From residue co-evolution to protein conformational ensembles and functional dynamics

Ludovico Sutto [*], **Simone Marsili** [†], **Alfonso Valencia** [†] **and Francesco L. Gervasio** [* ‡]

[*]Institute of Structural and Molecular Biology,University College London, 20 Gordon Street - London - WC1H 0AJ, UK,[†]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, Melchor Fernandez Almagro 3, 28029 Madrid, Spain, and [‡]Department of Chemistry, University College London, 20 Gordon Street - London - WC1H 0AJ, UK

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The analysis of evolutionary amino acid correlations has recently attracted a surge of renewed interest, also due to their successful use in de-novo protein native structure prediction. However, many aspects of protein function, such as substrate binding and product release in enzymatic activity, can be fully understood only in terms of an equilibrium ensemble of alternative structures, rather than a single static structure. In this paper we combine co-evolutionary data and molecular dynamics simulations to study protein conformational heterogeneity. To that end, we adapt the Boltzmann-learning algorithm to the analysis of homologous protein sequences and develop a coarse-grained protein model specifically tailored to convert the resulting contact predictions to a protein structural ensemble. By means of exhaustive sampling simulations, we analyze the set of conformations that are consistent with the observed residue correlations for a set of representative protein domains, showing that: i) the most representative structure is consistent with the experimental fold and ii) the various regions of the sequence display different stability, related to multiple biologically relevant conformations and to the cooperativity of the co-evolving pairs. Moreover, we show that the proposed protocol is able to reproduce the essential features of a protein folding mechanism as well as to account for regions involved in conformational transitions through the correct sampling of the involved conformers.

co-evolution network inference coarse-grained protein folding

Significance Statement

Evolutionary-related protein sequences have been selected to preserve a common function and fold. Residues in contact in this conserved structure are coupled by evolution and show correlated mutational patterns. The exponential growth of sequenced genomes make it possible to detect these co-evolutionary coupled pairs and to infer three-dimensional folds from predicted contacts. But how far can we push the prediction of native folds? Can we predict the conformational heterogeneity of a protein directly from sequences? We address these questions developing an accurate contact prediction algorithm and a protein coarse-grained model, and exploring conformational landscapes congruent with co-evolution. We find that both structural and dynamical properties can be already recovered using evolutionary information only.

Pairs of positions along a protein sequence can show strong correlations arising both from functional and structural constraints [1, 2, 3, 4, 5, 6, 7, 8, 9]. Earliest approaches for detecting interdependent residues and predicting three-dimensional contacts in proteins[1, 2, 3, 4, 8] analyzed alignments containing from tens to a few hundreds sequences. Given the small size of available sequences datasets, these works relied on an independent pair approximation: a co-evolutionary coupling between two residues was estimated independently for each pair, ignoring the rest of the network of residues. The number of known protein sequences, however, has grown dramatically in the last few years[10]. Such a large increase in the size of datasets has allowed to fit (either explicitly[11] or implicitly[12, 13]) pairwise models for protein sequences that take into account the whole network of correlated residues simultaneously, and are able to disentangle correlated positions from interacting positions by identifying the parameters of the model with the coupling constants in an Ising-like Hamiltonian[14, 15]. Despite their simplicity, these models have had remarkable success in the design of synthetic sequences preserving natural function[12, 13] and in the prediction of interacting pairs of residues from the knowledge of their sequence alone[16, 17, 18, 19, 20, 21].

In this paper, we tackle the problem of sampling an ensemble of structures compatible with the observed co-evolution between protein residues. We will follow a two-step procedure. The first step corresponds to an inverse problem: from a set of homologous sequences to the parameters of a model. Inverse problems are notoriously computationally hard. For large sets of variables, an exact evaluation of the normalizing constant of the variables joint distribution (the partition function, in the language of statistical mechanics) is impracticable. Previous works in the literature focused on efficiency, circumventing this problem by adopting different, approximated solutions[16, 17, 22, 18, 19, 23, 24, 25, 26], generically based on tractable approximations of the likelihood. However, given the success and the number of potential applications of co-evolutionary analysis, the study of reference and more quantitative approaches is necessary. In this regard, the Monte Carlo Markov Chain (MCMC)-based, maximum likelihood approach, albeit computationally demanding, is in principle exact given a sufficient sampling at each minimization step. In this work we adopted the Boltzmann learning algorithm[27, 11], whose accuracy in inferring the parameters of the pairwise model, at variance with all the previous approaches in the literature, is not biased *a priori* by the choice of a particular approximation scheme.

The second step is a direct problem: after translating the probabilistic model for sequences into an energy potential for protein structures, we can explore the resulting energy landscape using molecular dynamics. After extensive sampling, we can characterize the folding reaction and find the best candidate for the native fold as well as meta-stable intermediates and conformers that may have a functional role. Moreover, we can spot flexible regions, directly connecting co-evolution to function and dynamics. With this goal in mind,

Reserved for Publication Footnotes

we introduce a coarse-grained model particularly apt to translate predictions of contacts to a structural ensemble. Thanks to the great reduction in the number of degrees of freedom, coarse-grained models have been widely used to study many aspects of proteins[28, 29, 30, 31, 32, 33, 34]. Due to their simplicity, C_α models in particular have already been used to predict protein folds from co-evolutionary data[35, 36, 37]. Here, in the same spirit as the model presented in [35] where co-evolutionary information is used with a C_α coarse-grained protein model, we present a higher resolution coarse-grained model that combines the pairwise predictions with an adapted all-atom force-field for the heavy backbone atoms, similarly to the approach used in [38]. The predicted contacts are introduced as favorable interactions between C_β atoms of a coarse-grained side-chain, while the protein backbone is modeled with all the heavy atoms in order to capture the secondary structure conformation with high resolution. Indeed, we show through extensive molecular dynamics simulations on a set of 18 proteins, that the final accuracy of structure prediction, measured as RMSD from the native experimental structure, is determined solely by the accuracy of contact predictions.

However, besides recovering a protein native fold, the main advantage of the proposed approach is its ability to capture the conformational heterogeneity and the thermodynamical features of the folding reaction as implied by co-evolutionary information only. In contrast to more expensive approaches like all-atom MD or more refined coarse-grained potentials[39], we can afford an extensive equilibrium sampling of the conformational space. We illustrate this point by applying our approach to analyze two energy landscapes, related to the folding of the Ras protein and the conformational dynamics of a tyrosine kinase. As expected Ras folds cooperatively and we find and characterize a folding intermediate. The protein kinase correctly samples an ensemble of active-like and inactive-like structures that are biologically relevant for its function and shows a flexibility pattern compatible with experimental observations.

Results

Contact predictions via Boltzmann learning. Entropy maximization provides a simple procedure for building a probabilistic model that is consistent with a set of available measures. If we know the average values F_i of a set of variables x_i, the maximum entropy distribution is given by $P \propto \exp(\sum_i \lambda_i x_i)$, with a Lagrange multiplier λ_i for each variable x_i[40]. Fixing the frequencies for single and pairs of amino acids in a multiple sequence alignment (MSA), P takes the form[11]: $P(a) = Z^{-1} \exp\left[\sum_i h_i(a_i) + \sum_{i<j} J_{ij}(a_i\ a_j)\right]$ over protein sequences. The parameters $h_i(\alpha)$ and $J_{ij}(\alpha\ \beta)$ are the Lagrange multipliers that fix the averages of the model, $f_i(\alpha)$ and $f_{ij}(\alpha\ \beta)$, to the empirical frequencies $F_i(\alpha)$ and $F_{ij}(\alpha\ \beta)$ computed from the MSA, where α and β denote two particular amino acids and i, j two particular residues along the protein sequence. Due to the Boltzmann-like form of the previous equation, the parameter $J_{ij}(\alpha\ \beta)$ can be interpreted as the direct interaction between amino acids α and β at positions i and j, after the contributions from the interaction with other positions through indirect pathways have been disentangled[11, 16, 22].

To our knowledge, since the early work of Lapedes[11], the numerical route of likelihood maximization via importance sampling, or Boltzmann learning[27], has not been explored, probably due to its computational complexity. As outlined in the Introduction, this approach (see for example Roudi *et al.*[41] for an extensive comparison between Boltzmann learning and various approximated schemes), has the advantage of having unbounded precision in retrieving the parameters of a maximum entropy model. We tested Boltzmann learning on an assorted set of 18 proteins with varying length, from 63 to 216 residues, and different secondary structure composition (see Table 1). For each alignment, we inferred a pairwise model by maximizing a regularized version of the log-likelihood of the sequences with respect to parameters h and J (see the Materials and Methods section for details). In all the 18 cases, we were able to reproduce the empirical frequencies F within a reasonable error, as we checked through extensive sampling from the final models distribution(Fig. S1, SI Appendix). Depending on the size of the protein sequence, we obtained mean absolute relative errors $F_{ij} - f_{ij}$ F_{ij} between the model pair frequencies f_{ij} and the empirical F_{ij} ranging from 1% to 3% (0.2% to 2% for $F_{ij} > 0\ 01$) for the different protein families. Being our approach based on importance sampling, the presence of many isolated modes in the distribution of sequences could lead to poor mixing of the Markov chain and, consequently, to a large error in the estimate of the gradient of the likelihood function. Indeed, clusters of sequences in multiple sequence alignments are common and are known to reflect potential functional subfamilies among the members of a single protein family[7]. As a cross-check, we verified that the fitted models capture the organization of the original alignment in clusters of sequences (see Fig. 1 and details in SI Appendix). We point out that external sources of variation - such as changes in functional requirements within the same protein family - should be explicitly taken into account in future, improved, models in order to discriminate between intrinsic and extrinsic correlations in the sequence alignment.

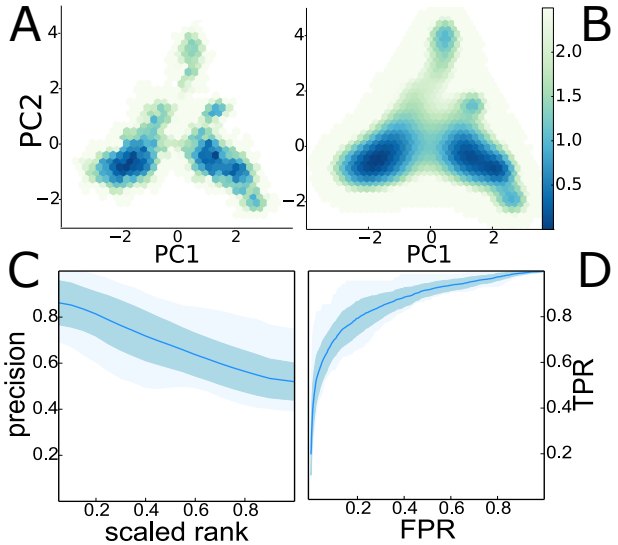

Figure 1: **Panels A, B***: energy surfaces obtained projecting sequences from the ADH_zinc_N domain family (panel A) and sequences simulated from the model (panel B) over the first two principal components of the MSA[7], and taking the negative logarithm of the resulting distribution. High probability regions in sequence space are in dark blue. The cluster structure of the alignment is clearly reproduced by the simulated trajectories.* **Panel C, D***: the mean precision of the top ranked predictions for different values of the scaled rank (rank/total number of contacts), and the mean true positive rate for different values of false positive rate (*ROC *curve). The color bands show the standard deviation and the interval between the minimum and maximum values.*

For each protein, from the estimated set of $\{J\}$ we computed a matrix of co-evolutionary couplings using a protocol first proposed by Ekeberg *et al*[25]. Assuming that a strong, direct co-evolution is an evidence of physical contact between two residues, we finally ranked the pairs of residues using the value of co-evolutionary coupling as a score. The accurate determination of couplings through the Boltzmann learning algorithm resulted in high quality predictions. Figure 1 summarizes the performance in terms of mean precision against the top scoring predictions rank, and as mean true positive rate as a function of false positive rate. We defined a pair of residues to be in contact when the distance between their C_β atoms (C_α in case of GLY) is smaller than 8 Å[42, 43], and included in the analysis all the pairs with a sequence separation larger than four. For the top N predictions, where N denotes the number of amino acids in a protein structure, the algorithm obtained a mean precision of 0.67 ± 0.03, with a maximum of 0.89 for the cNMP_binding domain (PDB:3FHI). A comparison with predictions obtained through a more standard mean-field solution[18, 19] is included in the SI Appendix.

Table 1: For each of the 18 protein domains analyzed in this work, the table shows: i) the number of effective sequences in the corresponding family MSA (Meff)*, ii) the fold class according to CATH classification[44]* (Class)*, iii) the PDB code for the representative structure of the family* (PDB)*, iv) the precision for the top N predictions, where N denotes the number of amino acids in the corresponding protein structure* (prec.)*, v) the distance RMSD* (dRMSD) *to the native conformations of the best (on the left) and minimum energy (on the right) sampled structure. The dRMSD are calculated from the coordinates of the C_α atoms corresponding to positions with less than 5% gaps in the (reweighted) MSA. The units are Angstroms (Å). vi) the number of amino-acids in the structure* (N) *and the number of positions with less than 5% gaps* (Ng, *as a subscript*) *that were included in the dRMSD calculation. Each domain is marked by its Pfam identifier* (Pfam_ID)*; the domains are ordered by size.*

Pfam_ID	Meff	Class	PDB	prec.	dRMSD	$N_{/Ng}$
Thioredoxin	4388	α/β	1RQM	0.75	1.9/2.2	63_{62}
HTH_31	2901	α	3F52	0.69	1.3/1.9	64_{49}
Sigma70_r2	8008	α	1OR7	0.74	1.0/1.9	68_{53}
RRM_1	7076	α/β	1G2E	0.80	1.2/2.0	71_{50}
Trans_Reg_C	6458	α	1ODD	0.55	2.1/3.1	76_{65}
cNMP_binding	7539	α/β	3FHI	0.89	1.6/2.1	81_{72}
CMD	1488	α	3D7I	0.39	2.4/3.8	85_{61}
HxlR	1674	α	3DF8	0.48	2.2/2.6	87_{77}
fn3	8862	β	1BQU	0.58	2.1/2.9	88_{57}
Cadherin	6219	β	2O72	0.69	2.5/3.1	90_{66}
OmpA	4081	α/β	1OAP	0.63	1.8/2.4	96_{78}
Response_reg	36372	α/β	1KGS	0.70	2.5/3.1	111_{99}
PAS	3350	α/β	2GJ3	0.58	3.6/7.3	112_{80}
Peptidase_M23	2975	β	3NYY	0.71	2.7/3.5	112_{82}
TrkA_N	2630	α/β	3FWZ	0.73	2.1/3.0	116_{100}
ADH_zinc_N	5932	α/β	1A71	0.66	2.9/3.6	119_{99}
Ras	2528	α/β	5P21	0.73	2.7/3.3	160_{144}
Trypsin	4703	β	3TGI	0.78	2.8/4.0	216_{167}

A coarse-grained model for structure prediction and conformational sampling. To take full advantage of the accurate contact predictions resulting from the Boltzmann learning algorithm, we propose a coarse-grained model that combines an all-heavy-atom description of the protein backbone with a C_β description of the side-chains (see Methods). Similar coarse-grained approaches have already been successfully applied to study the thermodynamics of model proteins [30]. Since we expect the residues to be evolutionary coupled through their side chains, we set the predicted interactions to act between the C_β atoms, that is the first atom to branch out of the main chain. More precisely, in our model we used the N best predicted contacts, where N is the number of residues (see Methods). Moreover, the inclusion of all the heavy atoms of the main chain permits to use a transferable potential that acts on the actual degrees of freedom of the backbone and allows to correctly reproduce the experimental population of Ramachandran backbone angles. To complement long-range predictions, we estimated the helical propensity of each residue solely from the predicted contacts and translated it in a stabilizing hydrogen bond-mimicking interaction between residues $i, i+4$ (see SI Appendix for details). For all of the α and α/β proteins, all the helices are correctly predicted with the exception of h1 for 3D7I, h5 for 2GJ3 and h2 for 5P21 structures (Fig. S2, SI Appendix).

Using this simplified model, we investigated the conformational space accessible to the protein domains in Table 1. The 9 best predicted structures obtained in the folding runs are shown in Fig. 2 superimposed to their respective native folds. Those structures correspond to the conformations with minimum distance root mean square deviation of the α carbon atoms (C_α-dRMSD, see Methods) to the native conformation. For all of them not only is the global fold correctly predicted, but also most of the secondary structure elements are present and correctly packed.

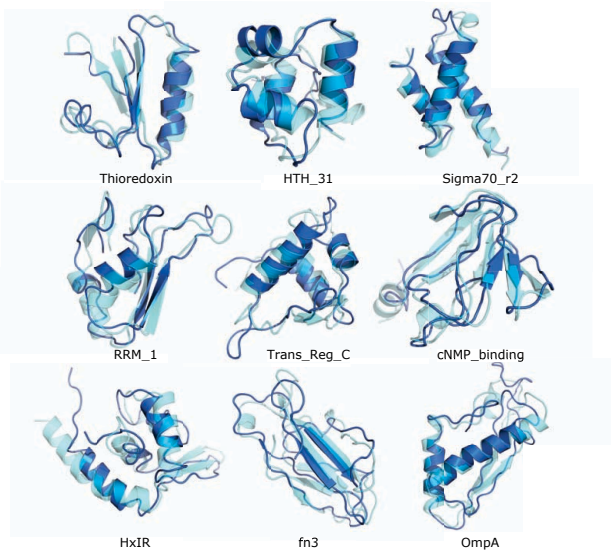

Figure 2: The 9 best predicted protein structures out of the 18 simulated are shown in blue, overlaid to the native conformation in cyan (transparent). The proteins are disposed with increasing size from left to right and from top to bottom. Their Pfam namecode is shown below.

For domains of similar size, we found a clear correlation between the precision of the contact predictions and the final dRMSD values (for example, domains Trans_Reg_C, CMD and PAS have lower precision in contact prediction and higher dRMSD values than other domains with a similar number of amino acids, see Table 1), confirming that the precision of contacts prediction is the main determinant of the quality of the final fold reconstruction. On the other hand, all the 18 proteins fold within 3 Å to the native protein, except for the

2GJ3 structure (PAS domain) that performs worse (see Table 1 and Fig. S6, SI Appendix). We found that the model is tolerant to both a conspicuous presence of non-native contacts among the set of predicted contacts as well as to the absence of a large proportion of native contacts among the predicted ones.

To have a reference baseline to compare with, we also simulated the proteins using the same conditions but replacing the set of predicted contacts with the set of native contacts (see details in the SI Appendix). The best dRMSDs obtained in this case are always below ∼2 Å (see SI Appendix, Fig. S6, black curve). These values quantify the theoretical limit of the present model with perfectly predicted contacts. The quality of the predicted structures is being as good as those predicted by the structure-based reference simulation in several cases (HTH_31, Sigma70_r2, RRM_1, Ras). Interestingly, the dRMSD of the minimum energy structure (dashed curve in Fig. S6, SI Appendix) does not deviate much from the absolute minimum dRMSD structure sampled in the trajectory. Indeed, for 17 out of 18 proteins the minimum energy structure is below 4 Å to the native structure. This indicates that the model and the contact predictions are sound and lead to a funnel-shaped energy landscape whose minimum corresponds to the crystallographic structure.

In the single case of the PAS domain, we observe a larger deviation, 7 Å vs 3.6 Å for the dRMSD of the minimum energy structure and the absolute minimum respectively (see Table 1), that can be only partially explained by the presence of an associated cofactor in the experimental structure. Such a large difference indicates that while the minimum dRMSD structure still belongs to the native basin, it does not coincide with the minimum energy structure which is the defining property of the native fold. Inspecting the set of predicted contacts, we found that 5 of them are in the dimeric interface of the corresponding homo-dimer (see Fig. S7, SI Appendix). After removing these inter-domain contacts, the minimum energy dRMSD of the monomeric structure decreases to 4.7 Å (see Fig. S6, SI Appendix), showing that these few contacts were responsible for a large displacement from the native conformation. This result is also corroborated by the fact that when 5 randomly picked false positive contacts are removed, the dRMSD of the minimum energy structure does not improve, as we verified running 20 independent simulations (see SI Appendix). As far as we know, the effect on fold reconstruction of strongly coupled pairs of residues at homo-dimeric (or homo-oligomeric) interfaces, when incorrectly classified as contacts in the monomeric structure, have not been described before. Here we show that this effect can be large, depending on the relative position of the pairs of residues at the dimer interfaces. Reasonably, similar but weaker effects are present for other cases, being 10 over 18 proteins in our set homo-dimeric in the corresponding PDB structure. On the other hand, the analysis demonstrates that the ability of PAS domains to form homo-dimers[45] is subjected to evolutionary pressure, and identifies a set of 5 pairs of positions involving the N-terminal helix (I40-F27; Y49-F27; A117-P35; L130-Q29; M132-A34 in the 2GJ3 PDB numbering) that emerge as crucial for the stabilization of the PAS homodimer across the protein family, as supported by their proximity in the crystal structure, strong co-evolutionary coupling and simulation.

Conformational heterogeneity and residue co-evolution

To investigate the ability of co-evolutionary couplings to also encode for dynamical and functional information, we analyzed the conformational space close to the native fold in the case of the catalytic domain of SRC tyrosine kinase and characterized the full folding reaction of the Ras domain from a thermodynamical point of view.

Protein kinases are known to undergo a large conformational rearrangement of a centrally located loop during activation, called activation loop or "A-loop". In the case of the catalytic domain (CD) of the SRC tyrosine kinase, the A-loop spans more than 20 residues and the structural rearrangement moves the backbone across ≈25 Å from the inactive conformation where the A-loop is folded to the fully active structure where the A-loop is in an extended conformation. The A-loop is known to be flexible to the point of being invisible in many crystal structures. It is thus interesting to see if traces of this conformational transition can be observed by an exhaustive sampling of the native fold basin with our model. We performed a 500 ns-long parallel tempering (PT) simulation of the 250 residues CD starting from the inactive conformation with the A-loop in the so-called half closed conformation, corresponding to the structure with PDB code 2SRC. We used the 250 best predicted contacts calculated over 3812 effective sequences from PFAM Pkinase_Tyr family, without adding any structure-based bias. Indeed, only 130 out of the 250 contacts are native contacts (where a native contact is defined between CB atoms within 8 Å in the native structure). Consistent with the fact that no contacts have been predicted for the A-loop, we observe an extensive sampling of the conformational space available to the loop while the rest of the protein correctly maintains the native fold.

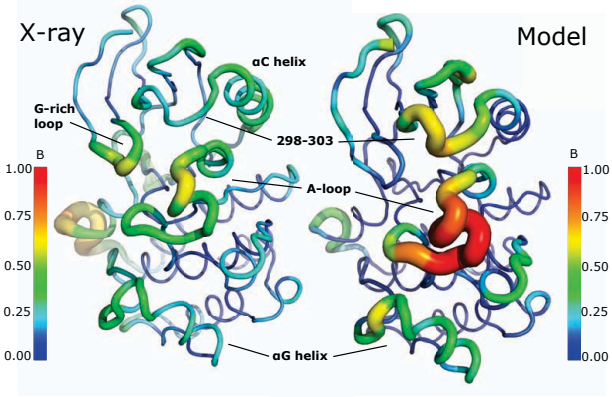

Figure 3: Comparison of the fluctuations of the SRC catalytic domain. The experimental X-ray structure (PDB: 2SRC) is shown on the left with the normalized B-factor color and thickness coded. The same structure with the B-factor calculated from the simulation of the SRC catalytic domain is shown on the right. For easier comparison, the scales have been normalized. Higher values correspond to larger fluctuations.

In Fig. 3 we compare the experimental and simulated structural flexibility of different positions along the chain. Even though we expect a sequence-dependent effect on structural order/disorder, we note that flexible regions (A-loop, 298-303 loop and αG-helix) are captured by our model with few exceptions (αC-helix, G-rich loop), suggesting that chain flexibility itself is partially encoded in the sequences of the kinase protein family (see also Fig. S8, SI Appendix). Indeed these regions are known to be involved in the activation or in protein-protein interaction [46]. We note that flexibility cannot be trivially deducted from the number of predicted contacts involving each residue, as shown in Fig. S8, but is

rather a consequence of the whole fold and network of interactions. Moreover, both the active and inactive conformations of the A-loop are repeatedly sampled within 5 Å C_α-RMSD to the crystal structure (see Fig. S8). The ability to reach both endpoints of the complex conformational transition in the catalytic domain is very encouraging.

The Ras protein is an α/β globular protein and is a crucial mediator of cellular proliferation and differentiation involved in several signaling pathways[47, 48]. We performed a 180 ns long PT simulation to explore a wide range of temperatures and to have a solid sampling. The simulated folding transition is highly cooperative with a clear peak in the heat capacity at constant volume at the folding temperature T_f (see SI Appendix, Fig. S9). In Fig. 4, we show the free energy as a function of dRMSD at three different temperatures: $T_{low} < T_f$, $T \approx T_f$, $T_{high} > T_f$. In the unfolded state (U) the beta strands β_2 and β_3 are unfolded and lead to an extended tail departing from a rather structured core around the partially formed α_3 and α_4 helices. The collapse of this unfolded tail and its correct positioning lead to an intermediate, partially folded state (I, dRMSD=6.5 Å) where helices α_4 and α_3 are formed. Eventually, the native basin (N, dRMSD=3.8 Å) is reached with the formation of helices α_3 and α_5 and their correct packing by crossing a free energy barrier of 4 kJ/mol.

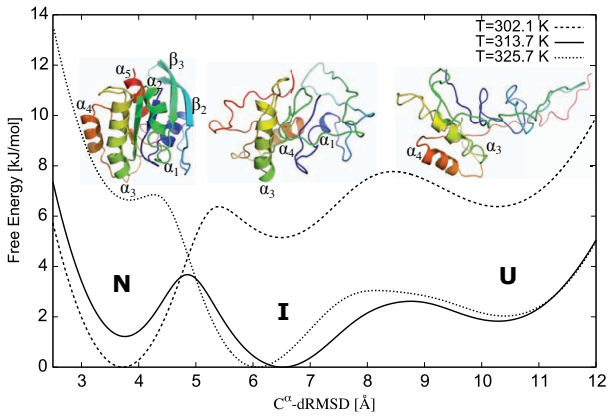

Figure 4: Folding free energy of Ras protein at three different temperatures around the folding temperature $T_f = 317K$ as a function of the C_α-dRMSD. Representative structures of the native (N), intermediate (I) and unfolded (U) states are shown above with the secondary structure elements labeled.

These features are not simply encoded in the native fold, for a coarse-grained C_α structure-based model[49] is unable to capture them (see SI Appendix). It is interesting to note that the existence of a folding intermediate has been reported at high pressure and in denaturing conditions[50]. Albeit we cannot directly compare our result with the experimental structures and the different experimental conditions, it is encouraging to observe the emergence of typical folding features, such as intermediate folding states, from a sequence-derived coarse-grained model.

Discussion and outlook

Among proteins sharing a common ancestor, secondary and tertiary structure is generally much more conserved than protein sequence[51]. As a result, a protein family is free to explore a large space of possible sequences, while at the same time preserving a common structural framework. As shown by previous works[3, 4, 8, 11, 22, 16, 17, 18, 19, 20, 21, 24, 23, 25, 26], the information contained in sequence variability can be translated to a series of couplings between pairs of protein residues, restraining the set of conformations compatible with the observed sequence alignments. The main objective of this paper is to study such a evolutionary-restrained space of conformations. We extracted the optimal couplings by forgoing unnecessary approximations and using a standard reference algorithm, Boltzmann learning[52], based on MCMC simulations of trajectories in sequence space. This scheme allowed us to verify, at variance with previous works, the simple but important fact that the fitted maximum-entropy models reproduce the observed correlations between residues in multiple sequence alignments. We developed a simplified physical model for protein dynamics, showing that the combination of accurate contact predictions and of a coarse-grained, yet biologically meaningful, protein model allows for a full sampling of the structural ensemble associated to a protein family. Our simulations support the finding[19] that the structural landscape dictated by residue co-evolution is dominated by the energy minimum corresponding to the native fold, that we recover with high accuracy for a set of unrelated protein domains. We show that the presence of conserved quaternary structure in the family, as a biologically relevant homo-dimer or homo-oligomer, can lead to misclassification of co-evolving pairs of residues as contacts in the monomeric structure, compromising the quality of the folding reconstruction.

Furthermore, the main end of our protocol is to explore the connection between co-evolutionary couplings at the residues level and protein dynamics. An issue that has been mentioned in the literature[18, 53, 36] but not directly addressed. We show in two significant cases that co-evolutionary information can be used to reproduce and predict increasingly complex protein features: from identifying flexible regions as in the case of SRC, to the sampling of conformers crucial for the kinase activation and function up to a full characterization of the folding reaction as in the case of the RAS protein. It is worth noting that in all these cases, the recovered information is not easily accessible by other approaches such as elastic-network models or without explicit knowledge of the structures involved.

The approach we propose paves the way to further development in combining experimental and genomic data, going beyond the rigid structure paradigm and taking on the exploration of a protein energy landscape and its biologically relevant conformations.

Materials and Methods

Inference of co-evolutionary couplings and contact predictions. Full MSAs for the 18 families were downloaded from the Pfam database[10]. The alignments were filtered removing unaligned insertions, and keeping the remaining aligned positions. Repeated sequences, sequences containing non-natural amino acids or with a fraction of gaps greater than 0.2 were removed. The empirical frequencies for single and pairs of amino acids were computed from the final alignments as weighted averages to account for sampling biases (see SI Appendix for details). The parameters $\{J_{i,j}(\alpha,\beta)\}$ were obtained by finding the maximum of the (l_2-regularized) likelihood of the parameters as discussed in the SI Appendix. Numerical maximization of the likelihood requires the calculation of the frequencies $\{f_k\}$ as averages over the model distribution, that we estimated through multiple (20 to 64) parallel Metropolis Monte Carlo simulations, and a number of sweeps per gradient estimation ranging from 10^4 to 10^5 per simulation, depending on the system. The final co-evolutionary couplings $C_{i,j}$ between each pair of residues were calculated from the estimated coupling parameters $\{J_{i,j}(\alpha,\beta)\}$ using a protocol proposed by Ekeberg *et al.* [25] (see SI Appendix for details).

Protein coarse-graining model. The protein chain is described by the heavy atoms of the backbone plus the C_β atoms of the side chains. See the inset in Fig. S5, SI Appendix for a schematic illustration of all the non-bonded potentials and the protein coarse-graining. The complete potential is fully described in the SI Appendix and comprise contributions from the AMBER99SB-ILDN force field [54] to account for a correct backbone geometry, a non-bonded term in the form of a 12-6 Lennard-Jones function between C_β atoms with parameters ($r_{\beta\beta} = 0.55$ nm, $\epsilon = 15$ kJ/mol), that accounts for the co-evolution predictions and an hydrogen bond mimicking potential between O and N atoms in the form of a 12-6 Lennard-Jones function with parameters ($r_{ON} = 0.3$ nm, $\epsilon = 15$ kJ/mol), derived from the sequence analysis, to stabilize the helices. The simulation protocols and the analysis methods are detailed in the SI Appendix.

ACKNOWLEDGMENTS. We thank the following for computing time: PRACE Research Infrastructure resources MareNostrum based in Spain at the BSC and Curie in France at the TGCC-CEA (FP7 RI-283493), PRACE-3IP project (FP7 RI-312763) resources and the HECBioSim resource Archer in the UK. FLG acknowledges the Engineering and Physical Sciences Research Council for partial support [grant number EP/M013898/1]

1. Altschuh D, Vernet T, Berti P, Moras D, Nagai K (1988) Coordinated amino acid changes in homologous protein families. Protein engineering 2:193–199.
2. Korber B, Farber RM, Wolpert DH, Lapedes AS (1993) Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proceedings of the National Academy of Sciences 90:7176–7180.
3. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Bioinformatics 18:309–317.
4. Shindyalov I, Kolchanov N, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Engineering 7:349–358.
5. Taylor WR, Hatrick K (1994) Compensating changes in protein multiple sequence alignments. Protein Engineering 7:341–348.
6. Neher E (1994) How frequent are correlated changes in families of protein sequences? Proceedings of the National Academy of Sciences 91:98–102.
7. Casari G, Sander C, Valencia A (1995) A method to predict functional residues in proteins. Nature structural biology pp 171–8.
8. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. Journal of molecular biology 271:511–523.
9. de Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. Nature Reviews Genetics 14:249–261.
10. Finn RD, et al. (2013) Pfam: the protein families database. Nucleic acids research p gkt1223.
11. Lapedes A, Giraud B, Jarzynski C (2002) Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. http://library.lanl.gov/cgi-bin/getfile?01038177.pdf.
12. Socolich M, et al. (2005) Evolutionary information for specifying a protein fold. Nature 437:512–518.
13. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R (2005) Natural-like function in artificial ww domains. Nature 437:579–583.
14. Bialek W, Ranganathan R (2007) Rediscovering the power of pairwise interactions. arXiv preprint arXiv:0712.4397.
15. Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440:1007–1012.
16. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. Proceedings of the National Academy of Sciences 106:67–72.
17. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. Proceedings of the National Academy of Sciences 106:22124–22129.
18. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108:E1293–E1301.
19. Marks DS, et al. (2011) Protein 3d structure computed from evolutionary sequence variation. PloS one 6:e28766.
20. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Elife 3:e02030.
21. Hopf TA, et al. (2014) Sequence co-evolution gives 3d contacts and structures of protein complexes. Elife 3:e03430.
22. Burger L, Van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS computational biology 6:e1000633.
23. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics 79:1061–1078.
24. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28:184–190.
25. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. Physical Review E 87:012707.
26. Lui S, Tiana G (2013) The network of stabilizing contacts in proteins studied by coevolutionary data. J. Chem. Phys. 139:155103.
27. Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for boltzmann machines. Cognitive science 9:147–169.
28. Whitford PC, et al. (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. Proteins 75:430–441.
29. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH (2007) The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. J Phys Chem B 111:7812–7824.
30. Irback A, Sjunnesson F, Wallin S (2000) Three-helix-bundle protein in a Ramachandran model. Proc Natl Acad Sci USA 97:13614–13618.
31. Pasi M, Lavery R, Ceres N (2013) PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties. J. Chem. Theory Comput. 9:785–793.
32. Kim YC, Hummer G (2008) Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. J Mol Biol 375:1416–1433.
33. Camilloni C, Sutto L (2009) Lymphotactin: how a protein can adopt two folds. J Chem Phys 131:245105.
34. Saunders MG, Voth GA (2013) Coarse-Graining Methods for Computational Biology. Annu Rev Biophys 42:73–93.
35. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. Proc Natl Acad Sci USA 109:10340–10345.
36. Morcos F, Jana B, Hwa T, Onuchic JN (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci USA 110:20533–20538.
37. Cheng RR, Morcos F, Levine H, Onuchic JN (2014) Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proc Natl Acad Sci USA 111:E563–71.
38. Sutto L, Mereu I, Gervasio FL (2011) A Hybrid All-Atom Structure-Based Model for Protein Folding and Large Scale Conformational Transitions. J. Chem. Theory Comput. 7:4208–4217.
39. Han KF, Bystroff C, Baker D (1997) Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. Protein Sci 6:1587–1590.
40. Jaynes ET (1957) Information theory and statistical mechanics. Physical review 106:620.
41. Roudi Y, Tyrcha J, Hertz J (2009) Ising model for neural data: model quality and approximate methods for extracting functional connectivity. Physical Review E 79:051915.
42. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2014) Critical assessment of methods of protein structure prediction (casp)round x. Proteins: Structure, Function, and Bioinformatics 82:1–6.
43. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A (2014) Evaluation of residue–residue contact prediction in casp10. Proteins: Structure, Function, and Bioinformatics 82:138–153.
44. Sillitoe I, et al. (2012) New functional families (funfams) in cath to improve the mapping of conserved functional sites to 3d structures. Nucleic acids research p gks1211.
45. Möglich A, Ayers RA, Moffat K (2009) Structure and signaling mechanism of per-arnt-sim domains. Structure 17:1282–1294.
46. Kalaivani R, Srinivasan N (2015) Molecular BioSystems. Molecular BioSystems pp 1–17.
47. Downward J (2003) Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer 3:11–22.
48. Rojas AM, Fuentes G, Rausell A, Valencia A (2012) The ras protein superfamily: evolutionary tree and role of conserved amino acids. The Journal of cell biology 196:189–201.
49. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 298:937–953.
50. Kalbitzer HR, Spoerner M, Ganser P, Hozsa C, Kremer W (2009) Fundamental Link between Folding States and Functional States of Proteins. J Am Chem Soc 131:16714–16719.
51. Holm L, Rosenström P (2010) Dali server: conservation mapping in 3d. Nucleic acids research 38:W545–W549.
52. Hinton GE, Sejnowski TJ (1986) Learning and relearning in boltzmann machines. MIT Press, Cambridge, Mass 1:282–317.
53. Dago AE, et al. (2012) Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. Proceedings of the National Academy of Sciences 109:E1733–E1742.
54. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78:1950–1958.