapes: from sampling to low-dimensional characterization

Eugenio Mauri, Simona Cocco, Rémi Monasson

Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 and PSL Research, Sorbonne Université, 24 rue Lhomond, 75231 Paris cedex 05, France

(Dated: April 25, 2022)

Understanding how protein functionalities vary along mutational paths is an important issue in evolutionary biology and in bioengineering. We here propose an algorithm to sample mutational paths in the sequence space, realizing a trade-off between protein optimality and path stiffness. The algorithm is benchmarked on exactly solvable models of proteins in silico, and applied to data-driven models of natural proteins learned from sequence data. Using mean-field theory, we monitor the projections of the sequence on relevant modes along the path, allowing for an interpretation of the protein sequence trajectory. Qualitative changes observed in paths as their lengths are varied can be explained by the existence of a phase transition in infinitely-long strings of strongly coupled Hopfield models.

Introduction. Designing proteins with controlled properties, such as stability, binding affinity and specificity is a central goal in bioengineering. Directed evolution setups result in the discovery of new proteins with enhanced activities or affinities to a specific substrate [1]. Over the past years, much progress was made using data-driven models, intended to capture the relation between protein sequences and functionalities. In particular, unsupervised machine-learning approaches such as Boltzmann Machines (BM) or Variational Auto-Encoders trained on homologous sequence data (defining a protein family) were shown to be robust generative models, able to design new proteins with functionalities comparable to natural proteins [2, 3].

By comparison, the (even) harder problem of designing paths of sequences, interpolating between two homologous proteins has received little attention (Fig. 1), see however [4]. Yet solving this problem would be important from an evolutionary point of view, and would shed light on the navigability of the sequence landscape [5], and on how specificity emerged from ancestral, promiscuous proteins [6]. Informally speaking, a path is a succession of mutations interpolating between two fixed sequences at the edges, such that the intermediate proteins maintain good functionality. Due to the huge number of possible paths mutagenesis experiments generally restrict to direct paths going through the \(2^D\) mutants containing the amino acids appearing in the two edge sequences (differing on \(D\) sites), see Fig. 1 [7]. However, constraining paths to be direct may preclude the discovery of much better global paths, involving mutations and their reversions and reaching more favorable regions in the sequence space (Fig. 1).

While various methods exist for building transition paths between the minima of a multi-dimensional continuous landscape [8, 9] they cannot be easily adapted to the case of discrete configurations. We hereafter propose a Monte Carlo algorithm to sample mutational paths in sequence space. We first benchmark our sampling procedure on an exactly solvable model of lattice proteins [10], and demonstrate its capability to find high-quality paths between two proteins belonging to different sub-families. We then apply our algorithm to the WW domains, a small binding module involved in the regulation of protein complexes [11] and studied in early works on sequence-based design [12]. The paths obtained between proteins with different specificities have high likelihoods and folding scores according to AlphaFold [13]. Furthermore, we show how mean-field theory can be applied to track informative projections of the multi-dimensional trajectories of sequences along the paths. Last of all, we observe that global paths, if long enough, significantly outperform direct paths. This crossover is related to the existence of a thermodynamic phase transition in infinitely-long strings of coupled Hopfield-like models that we analytically solve.

Algorithm for mutational path sampling. We assume the sequence landscape is modeled through a probability distribution \(P_{\text{model}}(\mathbf{v})\) over amino-acid sequences \(\mathbf{v}\) of length \(L\). Informally speaking, \(P_{\text{model}}\) quantifies the probability that \(\mathbf{v}\) is a member of the protein family of interest, i.e. share its common structural and functional properties, and can be learned from homologous sequence data [14, 15]. For natural protein families, exact
expressions for $P_{\text{model}}$ are not available, but approximate distributions can be inferred from multi-sequence alignments (MSA) using unsupervised learning techniques.

Hereafter, we use Restricted Boltzmann Machines (RBM) [16], a class of generative models based on two-layer graphs [17]. RBM define a joint probability distribution of the protein sequence $\mathbf{v}$ (carried by the visible layer) and of its $M$-dimensional latent representation $\mathbf{h}$ (present on the hidden layer) as

$$P_{\text{RBM}} \propto \exp \left( \sum_i g_i(v_i) + \sum_{\mu} h_{\mu} I_{\mu}(\mathbf{v}) - \sum_{\mu} U_{\mu}(h_{\mu}) \right),$$

where $I_{\mu}(\mathbf{v}) = \sum_i w_{i,\mu} v_i$ is the input to hidden unit $\mu$. The $g_i$'s and $U_{\mu}$'s are local potentials acting on, respectively, visible and hidden units, and the $w_{i,\mu}$'s are the interactions between the two layers. They are learned by maximizing the marginal probabilities $P_{\text{model}}(\mathbf{v}) = \int d\mathbf{h} P_{\text{RBM}}(\mathbf{v}, \mathbf{h})$ over the sequences $\mathbf{v}$ in a multi-sequence alignment of the family. While other unsupervised procedures providing approximate $P_{\text{model}}$ can be used, such as Direct Coupling Analysis [14, 15], RBM offer a convenient way to interpret and to visualize the changes in sequences along mutational paths, as we will see below.

We define the probability of a mutational path of $T$ sequences, $\mathcal{V} = \{v_1, v_2, ..., v_T\}$ through

$$\mathcal{P}[\mathcal{V} | v_{\text{start}}, v_{\text{end}}] \propto \prod_{t=1}^{T} P_{\text{model}}(v_t) \times \prod_{t=1}^{T-1} \pi(v_t, v_{t+1}) \times \pi(v_T, v_{\text{end}}),$$

where $\pi(v, v') = 1$ if the sequences $v$ and $v'$ are identical, $e^{-\Lambda}$ if they differ by one mutation (with $\Lambda > 0$), and 0 if they are two or more mutations apart. The probability $\mathcal{P}[\mathcal{V}]$ can be sampled as follows. Starting from a path $\mathcal{V}^0$, we randomly pick up an intermediate sequences $v_t$ and attempt mutating one amino acid, under the constraint that the Hamming distances of the trial sequence $v'$ with $v_{t-1}$ and $v_{t+1}$ be at most 1. The mutation is then rejected or accepted, i.e., $v_t \leftarrow v'$ according to detailed balance. Note that for global paths amino acids can take any values. For direct paths each amino acid has to coincide with the one either in $v_{\text{start}}$ or in $v_{\text{end}}$ on the same site, and the length $T$ of the path matches the Hamming distance $D$ between the two edge sequences.

To improve the quality of the sampled mutational paths we introduce a fictitious inverse temperature $\beta$ and resort to simulated annealing. We then sample paths from $\mathcal{P}[\mathcal{V}]^\beta$, where the value of $\beta$ is initially very small and progressively ramped up to some target value. The complete procedure and the proof of detailed balance are given in Supplemental Material, Sec. ??.

**Benchmarking mutational path sampling on in silico proteins.** We benchmark the performances of our MC procedure on a model of Lattice Proteins (LP) [10, 18]. In LP a sequences of 27 amino acids may fold into $\simeq 10^5$ different self-avoiding conformations going through the nodes of a $3 \times 3 \times 3$ cubic lattice. The sequence landscape associated to a conformation $\mathbf{S}$ (Fig. 2(a)) is defined by the probability $p_{\text{nat}}(\mathbf{v}|\mathbf{S})$ that a sequence $\mathbf{v}$ has $\mathbf{S}$ as its native fold; $p_{\text{nat}}$ can be exactly computed from the energies of interactions between adjacent amino acids, see Supplemental Material, Sec. ?? for details.

We first generate many sequences $\mathbf{v}$ with high $p_{\text{nat}}$ values for the fold $\mathbf{S}$ of Figs. 2(b,c) following the procedure of [19]. We next compute the top two Principal Components (PC) of these sequence data (Figs. 2(d,e)); PC1 corresponds to an extended electrostatic mode, and PC2 identifies possible Cys-Cys bridges. Projecting the sequences onto these two PCs reveals two sub-families separated along PC1 (Fig. 2(a)), associated to opposite chains of alternating charges along the electrostatic mode (Figs. 2(b,c)). We will use our path sampling procedure to interpolate between the two sub-families, see start (white star) and end (black star) sequences in Fig. 2(a).

To mimick the procedure followed for natural proteins we train a RBM on the LP sequence data generated above, to infer an approximate expression for $p_{\text{nat}}$ from the data; see Supplemental Material, Sec. ?? for details about the inference of the RBM model. We then use our sampling algorithm to produce global mutational paths,
The algorithm is able to find excellent global mutational paths in terms of the ground truth folding probability $p_{nat}$. By fixing the target inverse temperature $\beta$ to a value larger than one, we are able to obtain $p_{nat}$ values along the path higher than those of the extremity sequences. Repeated runs of the sampling procedure give different paths that cluster into two classes, shown in red and maroon in Fig. 2(a). While few global paths boost $p_{nat}$ by transiently introducing Cys-Cys interactions (maroon cluster), most (red cluster) stabilize the structure by realizing more contacts between positively and negatively-charged amino acids than direct paths (Supplemental Material, Sec ??).

**Mutational path sampling from data-driven models of natural proteins.** We next show that our path sampling procedure can be applied to natural proteins. To do so we train a RBM from MSA data of the WW family, a protein domain binding specifically proline-rich peptides [11, 20] and sample mutational paths, either global or direct, between the Human YAP1 domain and three natural sequences known to have different binding specificities [22]. The quality of the sequences along the path is assessed from their probabilities $P_{RBM}$ within the RBM model, and from 3D structure predictions obtained using AlphaFold [13]. Figure 3(a) shows some sampled paths in the 2-dimensional space spanned by the inputs $I(v)$ to two RBM hidden units chosen to cluster natural WW sequences depending on their binding specificities [17]. Figures 3(b,c) show the probabilities of sequences along global and direct paths are comparable to the ones of natural proteins, with significantly higher values for global paths. We report AlphaFold’s confidence scores of intermediate sequences along global paths in Fig. 3(d), indicating that these sequences have well defined folds. Furthermore, we compare these predicted folds to that of natural WW using Template Modelling scores (TM-score) [23], which measure structure similarity from 0 - unrelated proteins- up to 1 - perfect match. We obtain TM-score $> 0.5$, indicating a high similarity between the folds of sequences along the path and of natural WW.

**Mean-field characterization of mutational paths.** To understand how mutational paths explore the sequence space we introduce a mean-field theory exploiting the bipartite nature of the RBM architecture. Mean field allows us to introduce two sets of order parameters characterizing the paths $\mathcal{V}$: the mean values of the hidden-unit inputs, $m_i^\mu = \frac{1}{N} \langle \mu_i(v) \rangle$, and of the overlaps (fraction of conserved amino acids between successive sequences), $q_i = \frac{1}{5} \sum_k \langle \delta_{v_i, v_{i+1}} \rangle$; here, $\langle \cdot \rangle$ denotes the average over $P(\mathcal{V})$.

In the mean-field framework a step along the path can involve multiple mutations. The transition factor $\pi$ in Eq. (2) is defined through

$$\pi(\mathbf{v}, \mathbf{v'}) = e^{-N \Phi(q)}$$

where

$$q = \frac{1}{N} \sum_i \delta_{v_i, v'_i}$$

the potential $\Phi$ fords small overlaps $q \ll 1$, i.e. discontinuous jumps along the paths. We impose $q > q_c = 1 - \gamma/T$, allowing for the path to explore at most $T \times N(1-q_c) = \gamma N$ mutations in $T$ steps. Choosing $\gamma > D/N$ is therefore sufficient to interpolate between the two edge sequences, with larger values of $\gamma$ authorizing more flexible paths. In practice we set $\Phi(q) = 1/(T^2 (q - q_c))$; Other choices of potentials with hard wall constraints give similar results.

The $T \times (M + 1)$ order parameters $m_i^\mu$ and $q_i$ are determined through minimization of the path free-energy density $f_{path}$, see Supplemental Material, Sec ??, with

$$f_{path}(\{m_i^\mu\}, \{q_i\}) = -\sum_{t, \mu} (\Gamma_\mu(m_t^\mu) - m_t^\mu \Gamma'_\mu(m_t^\mu))$$

$$+ \sum_t (\Phi(q_t) - q_t \Phi'(q_t)) - \frac{1}{\beta N} \sum_i \ln Z_i(\{m_i^\mu\}, \{q_i\}).$$

Here, $\Gamma_\mu(m) = \frac{1}{\beta N} \ln \int dh e^{N m h - U_\mu(h)}$ and $Z_i$ is the following site-dependent partition function,

$$Z_i(\{m_i^\mu\}, \{q_i\}) = \sum_{\{v_i\}} \exp \left( \sum_t \left( g_i(v_t) + \beta \sum_{t, \mu} \Gamma'_\mu(m_t^\mu) w_{i, \mu}(v_t) - \beta \sum_t \Phi'(q_t) \delta_{v_i, v_{i+1}} \right) \right).$$

$Z_i$ can be efficiently estimated through products of $A \times A$-dimensional transfer matrices, where $A$ is the number of Potts states. For global paths, $A = 21$ (20 amino acids plus the gap symbol), while $A = 2$ for direct paths. The
FIG. 4. Mean-field description of mutational paths in lattice proteins. (a)-(b) Values of two inputs similar to PC in Fig. 2 vs. number of mutations along paths of length $T = 40$. Red and green lines correspond to, respectively, global and direct paths. Parameters: $\beta = 3$, $\gamma = 3.5$ chosen so that the average distance to the direct space of the mean-field solutions is the same as in Fig. 2. (c)-(d) Logos of the attached weights $w_{i,j}(v)$. (e) Log-likelihoods of sequences along the same paths as in panels (a),(b). (f) Overlap (right scale) between sequences at steps $t$ and average number of mutations (left scale) along direct paths are distinct $q_t$ (left scale) and average number of mutations $D_H = N(1 - q_t)$ (right scale) between sequences at steps $t$ and $t + 1$ vs. $t$. The dark line shows $q_c$.

derivation of $f_{\text{path}}$ is exact when the sequence length $N \to \infty$ and the numbers of hidden units, $M$, and of steps, $T$ remain finite, and is an accurate approximation even in the cases of LP ($N = 27$) and WW ($N = 31$), as shown below.

The trajectories of the inputs $m_i^a$ and of the overlaps $q_t$ reveal which and when latent factors of RBM enter into play throughout the interpolation between the initial and final sequences. Figures 4(a,b) show the trajectories of inputs associated to the weights in Figs. 4(c,d) for the lattice protein studied in Fig. 2. The dynamics explains how optimal paths exploit Cysteine-Cysteine interactions (not present in the initial and final sequences) in order to maintain the structure of the protein when the signs of the charge along the electrostatic chain are reversed (Fig. 4(a,b)), and agrees with the average behaviour of the paths sampled in Fig. 2(a), see Supplemental Material, Sec. ??.

Sequences along global paths have substantially higher probabilities than along direct paths (Fig. 4(e)). The exploration of favourable regions in the landscape is made possible by the slightly higher number of mutations between successive sequences in the former case than in the latter, see Fig. 4(f). Along global paths, most of the intermediate mutational steps do not abruptly affect the inputs nor the probability, with the exception of the bump in the overlap $q$ at step $\sim 10$, possibly related to the presence of preparatory mutations for the Cys-related transition in Fig. 4(b,d).

Crossover between global and direct paths. The results reported in Figs. 2, 3, 4 indicate that sequences along global mutational paths have larger scores than along direct paths, see Fig. 3(b,c). To better understand the differences between global and direct paths we introduce and analyze in details a toy-model capturing the effects of extra dimensions with respect to the direct subspace of sequences. This Hopfield-Potts (HP) model includes $M = 2$ patterns, and $A \geq 3$ Potts symbols. Denoting the first three symbols by $a, b, c$, the patterns are set to $w_{i,j}(v) = \delta_{a,a} + \omega \delta_{c,a}$ and $w_{2,j}(v) = \delta_{c,b} + \omega \delta_{c,c}$, uniformly over sites $i$, and define the sequence distribution $P_{HP}(v) \propto \exp \left[ \frac{1}{2} \sum_{i,j} \sum_{\mu=1,2} w_{i,j}(v_i) w_{\mu,j}(v_j) \right]$. The initial and final sequences are chosen, respectively, as $v^\text{start} = a$ and $v^\text{end} = b$; hence, $\omega$ quantifies the attractiveness of the global direction $c$ (orthogonal to the direct subspace spanned by $a, b$). We then couple $T$ such HP models to form a 1D-string with controlled stiffness (through the transition factors $\pi$ in Eq. (3)), and anchored in $v^\text{start}$ and $v^\text{end}$.

As HP models are a special case of RBM with quadratic potentials $U(h) \propto h^2$ [24] the path free-energy for trajectories over the inputs and the overlaps in Eq. (5) is exact when $N \to \infty$. The optimal trajectories can be analytically studied in great details, see Supplemental Material, Sec. ??, with the following results. For $\omega < \omega_c = \frac{1}{2}$, mutational paths typically lie within the direct subspace (Fig. 1): the attraction along the $c$ direction is too weak to counterbalance the stiffness of the path imposed by the potential $\Phi$. For $\omega > \omega_c$ optimal paths leave the direct subspace and explore the global space if their lengths exceed $T_{c,o} = \max \left( \frac{1}{\gamma^2 \Theta (\omega - \frac{1}{3})}, \frac{1}{(\gamma - 1)^2 (\omega - \frac{1}{3})} \right)$, where $1/\gamma < \Theta \leq 1$ is the fraction of the $T$ steps in which the sequences $v^t$ along direct paths are distinct from $v^\text{start}$ and $v^\text{end}$, see Supplemental Material, Sec. ??.

The resulting phase diagram is shown in Fig. 5(a). As expected, for lower values of $\gamma$, paths become stiffer, and $T_{c,o}$ increases. The average distance to the direct subspace, $d_{DS} = \frac{1}{N} \sum_i (1 - \delta_{\delta_{c,i}^\text{start}}, 1 - \delta_{\delta_{c,i}^\text{end}})$, and the average log-probability of intermediate sequences, $\langle \frac{1}{N} \log P_{HP} \rangle$ are shown in Fig. 5(b); their behaviours confirm the existence of the crossover at $T_{c,o}$. This crossover is also observed for natural proteins, such as WW, when keeping the length $T$ fixed and varying the flexibility $\gamma$ of the path, see Fig. 5(c). As in the HP model case the non-orthogonality of the weight vectors learned by the RBM provides multiple opportunities for mutational paths to escape the direct subspace.

Conclusion. In this work we have shown how data-driven, in particular, RBM models of protein sequence data can be used to sample mutational paths. Though our sampling procedure was illustrated on short proteins, it can be easily applied to longer enzymes, with $> 100$
FIG. 5. Crossover between direct and global mutational paths. (a). Behavior of $T_c, \omega$ vs. $\omega$ for the HP model and two values of $\gamma$, see Eq. (6). The black dots show the crossovers for $\omega = 3/4$. (b) $d_{DS}$ (Top) and $(\log P_{MF})/N$ averaged over intermediate sequences (Bottom; solid line: global, dashed: direct) vs. path length $T$; same parameters as in (a). (c) Mean-field estimates of $d_{DS}$ (Top) and of $(\log P_{RBM})/N$ (Bottom; red: global paths, green: direct) vs. $\gamma$ for mutational paths of the WW domain of length $T = 10$. Initial sequence: YAP 1, final sequence: green triangle in Fig. 3. In all panels $\beta = 3$.

amino acids, whose functionalities could be experimentally tested.

The analytical study of a toy Hopfield-Potts model reveals the existence of a qualitative change in mutational paths with their length. Long paths can explore favorable detours in the global sequence landscape, which are not accessible to shorter paths. An illustration, in the WW case, is the high-quality global green path going through a region with few natural sequences in Fig. 3(a). It would be very interesting to test this striking prediction experimentally.

In addition, the use of mean-field theory allows us to follow the dynamics of relevant latent factors along the paths, and understand how the transition from one functionality to another is implemented through sequential changes of few residues at a time. Extending our mean-field analysis to the case of an extensive number of RBM weight vectors (finite $M/N$) would allow for better monitoring the dynamics of the few inputs of interest along paths interpolating between subfamilies.

Acknowledgements. We are grateful to M. Bisardi, A. Di Gioacchino, A. Murugan, R. Ranganathan, J. Tubiana and F. Zamponi for discussions. This work was supported by the ANR-17 RBMPro and ANR-19 Decrypted CE30-0021-01 projects. E.M. is funded by a ICFP Labex fellowship of the Physics Department at ENS.
Supplemental Material

Mutational paths in protein-sequence landscapes: from sampling to low-dimensional characterization

Eugenio Mauri, Simona Cocco, Remi Monasson
April 25, 2022

1 Path sampling algorithm

1.1 Detailed description

We introduce below the details of the sampling procedure we use to obtain paths connecting two fixed sequences in a landscape described by the probability distribution \( P_{\text{model}} \). Starting from a path \( \{v_t\} \), we look at intermediate sequences (starting from \( t = 1 \)) and propose a mutation with the constraint that the Hamming distance \( D_H \) between \( v_{t-1} \) and \( v_{t+1} \) is not greater than 1.

We accept this move with a probability fixed to ensure detail balance. Different cases have to be considered, depending on the Hamming distance \( D_H \) between the new attempted sequence and existing ones:

- **\( D_H(v_{t-1}, v_{t+1}) = 0 \).** In this case the new sequence \( v_t' \) can have a single mutation at any site, compared with the two adjacent sequences along the path. Hence, we first draw a random site \( i \), then we propose a new sequence \( v_{t-1}' \) amino-acids for that site \( v_t' \) drawn from the distribution \( \propto P_{\text{model}}(v_t') \), where the amino acids are fixed to the same site \( i \). Then if the old sequence \( v_t \) already had a mutation with respect to \( v_{t-1} \) at given site \( j \), we accept the new mutated sequence \( v_{t+1}' \) (which is equal to \( v_{t-1} \) apart from the amino acid at site \( i \)) with a probability

\[
p_{\text{acc}}(v_t \rightarrow v_{t+1}) = \begin{cases} 
1, & \text{if } j = i \\
\frac{\pi_{\text{tr}}(v_{t-1}', v_{t+1})^{2\beta}}{\pi_{\text{tr}}(v_{t-1}, v_{t+1})^{2\beta}} \sum_{z=1}^{M} P_{\text{model}}^z(M_i^z v_{t-1}), & \text{otherwise}
\end{cases}
\]

where \( M_i^z \) indicates the mutation \( z \) at site \( i \). If \( v_t \) or \( v_{t+1}' \) are equal to \( v_{t-1} \), then the acceptance probability is \( p_{\text{acc}}(v_t \rightarrow v_{t+1}) = \min(1, \pi_{\text{tr}}(v_{t-1}', v_{t+1})^{2\beta}/\pi_{\text{tr}}(v_{t-1}, v_{t+1})^{2\beta}) \).

- **\( D_H(v_{t-1}, v_{t+1}) = 1 \).** In this case the new sequence \( v_t' \) can have a single mutation only at the site \( i \) where \( v_{t-1} \) and \( v_{t+1} \) are different. At that site, we propose a new mutation from the distribution \( \propto P_{\text{model}}(v_t') \) and accept it with probability \( p_{\text{acc}} = \exp[-\Lambda \beta(D_H(v_t, v_{t-1}) + D_H(v_t, v_{t+1}) - D_H(v_t, v_{t-1}) - D_H(v_t, v_{t+1}))] \).

- **\( D_H(v_{t-1}, v_{t+1}) = 2 \).** In this case the previous and subsequent sequence present two mutations at site \( i \) and \( j \). The new sequence \( v_t' \) can be of two forms: it can have the same mutation of \( v_{t+1} \) (with respect to \( v_{t-1} \)) at site \( i \) or at site \( j \). Hence, we extract one of the two possibilities with a probability weighted accordingly with \( P_{\text{model}}^z(v_t') \).

1.2 Proof of detailed balance

To prove that dynamics given by our algorithm of path sampling converges to the target distribution we have to prove that it respects detailed balance, i.e. the reversibility of each Markov step. We consider the transition from a path \( \{v_t\} \) to a new path that differ only by one sequence \( v_t' \) at time \( t \). We write the detailed balance condition as
\[ P_{\text{path}}(\{v_t\}) \rhd P_{\text{trans}}(v_t \rightarrow v'_t) = P_{\text{path}}(\{v'_t\}) \rhd P_{\text{trans}}(v'_t \rightarrow v_t) \]

\[ \pi_\lambda(v_{t-1}, v_t) \pi_\lambda(v_t, v_{t+1}) P_{\text{model}}(v_t) P_{\text{trans}}(v_t \rightarrow v'_t) = \pi_\lambda(v_{t-1}, v'_t) \pi_\lambda(v'_t, v_{t+1}) P_{\text{model}}(v'_t) P_{\text{trans}}(v'_t \rightarrow v_t) \] (2)

If \( D_H(v_{t-1}, v_{t+1}) = 0 \), the new sequence can have a mutation at any site \( i \) compared to its neighbour \( v_{t-1} \), while \( v_t \) will have the mutation at another site \( j \) (note that \( i \) and \( j \) can be equal). Hence the transition probability in this case will be

\[ p_{\text{trans}}(v_t \rightarrow v'_t) = \frac{1}{N} \sum_{z=1}^Q P_{\text{model}}(v'_t) \pi_{\text{acc}}(v_t \rightarrow v'_t), \quad p_{\text{trans}}(v'_t \rightarrow v_t) = \frac{1}{N} \sum_{z=1}^Q P_{\text{model}}(v_t) \pi_{\text{acc}}(v'_t \rightarrow v_t). \] (3)

By substituting everything in the detailed balance condition we obtain the acceptance probability described in the main paper. This will hold similarly when \( D_H(v_{t-1}, v_{t+1}) = 1 \) (with \( i = j \)). For \( D_H(v_{t-1}, v_{t+1}) = 2 \), the sequences \( v_t \) and \( v'_t \) can either be equal to \( v_{t+1} \) or \( v_{t-1} \), from which the condition presented in the paper descends.

2 Lattice Proteins

To benchmark the performances of this MC procedure to find good transition path between two sequences, we test it on Lattice Proteins [LD89], a well known toy-model for protein structure. We consider a protein sequence of 27 amino acids folding into a 3D structure specified as a self-avoiding path over a 3x3x3 lattice where each amino acid occupies one node. The probability of a sequence \( v \) to fold into a specific structure \( S \) is given by the interaction energies between amino acids in contact in the structure (i.e. those who occupy neighbouring nodes of the lattice, but are not adjacent in the protein sequence). In particular, the total energy of a sequence with respect to a given structure is given by

\[ E_{\text{LP}}(v|S) = \sum_{i<j} c_{ij}^S E_{M,J}(v_i, v_j) \] (4)

where \( c_{ij}^S \) is the contact map (\( c_{ij}^S = 1 \) if sites are in contact and 0 otherwise), while the pairwise energy \( E_{M,J}(v_i, v_j) \) represents the amino-acid physico-chemical interactions given by the the Miyazawa-Jernigan knowledge-based potential [MJ96]. The probability to fold into a specific structure is written as

\[ p_{\text{nat}}(S|v) = \frac{e^{-E_{\text{LP}}(v|S)}}{\sum_{S'} e^{-E_{\text{LP}}(v|S')}} \] (5)

where the sum is over the entire set of self-avoiding path in the cubic lattice.

The function \( p_{\text{nat}} \) represents a suitable landscape that maps each sequence to a score measuring the quality of its folding. To study in more detail this landscape, we will consider an alignment of sequences folding into a specific structure (that we will call \( S \)) sampled from a low temperature MC sampling using \( -\beta \log p_{\text{nat}}(\cdot|S) \) (with \( \beta = 10^5 \)) as effective energy [JGS+16].

3 Restricted Boltzmann Machines and training parameter

To study the problem of transition paths we first need a model to infer a landscape from our sequence data set. At this scope, we are going to use Restricted Boltzmann Machines, an unsupervised energy-based model able to learn representations of the data in a two-layer bipartite graph [FI12]. The first "visible" layer represents the protein sequence \( v = \{v_1, ..., v_N\} \) where each unit takes one out of 21 possible states (20 amino acids + 1 alignment gap). The second is the "hidden" layer which displays the real-valued representations \( h = \{h_1, ..., h_M\} \). The joint probability distribution for \( v \) and \( h \) is

\[ P_{\text{RBM}}(v, h) \propto \exp \left[ N \sum_{i=1} g_i(v_i) + \sum_{i,\mu} w_{i\mu}(v_i) h_\mu - \sum_{\mu} \mathcal{U}(h_\mu) \right] \] (6)

up to a normalization constant. Visible and hidden units are coupled through the matrices \( w_{i\mu} \) and the value of each unit is biased by the local fields \( g_i \) and \( \mathcal{U}_\mu \). In [TCM19] it has been shown that this model is able to recover statistically relevant
sequence motifs playing crucial roles in the structure and functionality of different protein families. Following their approach, we choose $U_{\mu}$ to be double Rectified Linear Unit (dReLU) potentials of the form

$$U_{\mu}(h) = \frac{1}{2} \gamma_{\mu,+} h_+^2 + \frac{1}{2} \gamma_{\mu,-} h_-^2 + \theta_{\mu,+} h_+ + \theta_{\mu,-} h_-,$$

where $h_+ = \max(h, 0)$, $h_- = \min(h, 0)$, and we have defined the hyper-parameters $\gamma_{\mu,\pm}$, $\theta_{\mu,\pm}$.

At this point, we need a learning procedure to infer the hyper-parameters that best fit our data. We decided to use a Persistent Contrastive Divergence algorithm [Tie08] which has been shown to be sufficiently good and robust under cautious choice of the regularization hyper-parameters [Tub18a]. The code and the data used to train our RBMs for Lattice Proteins and WW domains can be found in [Tub18b]. The hyper-parameters used for learning are the following:

- For WW ($N=31$):
  - $M = 50$
  - Batch size = 100
  - Number of epochs = 500
  - Learning rate = $5 \times 10^{-3}$ (which has a decay rate of 0.5 after 50% of iterations)
  - $L_1 b$ regularization = 0.25
  - Number of MC step between each update = 10

- For Lattice Protein ($N=27$):
  - $M = 100$
  - Batch size = 100
  - Number of epochs = 100
  - Learning rate = $5 \times 10^{-3}$ (which has a decay rate of 0.5 after 50% of iterations)
  - $L_1 b$ regularization = 0.025
  - Number of MC step between each update = 5

4 Statistic of paths sampled in LP

Here we show the relation between the sequences sampled using the sampling procedure described above (using the RBM as model) with the mean-field solution described in the main paper. Looking at Figure S1 we see, consistently with the mean-field solution, that global solutions prefers to activate input 14 more than the direct solution. Furthermore, statistically this input is more likely to be reduced before activating input 4, which is consistent with the mean-field solution.

Analysing the difference between the direct and global paths along each inputs we notice that global solutions maintain high scores in terms of $p_{nat}$ by exploiting the interactions between amino-acids at sites 5,6,11 and 22 (see Figure 2 in the main paper). Global paths are divided in two classes corresponding to the chemical nature of the interaction used to bind these amino-acids. One cluster (shown in maroon in Figure 2 of the main paper and in Figure S1) uses Cys-Cys bridges to establish this interactions. Instead the most populated cluster (shown in red in the same figures) exploits electrostatic interactions that are not initially present in the target sequences (hence forbidden along direct paths). Some of these interactions are shown in Figure S2. This explains why the Principal Component Analysis shown in Figure 2 of the main paper does not discriminate between direct paths and most of the global ones (i.e. red cluster).

To deeper characterize the behaviour of global solutions, we also plot in Figure S3 the average distance from the direct space as a function of time obtained from the paths sampled using $P_{model} \propto P_{RBM}$, the likelihood from the trained RBM model.
Figure S1: Plot of some relevant inputs as function of time sampled with our Monte-Carlo procedure ($\Lambda = 2$ and target $\beta = 3$). The colors respect those presented in Figure 2 in the main paper. Black lines correspond to the average time at which the input switch value ($>0$ for $I_4$ and $<2$ for $I_{14}$).

Figure S2: Plot of some relevant inputs (and their respective weights logos) as function of time sampled with our Monte-Carlo procedure ($\Lambda = 2$ and target $\beta = 3$) exploiting relevant electrostatic interactions forbidden in the direct space. The colors respect those presented in Figure 2 in the main paper.
Figure S3: Average value of the distance from direct space as a function of the step along the path connecting the two modes in the LP landscape. Black and White stars refer to the same sequences as in Figure 2 of the main text.

5 Additional information about mutational paths of the WW domain

5.1 Lists of the tested sequences

Here we present the reference sequences sampled with the MC algorithm and tested using AlphaFold (note that the first sequence of each list represent the YAP1 wild-type sequence from [ES99], while the last is the natural wild-type specific for each class):

- From YAP1 to wild-type of class II/III:
  - LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP
  - LPAGWEMARTSD-GQVYFINHNTQTTTWQDP
  - LPPGWQEARTPD-GRVYYINHNTKTTTWTKP
  - -PPWEQARTPD-GRVYYMNHNTKTTTWTKP
  - ---EWQEARTPD-GRVYYYNHNTKQTTTWTKP
  - ---EWQEAKTPD-GRVYYYNKNTKQTTWEKP
  - ---EWQEFKTPA-GKKYYYNKNTKQSRWEKP

- From YAP1 to wild-type of class IV:
  - LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP
  - LPPGWEMARTSD-GQVYFINHNTQTTTWQDP
  - LPPGWEMARSD-GQVYFINHNTQTTTWQDP
  - LPPGWEMARSD-GQVYFINHNTQTTTWQDP
  - LPPGWEMARSD-GQVYFINHNTQTTTWQDP
  - LPPGWEMARSD-GQVYFINHNTQTTTWQDP
  - LPPGWEMARSD-GQVYFINHNTQTTTWQDP
  - LPPGWEMARSD-GQVYFINHNTQTTTWQDP
  - LPPGWEMARSD-GQVYFINHNTQTTTWQDP
  - LPPGWEMARSD-GQVYFINHNTQTTTWQDP

- From YAP1 to wild-type of class I:
  - LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP
  - LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP
  - LPPGWEMAYTPE-GERYFINHNTKTTTWLDP
  - LPPGWEMAYTPE-GERYFINHNTKTTTWLDP
  - LPPGWEMAYTPE-GERYFINHNTKTTTWLDP
  - LPPGWEMAYTPE-GERYFINHNTKTTTWLDP
  - LPPGWEMAYTPE-GERYFINHNTKTTTWLDP
5.2 details on the TM-score

To compare the inferred structures of the sampled sequences along the path with those of the target natural sequences we used the Template Modelling (TM) score developed in [ZS04] and represents a variation of the Levitt–Gerstein (LG) score [LG98]. Compared to other similarity score (like root-mean-square deviation (RMSD)) it gives a more accurate measure since it relies more on the global similarity of the full sequence rather than the local similarities. Practically, we consider a target sequence of length $L_{\text{target}}$ and a template one whose structure has to be compared with. First, we align the two sequences and we take the $L_{\text{common}}$ pairs of residues that commonly appear aligned. Then the score is computed as

$$
\text{TM-score} = \max \left\{ d_i \right\} \left[ \frac{1}{L_{\text{target}}} \sum_{i=1}^{L_{\text{common}}} \frac{1}{1 + \left( \frac{d_i}{d_0(L_{\text{target}})} \right)^2} \right],
$$

where $d_i$ is the distance between the $i$th pair of residues between the template and the target structures after alignment, and $d_0(L_{\text{target}}) = 1.24(L_{\text{target}} - 15)^{1/3} - 1.8$ is a distance scale that normalizes distances. This formula gives a score between 0 and 1. if TM-score< 0.2, the two sequences are totally uncorrelated, while they can be considered to have the same structure if TM-score> 0.5. In Figure S4 we show the tables of the TM-scores associated with the sequences tested above.

|           | YAP1 | wt class I | YAP1 | wt class IV | YAP1 | wt class IV/III |
|-----------|------|------------|------|-------------|------|----------------|
| YAP1      | 1    | 0.88       | 1    | 0.67        | 1    | 0.71           |
| 1         | 0.95 | 0.83       | 1    | 0.94        | 1    | 0.94           |
| 2         | 0.94 | 0.86       | 2    | 0.82        | 2    | 0.91           |
| 3         | 0.91 | 0.82       | 3    | 0.74        | 3    | 0.75           |
| 4         | 0.84 | 0.98       | 4    | 0.69        | 4    | 0.74           |
| wt class I| 0.99 |           | 5    | 0.67        | 5    | 0.75           |
| YAP1      |      | 0.57       |      | 1           |      | 0.71           |

Figure S4: Table of the TM-scores measured between the structures inferred from AlphaFold.
5.3 Visualization of mean-field paths

Figure S5: Mean-field paths for the RBM model trained over WW domain dataset. Solid and dashed lines stand, respectively, for global and direct solutions. Here $\beta = 3$, $\alpha = 1$, $\lambda = 1$, $\gamma = 2$ and $T = 30$. Same edge sequences as in Figure 3 of the main text.

6 Derivation of mean-field equations for path sampling with a RBM model

To exploit the nature of the RBM as a mean-field model we rewrite the probability distribution of the model as

$$P_{RBM}(v) = \frac{1}{Z_{RBM}} \int \prod_{\mu} dh_{\mu} \exp \left( \sum_i g_i(v_i) + \sum_{i,\mu} w_{i\mu}(v_i)h_{\mu} - \sum_{\mu} U_{\mu}(h_{\mu}) \right)$$

$$= \frac{1}{Z_{RBM}} \exp \left( \sum_i g_i(v_i) + N \sum_{\mu} \Gamma_{\mu} \left( \frac{1}{N} \sum_{i} w_{i\mu}(v_i) \right) \right)$$

(9)

(10)

where $\gamma$ is defined in the main text.

By introducing the order parameters $m_{\mu}^{t} = I_{\mu,t}/N$, $q_{t} = \frac{1}{N} \sum_{i} \delta_{v_{i,t},v_{i,t+1}}$ as well as the overlap potential $\Phi$, we can write the probability for a path in the order parameter space as (in the large $N$ limit)

$$P_{\text{path}}(\{m_{t},q_{t}\}_{t=1}^{T};\beta) \propto \sum_{\{v_{t}\}} \prod_{t,\mu} \delta \left( \frac{1}{N} \sum_{i} w_{i\mu}(v_{i,t}) - m_{\mu}^{t} \right) \prod_{t} \delta \left( \frac{1}{N} \sum_{i} \delta_{v_{i,t},v_{i,t+1}} - q_{t} \right) P_{RBM}^{\beta}(v_{1}) \pi^{\beta}(v_{t},v_{t+1})$$

$$= \exp \left[ \beta N \left( \sum_{t,\mu} \Gamma(m_{\mu}^{t}) - \sum_{t} \Phi(q_{t}) \right) \right] \times$$

$$\times \int \left( \prod_{\mu,t} dm_{\mu}^{t} \prod_{t} dq_{t} \right) \sum_{\{v_{t}\}} \exp \left[ \sum_{i,t} g_{i}(v_{i,t}) + N \sum_{t,\mu} m_{\mu}^{t} \left( \frac{1}{N} \sum_{i} w_{i\mu}(v_{i,t}) - m_{\mu}^{t} \right) + N \sum_{t} q_{t} \left( \frac{1}{N} \sum_{i} \delta_{v_{i,t},v_{i,t+1}} - q_{t} \right) \right]$$

$$\approx \exp \left[ \beta N \left( \sum_{t,\mu} \Gamma(m_{\mu}^{t}) - \sum_{t} \Phi(q_{t}) \right) + NS(\{m_{t},q_{t}\}_{t=1}^{T}) \right] = \exp \left( -N \beta f_{\text{path}}(\{m_{t},q_{t}\}) \right)$$

(11)

(12)

(13)

(14)
where
\[
S(({\hat{m}_t, \hat{q}_t})_{t=1}^{T}) = \min_{\{\hat{m}_t, \hat{q}_t\}} \frac{1}{N} \sum_i \log Z_i(({\hat{m}_t, \hat{q}_t})) - \sum_{t, \mu} m_i^\mu \hat{m}_i^\mu - \sum_t q_t \hat{q}_t \tag{15}
\]
and
\[
Z_i(({\hat{m}_t, \hat{q}_t})) = \sum_{v_1, \ldots, v_T} \exp \left[ \sum_t g_i(v_t) + \sum_{t, \mu} \hat{m}_i^\mu w_{i\mu}(v_t) + \sum_t \hat{q}_t \delta_{v_t, v_{t+1}} \right]. \tag{16}
\]

Under minimization we find the result shown in the main paper. To obtain numerically the set of magnetizations and overlap that minimize the free energy, we note that the saddle point equation for \( f_{\text{path}} \) leads to the following self-consistent equation:

\[
m_i^\mu = \frac{1}{N} \sum_i \frac{1}{Z_i} \sum_{v_1, \ldots, v_T} w_{i\mu}(v_t) \exp \left[ \sum_t g_i(v_t) + \sum_{t, \mu} \hat{m}_i^\mu w_{i\mu}(v_t) + \sum_t \hat{q}_t \delta_{v_t, v_{t+1}} \right] \tag{17}
\]

\[
q_t = \frac{1}{N} \sum_i \frac{1}{Z_i} \sum_{v_1, \ldots, v_T} \delta_{v_t, v_{t+1}} \exp \left[ \sum_t g_i(v_t) + \sum_{t, \mu} \hat{m}_i^\mu w_{i\mu}(v_t) + \sum_t \hat{q}_t \delta_{v_t, v_{t+1}} \right], \tag{18}
\]

where \( \hat{q}_t = -\beta \Phi'(q_t) \) and \( \hat{m}_i^\mu = \beta \gamma_i^\mu(m_i^\mu) \). We solve this set of equations using gradient descent. To compute the LHS we first compute the partition functions \( Z_i \) using the transfer matrix method and then we take their gradient using automatic differentiation technique built in the Python library JAX [BFH+18].

To obtain the average distance from the direct space, we need to compute at each time at each site the probability of a specific state \( a = 1, \ldots, A \). This can be computed as

\[
f_{i,t}(a|{\hat{m}_t, \hat{q}_t}) = \frac{1}{Z_{\text{path}}({\hat{m}_t, \hat{q}_t})} \sum_{\{v_i\}} \delta_{v_i, a} \prod_{t, \mu} \delta \left( \frac{1}{N} \sum_i w_{i\mu}(v_{i,t}) - m_i^\mu \right) \prod_t \delta \left( \frac{1}{N} \sum_i \delta_{v_i, v_{i,t+1}} - q_t \right) \tag{19}
\]

\[
= \frac{1}{Z_{\text{path}}({\hat{m}_t, \hat{q}_t})} \int \prod_{\mu, t} \text{d} \hat{m}_i^\mu \prod_t \text{d} \hat{q}_t \exp \left( -N \sum_{\mu, t} \hat{m}_i^\mu m_i^\mu - N \sum_t \hat{q}_t q_t \delta_{v_t, v_{t+1}} \right) \sum_{\{v_i\}} \delta_{v_i, a} \exp \left( \sum_{i,t, \mu} \hat{m}_i^\mu w_{i\mu}(v_{i,t}) + \sum_t \hat{q}_t \delta_{v_t, v_{t+1}} \right) \tag{20}
\]

\[
= \frac{1}{Z_{\text{path}}({\hat{m}_t, \hat{q}_t})} \partial_{g_i,a} \exp \left[ N \min_{\{m_i^\mu\}} \left( -\sum_{\mu, t} \hat{m}_i^\mu m_i^\mu - \sum_t \hat{q}_t q_t + \frac{1}{N} \sum_t \log Z_i^{g_i,a} \right) \right] \tag{21}
\]

\[
= \partial_{g_i,a} \log \sum_{v_1, \ldots, v_{T-1}} \exp \left[ \sum_{i,t,a} g_i(a) \delta_{v_i, a} + \sum_{i,t, \mu} \hat{m}_i^\mu w_{i\mu}(v_t) + \sum_t \hat{q}_t \delta_{v_t, v_{t+1}} \right] \bigg|_{g_i=0} \tag{22}
\]

Once \( f_{i,t} \) is computed, we obtain the average distance from direct space at time \( t \) through

\[
d_{DS}(t) = \frac{1}{N} \sum_i \sum_{a \neq \{v_i^t, v_i^{t+1}\}} f_{i,t}(a). \tag{23}
\]

Finally, up to now we have note defined \( \Phi \) in this section in order to leave the computation as general as possible. Below we will define the potential as

\[
\Phi(q) = \lambda |q - (1 - \gamma/T)|^{-\alpha} T^{-\alpha-1} \tag{24}
\]

(with \( \gamma > 1 \)) to ensure a good scaling as \( T \to \infty \). We note that this is a generalisation of the potential shown in the main paper where we have fixed the exponential \( \alpha = 1 \) and \( \lambda = 1 \).
7 Analysis of the Hopfield-Potts model

7.1 The model

To deeper explore the relation between direct and global path we consider an Hopfield-Potts model with $M = 2$ patterns and $A \geq 3$ states per site (called a, b and c and so on). We will consider the thermodynamic limit $N \to \infty$, each pattern $w_{\mu}$ is constructed as follows:

$$w_1(v_i) = \delta_{v_i,a} + \omega \delta_{v_i,c}$$
$$w_2(v_i) = \delta_{v_i,b} + \omega \delta_{v_i,c}$$

(25)

The energy of the model is given by

$$E(v) = -\frac{1}{2} \sum_{\mu} \sum_{i,j} w_{i\mu}(v_i)w_{j\mu}(v_j).$$

(26)

This model is equivalent to a RBM with $M$ hidden units and quadratic local potentials:

$$E(v, h) = -\sum_{i,\mu} w_{i\mu}(v_i)h_\mu + \frac{1}{2} \sum_{\mu} h_\mu^2.$$  

(27)

Hence we can compute the optimal path between the target sequences using the mean field approach described above. Denoting $m_i^\mu(v)$ (with $\mu = 1, 2, 3$) the projection along the vectors $\delta_{v_i,a}$, $\delta_{v_i,b}$, $\omega \delta_{v_i,c}$ and $q_t$ the usual overlap defined above, we rewrite the free energy of the path as

$$f_{\text{path}}(\{m_i\}, \{q_t\}) = \sum_t ((m_t^1)^2/2 + (m_t^2)^2/2 + (m_t^3)^2 + m_t^4(m_t^1 + m_t^2)) + \sum_t (\Phi(q_t) - q_t \Phi'(q_t)) - \frac{1}{\beta} \log Z_{1D},$$

(28)

where

$$Z_{1D} = \sum_{\{v_t\}} \exp \left[ \beta \sum_t m_t^1(\delta_{v_t,a} + \omega \delta_{v_t,c}) + m_t^2(\delta_{v_t,b} + \omega \delta_{v_t,c}) + m_t^3(\delta_{v_t,a} + \delta_{v_t,b} + 2\omega \delta_{v_t,c}) - \Phi'(q_t)\delta_{v_t,v_{t+1}} \right]$$

(29)

As boundary condition, we set the target sequences to be $v_0 = \{a\}^N$ and $v_{T+1} = \{b\}^N$.

As we shall see, this model undergoes a first order phase transition at $\omega = \omega_c$ in the limit $\beta \times T \to \infty$. In this limit, when $\omega < \omega_c$, the minimum of the free energy corresponds to the direct solution from $v_0$ to $v_{T+1}$ that one obtains by restricting the sum in $Z_{1D}$ over the first two colors only. We will refer to this solution as $\#2$. When $\omega > \omega_c$, this solution is no longer a minimum of the free energy, and the latter is minimized by global paths introducing novel mutations at intermediate steps with a non zero value of $m_t^3$. Moreover, we will see how $\omega_c$ will depend on the parameters of $\Phi$ in (24).

7.2 Minimization of the path free energy in the direct subspace

To do so, we first have to find a solution of the direct problem $\#2 = \{m_t^1, m_t^2, q_t\}$. The direct solution is given by solving the following coupled equations:

$$m_t^1 = \frac{1}{Z_{1D}^{\text{dir}}} \sum_{\{v_t\}} \delta_{v_t,a} \exp \left[ \beta \sum_t m_t^1 \delta_{v_t,a} + m_t^2 \delta_{v_t,b} - \Phi'(q_t)\delta_{v_t,v_{t+1}} \right]$$

(30)

$$q_t = \frac{1}{Z_{1D}^{\text{dir}}} \sum_{\{v_t\}} \delta_{v_t,v_{t+1}} \exp \left[ \beta \sum_t m_t^1 \delta_{v_t,a} + m_t^2 \delta_{v_t,b} - \Phi'(q_t)\delta_{v_t,v_{t+1}} \right],$$

(31)

where the partition function $Z_{1D}^{\text{dir}}$ is the same as in (29) but with the sum running over the first two colors a and b only. Moreover we have $m_t^2 = 1 - m_t^1$. 

9
First we guess that the direct solution is of the form:

\[ m_1^{\text{dir}} = \begin{cases} 
1 & \text{for } t/T < \hat{x} \\
1 - \frac{t/T - \hat{x}}{1 - 2\hat{x}} + \eta(t/T) & \text{for } t/T \in (\hat{x}, 1 - \hat{x}) \\
0 & \text{for } t/T > 1 - \hat{x}
\end{cases} \]

where we impose \( q_{\tau=1/T}^{\text{dir}} = 1 + \partial_\tau m_1^{\text{dir}}/T \), while \( \eta \) is a perturbation of the order of \( 1/T^{1/(\alpha+1)} \). We first note that \( \hat{x} \) is related to the value of \( \Theta \) shown in the main paper through \( \Theta = 1 - 2\hat{x} \). We then inject this Ansatz into the equation for \( m_1 \) and try to find \( \eta \) and \( \hat{x} \) which closes the equation at the zeroth order in \( T \). Plugging this Ansatz into the definition of the partition function \( Z_{1D}^{\text{dir}} \) we notice that \( -\Phi'(q^{\text{dir}}(\tau)) = \lambda\alpha|\gamma - 1/(1-2\hat{x}) + \eta(\tau)|^{-\alpha-1} \). By rewriting \( \gamma - 1/(1-2\hat{x}) + \eta(\tau) = \xi(\tau)/T^{\alpha+1} \), we have \( -\Phi'(q^{\text{dir}}(\tau)) \sim T \). The linear term in \( T \) in the coupling interactions of our 1D model forces the partition function to be dominated by the configurations \( v_t = a \) for \( t < \hat{x}T \) and \( v_t = b \) for \( t > \hat{x}T \). Hence the partition function can be rewritten as follows:

\[ Z_{1D}^{\text{dir}} = T \int_0^1 d\tau \exp \left[ \beta T \left( \int_0^\tau dy m_1^{\text{dir}} + \int_\tau^1 dy (1-m_1^{\text{dir}}) - \frac{\lambda\alpha}{|\xi(\tau)|^{\alpha+1}} \right) \right]. \]

We can neglect the first order correction in \( T \) and maximize the argument in the exponential to obtain the leading term of the partition function. This amounts to solving the following differential equation in \( \tau \in (\hat{x}, 1 - \hat{x}) \):

\[ -2 \frac{\tau - \hat{x}}{1 - 2\hat{x}} + \frac{1 + \lambda\alpha(\alpha+1)}{\xi(\tau)^{\alpha+2}} = 0. \]

Solving this differential equation leads to

\[ \xi(\tau) = \left( \frac{1}{\xi(\hat{x})^{\alpha+1}} - \frac{T^2 - \hat{x}^2 - (\tau - \hat{x})}{\lambda\alpha(1-2\hat{x})} \right)^{\frac{1}{\alpha+1}}. \]

In order to ensure the continuity of \( \Phi'(q_\tau) \) in \( \tau = \hat{x} \), we fix \( \xi(\hat{x}) = \gamma T^{1/(\alpha+1)} \). By definition of \( \xi \) we have \( \xi/T^{1/(\alpha+1)} = \gamma - 1/(1-2\hat{x}) + \eta' \). Rewriting \( \hat{x} = 1/2 - 1/(2\gamma) + \xi/T^{1/(\alpha+1)} + o(T^{1/(\alpha+1)}) \) we can fix \( \eta \) at the first order as

\[ \eta(\tau) = \frac{1}{T^{1/(\alpha+1)}} \left( \frac{1}{\gamma} \int_{\hat{x}}^\tau \xi(y)dy + 2\zeta \gamma^2 (\tau - \hat{x}) \right) + o \left( \frac{1}{T^{1/(\alpha+1)}} \right). \]

By imposing the boundary condition \( \eta(\hat{x}) = \eta(1 - \hat{x}) = 0 \) we can also fix \( \zeta \) at first order. It is easy to check that (30),(31) are fulfilled at zeroth order by this solution.

### 7.3 The direct-to-global phase transition

We now write the first derivative of the free energy along the third magnetization \( m_3 \):

\[ \frac{\partial f_{\text{path}}}{\partial m_3} \bigg|_{#_2} = 1 - \langle \delta_{v_t,a} + \delta_{v_t,b} + 2\omega\delta_{v_t,c} \rangle_{1D} \bigg|_{#_2} \]

By studying these derivatives we will show the existence of a critical \( \omega_c \) discriminating a regime where all the derivatives vanish (\( \omega < \omega_c \)) and a regime with negative derivatives (\( \omega > \omega_c \)).

As above, the average in the RHS of (37) is dominated, for \( T \to \infty \), by the ground state path. Two classes of competing configurations must be considered: the usual direct configurations that start in \( v_0 = a \) and turn into \( b \) at some point \( t \) such that \( t/T \in (\hat{x}, 1 - \hat{x}) \); another one starting in \( a \) then changing to \( c \) at some point \( t = xT \) (\( x \in (0,1/2) \)) and then turning into \( b \) when \( t = (1-x)T \). The energy of the first set of configurations (for \( T \gg 1 \)) is given by:

\[ E_1 = -T \left( \hat{x} + \frac{1}{2} \right) + \frac{\lambda\alpha}{\gamma^{\alpha+1}}, \]

while the second configuration has energy

\[ E_2(x) = \begin{cases} 
-T(2x + \omega(1 - 2x)) + \frac{2\lambda\alpha}{\xi(\hat{x})^{\alpha+1}} & \text{for } x \leq \hat{x} \\
-T \left( 2\hat{x} + 2 \int_\hat{x}^x dy (1 - \frac{y-\hat{x}}{1-2\hat{x}}) + \omega(1 - 2x) - \frac{2\lambda\alpha}{\xi(x)^{\alpha+1}} \right) & \text{for } x \in (\hat{x}, 1/2)
\end{cases} \]
Figure S6: Average value of the free energy derivative along $m_3^t$. Here $\beta = 3$, $\gamma = 2$, $\alpha = 1$ and $\lambda = 3$.

which is minimized for $x = \hat{x}$ when $\omega \in (1/4, 1)$ and at $x = 0$ when $\omega > 1$. Here the condition $E_2 < E_1$ leads to a phase transition at the critical value $\omega_c = 1/2 + \lambda \alpha/(T \gamma^{\alpha+1}(1 - 2\hat{x}))$.

The last argument holds until $\hat{x} > 0$. For finite $T$ we could be in a regime where $\hat{x} = 0$. This leads to a different solution for the phase transition. To study this regime we consider $\lambda = \hat{\lambda} T^{(\alpha+2)/(\alpha+1)}$. Injecting it into the expression for the partition function $Z_{TD}^{\text{dir}}$ and expanding the force in terms of $\eta' = \hat{\eta}'/T^{1/(\alpha+1)}$, we have

$$Z_{TD}^{\text{dir}} = T \int_0^1 d\tau \exp \left[ \beta T \left( \int_0^\tau dy m_1^{y, \text{dir}} + \int_\tau^1 dy (1 - m_1^{y, \text{dir}}) - T^{1/(\alpha+1)} \frac{\hat{\lambda} \alpha}{|\gamma - 1|^{\alpha+1}} + \frac{\hat{\lambda} \alpha (\alpha + 1) \hat{\eta}'(\tau)}{|\gamma - 1|^{\alpha+2}} \right) \right].$$

Maximization leads to the following solution for $\eta$:

$$\eta(\tau) = \frac{|\gamma - 1|^{\alpha+2}}{T^{1/(\alpha+1)} \lambda \alpha (\alpha + 1)} \left[ \frac{\tau^3}{3} - \frac{\tau^2}{2} + \frac{\tau}{6} \right],$$

where we have imposed the boundary condition $\eta(0) = \eta(1) = 0$. The condition $-\Phi'(q) > \frac{\lambda \alpha}{T^{1/(\alpha+1)}}$ leads to the condition

$$\lambda \alpha > \frac{T}{6} \left( \frac{1}{|\gamma - 1|^{\alpha+1}} - \frac{1}{\gamma^{\alpha+1}} \right)^{-1},$$

which is valid in the case $\lambda = \hat{\lambda} T^{(\alpha+2)/(\alpha+1)}$ and $T \to \infty$. The derivative of the free energy in eq. (37) can be done as above by considering the two classes of relevant configurations. The energy of the first class is

$$E_1 = -\frac{2T}{3} + \frac{\lambda \alpha}{|\gamma - 1|^{\alpha+1}}$$

while the second class corresponds to energy

$$E_2 = -T \left( \omega + \frac{1}{3} \right) + \frac{\lambda \alpha}{|\gamma - 1|^{\alpha+1}}.$$

The condition $E_2 < E_1$ leads to a new critical value $\omega_c = 1/3 + \lambda \alpha/(T |\gamma - 1|^{\alpha+1})$. Merging together the two regimes studied above, we find that the critical line at the first order in $1/T$ is given by:

$$\omega_c = \max \left( \frac{1}{2} + \frac{\lambda \alpha}{T^{\gamma^{\alpha+1} \Theta} \frac{1}{3} + \frac{\lambda \alpha}{T |\gamma - 1|^{\alpha+1}} \right),$$

where $\Theta = 1 - 2\hat{x}$. In Figure S6 we plot the behavior of the derivative of the free energy for different value of $\omega$, showing the instability at the corresponding critical $\omega_c$. 

Figure S6: Average value of the free energy derivative along $m_3^t$. Here $\beta = 3$, $\gamma = 2$, $\alpha = 1$ and $\lambda = 3$. 

Figure S7: Average log-likelihood and distances to direct space (inset) of intermediate sequences as a function of $\beta \times T$. Symbols stands for different $T$ (circles for $T=20$, diamonds for $T=30$ and pluses for $T=40$). Green symbols represents direct solutions (which are of course independent of $\omega$), Red symbols represents global solutions with $\omega = 0.4$ and maroon symbols represent global solutions for $\omega = 0.7$. The other parameters are set to $\alpha = 1$, $\gamma = 2$, $\lambda = 1$.

The crossover is visible in Figure S7, where can observe the coincidence of the average log-likelihoods of intermediate sequences along direct and global paths at large $T$ for small $\omega$, and the higher quality of global paths for large $\omega$. Notice that these results are valid when $T$ is sent to large values while keeping $\beta$ fixed. If $\beta$ is small, e.g. of the order of $\frac{1}{T}$, the domination of global paths on direct paths is due to the larger entropy of the former. Figure S7 shows that, for small $\beta \times T$, global paths are indeed of lesser quality (probability) than their direct counterparts, even at high $\omega$.

References

[BFH+18] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. Available online at: http://github.com/google/jax.

[ES99] Xavier Espanel and Marius Sudol. A single point mutation in a group i ww domain shifts its specificity to that of group ii ww domains. Journal of Biological Chemistry, 274(24):17284–17289, 1999.

[FI12] Asja Fischer and Christian Igel. An Introduction to Restricted Boltzmann Machines. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pages 14–36. Springer, Berlin, Germany, September 2012.

[JGS+16] Hugo Jacquin, Amy Gilson, Eugene Shakhnovich, Simona Cocco, and Rémi Monasson. Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. PLOS Computational Biology, 12(5):1–18, 05 2016.

[LD89] Kit Fun Lau and Ken A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules, 22(10):3986–3997, October 1989.
[LG98] Michael Levitt and Mark Gerstein. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of sciences*, 95(11):5913–5920, 1998.

[MJ96] Sanzo Miyazawa and Robert L. Jernigan. Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.*, 256(3):623–644, March 1996.

[TCM19] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. Learning protein constitutive motifs from sequence data. *eLife*, 8:e39397, March 2019.

[Tie08] Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML ’08: Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. Association for Computing Machinery, New York, NY, USA, July 2008.

[Tub18a] Jérôme Tubiana. *Restricted Boltzmann machines : from compositional representations to protein sequence analysis*. PhD thesis, Université Paris sciences et lettres, Paris, France, November 2018.

[Tub18b] Jerome Tubiana. Probabilistic graphical models (pgm), 2018. Available online at: https://github.com/jertubiana/PGM.

[ZS04] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.
Appendix A: Weights Logo for the WW domain

link to the RBM trained on WW domain data.
9 Appendix B: Weights Logo for the Lattice Proteins

link to the RBM trained on LP domain data.