Correlations from structure and phylogeny combine constructively in the inference of protein partners from sequences

Andonis Gerardos[1,2,¤], Nicola Dietler[1,2], Anne-Florence Bitbol[1,2*]

1 Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
2 SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

¤Current address: Aix Marseille Univ, CNRS, CINAM, Marseille, France
* anne-florence.bitbol@epfl.ch

Abstract

Inferring protein-protein interactions from sequences is an important task in computational biology. Recent methods based on Direct Coupling Analysis (DCA) or Mutual Information (MI) allow to find interaction partners among paralogs of two protein families. Does successful inference mainly rely on correlations from structural contacts or from phylogeny, or both? Do these two types of signal combine constructively or hinder each other? To address these questions, we generate and analyze synthetic data produced using a minimal model that allows us to control the amounts of structural constraints and phylogeny. We show that correlations from these two sources combine constructively to increase the performance of partner inference by DCA or MI. Furthermore, signal from phylogeny can rescue partner inference when signal from contacts becomes less informative, including in the realistic case where inter-protein contacts are restricted to a small subset of sites. We also demonstrate that DCA-inferred couplings between non-contact pairs of sites improve partner inference in the presence of strong phylogeny, while deteriorating it otherwise. Moreover, restricting to non-contact pairs of sites preserves inference performance in the presence of strong phylogeny. In a natural dataset, as well as in realistic synthetic data based on it, we find that non-contact pairs of sites contribute positively to partner inference performance, and that restricting to them preserves performance, evidencing an important role of phylogeny.

Author summary

In protein sequence data, the amino acid usages at different sites of a protein or of two interacting proteins can be correlated because of functional constraints. For instance, the need to maintain physico-chemical complementarity among two sites that are in contact in the three-dimensional structure of a protein complex causes such correlations. However, correlations can also arise due to shared evolutionary history, even in the absence of any functional constraint. While these phylogenetic correlations are known to obscure the inference of structural contacts, we show, using controlled synthetic data, that correlations from structure and phylogeny combine constructively to allow the inference of protein partners from sequences. We also show that pairs of amino acids that are not in contact in the structure have a major impact on partner inference in a natural dataset and in realistic synthetic ones. These findings explain the success of methods based on pairwise maximum-entropy models or on information theory at predicting protein partners from sequences.

Introduction

Most cellular processes are carried out by interacting proteins. Thus, mapping protein-protein interactions is a crucial goal. Since high-throughput experiments remain challenging [1], it is interesting to exploit the growing amount of available sequence data to identify candidate protein-protein interaction partners. The amino-acid sequences of interacting proteins are correlated, both because of evolutionary constraints arising from the need to maintain physico-chemical complementarity between amino acids that are in contact in the three-dimensional structure of protein complexes, and because of shared evolutionary history. On the one hand, correlations from structural contacts have received substantial interest, both within single proteins and across interacting protein partners. Global statistical models [2,3] using the maximum entropy principle [4] and designed to match the one- and two-body statistics of natural sequence data, often called Direct Coupling Analysis (DCA) [5], have been employed with success to determine three-dimensional protein structures from sequences [6–8], to analyze mutational effects [9–12], protein evolution [13] and conformational changes [14, 15], to design proteins [16], to find residue contacts between known interaction partners [5, 17–23], and to predict interaction partners among paralogs [24, 25] and protein-protein interaction networks [26, 27] from sequence data. On the other hand, correlations arise in protein sequences due to their common evolutionary history, i.e. phylogeny [28–30], even in the absence of structural constraints. Functionally related [31] and interacting [32] protein families tend to have similar phylogenies. This can arise from global shared evolutionary pressures on interacting partners, resulting in similar evolutionary rates [33–37], and from mere shared evolutionary history, including common timing of speciations and gene duplications [36]. Accordingly, methods based on sequence similarity, e.g. Mirrortree [38–42], or on the simultaneous presence and absence of genes, e.g. phylogenetic profiling [43–45] allow to predict which protein families interact. Mutual information (MI), which includes all types of statistical dependence between the sequences of interacting partners, slightly outperforms DCA at predicting interaction partners among paralogs [46]. While DCA allows to infer interaction partners in synthetic data that only comprises correlations from contacts [47], the inference of protein partners by DCA and MI also performs well on synthetic data that only includes phylogenetic correlations [48]. Therefore, correlations from contacts and from phylogeny are both useful to predict protein-protein interactions. This stands in contrast with the identification of structural contacts by DCA [5, 6, 30, 49–51], where phylogenetic correlations obscure structural ones, motivating the use of phylogeny corrections [52, 53], such as the Average Product Correction [54, 55], reweighting close sequences [6, 7, 55, 56], and Nested Coevolution [57].

How do DCA- and MI-based methods to infer protein partners perform in the presence of both phylogeny and structural contacts? Does successful inference mainly rely on one or the other of these two sources of correlations? Do they combine constructively or hinder each other? What changes when one dominates over the other? Answering these questions is important to understand the performance of DCA- and MI-based methods to infer protein partners in natural data, and should help to develop new methods that combine information from both phylogeny and contacts in an optimal way. To address them, we generate synthetic data in a minimal model that allows us to control the amounts of structural constraints and phylogeny. We also consider a data set of natural sequences, as well as synthetic data generated using models inferred on this natural data. Our focus is on predicting partners among paralogs, starting from a training set of known interaction partners, but our methods can be extended to the case where there is no training set via an Iterative Pairing Algorithm (IPA) [24, 46].

We find that correlations from structural contacts and from phylogeny add constructively in partner inference by DCA or MI. Furthermore, the signal from phylogeny can rescue partner inference in regimes of relatively weak selection and in the realistic case where inter-protein contacts are restricted to a small subset of sites. We show that DCA-inferred couplings between non-contact sites improve partner inference in the presence of strong phylogeny, while deteriorating it otherwise, and that they suffice to obtain good performance in the presence of strong phylogeny. In a natural dataset, as well as in realistic synthetic data, we find that non-contact pairs of sites contribute positively to partner inference performance, and that restricting to them preserves performance, evidencing an important role of phylogeny.

Model and methods

Modeling structural constraints with Potts models

General approach. We model the constraints stemming from the physicochemical complementarity of amino acids that are in contact in the three-dimensional structure of protein complexes by pairwise interactions in a Potts model. We consider concatenated sequences composed of two interacting partners A and B with respective lengths L_A and L_B. We denote by $\alpha_i \in \{1, \ldots, q\}$ the state of site $i \in \{1, \ldots, L_A + L_B\}$, where q is the number of possible states. The Hamiltonian of a concatenated sequence $\vec{\alpha} = (\alpha_1, ..., \alpha_{L_A+L_B})$ reads:

$$H(\vec{\alpha}) = -\sum_{i=1}^{L_A+L_B} h_i(\alpha_i) - \sum_{j=1}^{L_A+L_B} \sum_{i=1}^{j-1} e_{ij}(\alpha_i, \alpha_j)\,, \tag{1}$$

where fields h_i yield conservation, while (direct) couplings e_{ij} model pairwise interactions. Pairwise maximum entropy inference (DCA) yields the Potts Hamilitonian in Eq. 1 [5, 58].

Minimal model. In our minimal model, sequences are strings of binary variables represented by "Ising spins" taking values -1 or 1 ($q = 2$). Equivalently, one could take values 0 or 1, which is more usual for proteins [29, 59, 60], but we choose the spin convention to make the link with statistical physics [47]. These spins are coupled via uniform ferromagnetic couplings, set to unity, on all edges of an Erdős-Rényi random graph, all other couplings being zero, and all fields being zero. For simplicity we assume $L_A = L_B = L$. The Erdős-Rényi graph we mainly consider possesses $2L = 200$ vertices, and any two vertices are connected with probability $p = 0.02$. The corresponding Hamiltonian reads

$$H(\vec{\alpha}) = -\sum_{(i,j)\in\mathcal{E}} \alpha_i \alpha_j\ , \tag{2}$$

where the sequence $\vec{\alpha} = (\alpha_1, ..., \alpha_{2L}) \in \{\pm 1\}^{2L}$ is a string of $2L$ Ising spins, and $\mathcal{E}$ the set of edges of the Erdős-Rényi graph, which is fixed throughout, as it models the set of contacts of two given interacting protein families assembling into a specific complex structure. Because inter-protein contacts are generally sparser than intra-protein ones, we also study other graphs satisfying this constraint.

Models inferred from real data. We also generate more realistic synthetic data from Potts models inferred from a natural sequence data set composed of 23,633 pairs of natural sequences of interacting histidine kinases (HK) and response regulators (RR) from the P2CS database [61, 62]. There are $q = 21$ states, namely the 20 natural amino acids and the alignment gap. We use state-of-the-art methods that have good generative properties, namely bmDCA [16, 63] and arDCA [64]. In practice, we employ bmDCA with its default parameters for $q = 21$, and with default parameters except $t_{wait,0} = 1000$ and $\Delta t_0 = 100$ for $q = 2$ (motivated by the faster equilibration observed for $q = 2$). For arDCA, we use default parameters, apart from the reweighting parameter $\theta = 0.2$ (chosen to match the bmDCA value) and the regularization strengths $\lambda_J = 2 \times 10^{-4}$ and $\lambda_h = 10^{-5}$ for $q = 21$ or $\lambda_J = 2 \times 10^{-3}$ and $\lambda_h = 10^{-4}$ for $q = 2$ (chosen to reproduce one- and two-body frequencies well, see Figs. S8, S9 and S11).

Generating synthetic data with controlled amounts of structural constraints and phylogeny

General approach. We generate synthetic data using Markov Chain Monte Carlo sampling along a phylogenetic tree [30, 49], employing the Potts model Hamiltonian H in Eq. 1 or Eq. 2 to model structural constraints. As we focus on pairs of protein families with given structures, we assume that the ancestral protein complex already had the same structural constraints, and we take as our ancestral concatenated sequence AB an equilibrium sequence under the Hamiltonian H at sampling temperature

T. We then simulate evolution along the chosen phylogenetic tree (see below): random mutations are proposed at sites chosen uniformly at random, independently on each branch of the tree. Each proposed mutation is accepted with a probability p given by the Metropolis criterion at sampling temperature T:

$$p = \max\left[1, \exp\left(-\frac{\Delta H}{T}\right)\right], \tag{3}$$

where ΔH is the difference between the value of H after the mutation and before it. This models natural selection for maintaining structure [47]. Indeed, all mutations that decrease H are accepted ($p = 1$), while those that increase H can be rejected ($p < 1$), and will generally be rejected if T is small.

Minimal model. In our minimal model, a simple phylogeny is introduced via a binary branching tree with a fixed number n of "generations" (duplication events) and a fixed number μ of accepted mutations on each branch (between two subsequent duplication events; see Fig. 1). It gives rise to 2^n concatenated sequences AB on the leaves of the tree, which constitute a synthetic data set of paired sequences, where partners A and B evolved together along the tree. In practice, we choose $n = 10$, and thus $2^n = 1024$, ensuring that inference works well without phylogeny [47], and is computationally fast.

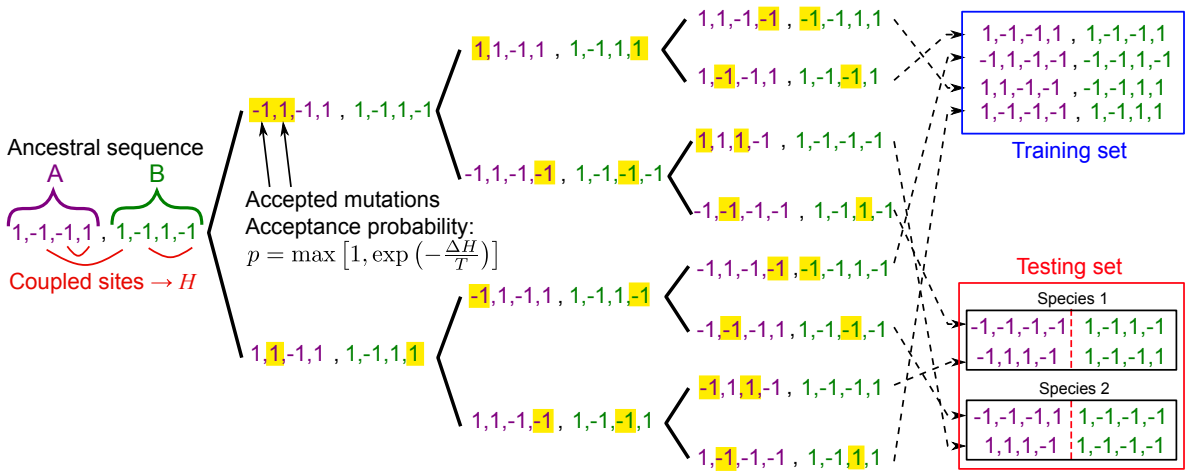

Fig 1. Construction of a synthetic dataset in the minimal model. A binary branching tree provides a minimal phylogeny. Structural constraints are represented by pairwise couplings on the edges of an Erdős-Rényi graph (Hamiltonian H in Eq. 2). Starting from an ancestral concatenated sequence AB of $2L$ Ising spins (here $L = 4$), a series of n duplication events ("generations", here $n = 3$) are performed. Between these duplications, and independently on each branch of the tree, μ mutations take place (here $\mu = 2$, and mutated sites are highlighted). Each mutation is an accepted spin flip at one site, and the acceptance criterion is given by Eq. 3. This process yields $2^n = 8$ chains AB at the tree leaves. This data is then randomly split into a training set and a testing set. In the testing set, groups of m sequences modeling species are randomly formed (here $m = 2$). Next, the evolutionary and functional pairings between chains A and B are blinded within each species of the testing set.

Controlling the importance of structural constraints and phylogeny. Tuning the sampling temperature T and the number μ of mutations per branch allows us to control the relative importance of structural constraints and phylogeny. First, Eq. 3 shows that when $T \to 0$, mutations that increase the value of H are all rejected, yielding strict selection for structure. Conversely, when $T \to \infty$, all mutations are accepted and evolution is neutral, so that all correlations in the data arise from phylogeny (and finite-size noise). Second, if μ is small, then all sequences resemble the ancestral one,

yielding extreme phylogenetic correlations. If μ is very large, even sister sequences arising from the last branching event become independent [48], leaving only correlations from structure.

Pure structural constraints limit and pure phylogeny limit. First, to consider the limiting case that only involves structural constraints, independent equilibrium sequences are generated using the Markov Chain Monte Carlo sampling scheme explained above (see Eq. 3). Each sequence is generated starting from a different initial random sequence [47]. The equilibration time is determined by the convergence of the Hamiltonian value, see Fig. S2 (note that the convergence of the absolute magnetization correlation function gives similar results [47]). Second, to consider the limiting case that only involves phylogeny, synthetic sequences are evolved along a phylogenetic tree, and all proposed mutations are accepted, simulating neutral evolution.

Models inferred from real data. We employ generative models inferred on natural paired HK-RR sequences (see above) to generate realistic synthetic data either without and with phylogeny. While bmDCA infers a Potts model, arDCA directly infers the distribution of probability of sequences [64]. First, to generate contact-only data, we employ a Markov Chain Monte Carlo procedure for bmDCA (equilibrium is considered reached after 10^6 accepted mutations for $q = 21$, or 10^5 for $q = 2$), while we directly sample independent sequences from the inferred distribution for arDCA. Second, to generate data that incorporates both phylogeny and contacts [49], we employ a tree inferred on the dataset of natural paired sequences via FastTree2 [65]. As the length b of a branch gives the mutation probability per site along it [65], we generate data by making $\lfloor b \times (L_A + L_B) \rfloor$ mutations on a branch of length b.

Inference problem

Question. Given two protein families A and B that interact, and starting from a training set of known AB partners, we aim to find, in each species of a testing set (see Fig. 1), which specific proteins A and B are evolutionary and functional partners. We assume for simplicity that there is a strict one-to-one pairing between each A and its partner B.

Species in the minimal model. In our minimal model, we randomly group concatenated AB sequences into sets of equal size m, representing species. The m different sequences A (or B) within a species represent paralogs. This minimal model, where species contain random assortments of sequences, is realistic if exchange between species (horizontal gene transfer) is sufficiently frequent. In Ref. [48], this random-species model was compared to more realistic ones, and qualitative conclusions were not affected, although the minimal model yields higher phylogenetic signal. In the testing set, within each species, we blind the pairings of the chains A and B. We then aim to infer these pairings, i.e. to recover for each A chain its evolutionary and functional partner, which is the B chain that coevolved with it.

Species in the model inferred from natural data. In our more realistic model where synthetic sequences are generated employing Hamiltonians and phylogenies inferred from natural sequence data, we rely on the inferred tree to define species. On each leaf of the inferred tree lies a natural paired sequence, coming from a given species. When generating data along this tree, we put the generated sequence on a given leaf of the tree into the species associated to this leaf in the natural data. The distribution of the number of sequence pairs per species in the synthetic data then exactly matches that of the natural data, and sequences in each species have the same positions in the tree as in the natural data. The inference question is then asked in the exact same way as in the minimal model.

Inference methods

Training set statistics. The statistics of the training set of paired chains AB, of total length L_A+L_B, are described using the empirical one-site frequencies of each state α_i at each site $i \in \{1, \ldots, L_A + L_B\}$,

denoted by $f_i(\alpha_i)$, and the two-site frequencies of occurrence of each ordered pair of states (α_i, α_j) at each ordered pair of sites (i, j), denoted by $f_{ij}(\alpha_i, \alpha_j)$. Covariances are computed as $C_{ij}(\alpha_i, \alpha_j) = f_{ij}(\alpha_i, \alpha_j) - f_i(\alpha_i) f_j(\alpha_j)$. When we employ mean-field DCA (mfDCA) and mutual information (MI), pseudocounts with weights denoted by λ and defined as in Refs. [24, 46, 48] are introduced [6, 7, 17, 46]. The value $\lambda = 0.5$ is usually employed in mfDCA [6, 7, 24], while smaller values have proved better for MI [46, 48]. Thus, we always take $\lambda = 0.5$ for mfDCA and $\lambda = 0.01$ for MI. However, we do not employ any phylogenetic reweighting (except where noted) because our aim is to investigate the effect of phylogeny, and also because this reweighting has very little impact on the inference of partners [24].

DCA-based inference method. DCA is based on building a global statistical model consistent with the empirical one- and two-body frequencies of the training set [5–7, 58], through the maximum entropy principle [4]. This results in a probability of observing a given sequence reading [58]:

$$P(\alpha_1, \ldots, \alpha_{L_A+L_B}) = \frac{\exp[-H]}{Z}\,, \tag{4}$$

where the Hamiltonian H is given by Eq. 1, i.e. by the Potts model, and Z is a normalization constant. Inferring the couplings and the fields that appropriately reproduce the empirical covariances is a difficult problem [66]. Within the mean-field approximation (mfDCA), which we employ for partner inference as in [24, 47, 48], inferred coupling strengths can be simply approximated by $\hat{e}_{ij}(\alpha_i, \alpha_j) = C^{-1}_{ij}(\alpha_i, \alpha_j)$ in the reference-state gauge [6, 7, 67]. One then makes a gauge change to the zero-sum (or Ising) gauge [24, 48, 54], which attributes the smallest possible fraction of the energy to the couplings, and the largest possible fraction to the fields [5, 54].

The effective interaction energy E_{AB} of each possible pair AB in the testing set is given by [24]

$$E_{AB} = -\sum_{i=1}^{L_A} \sum_{j=L_A+1}^{L_A+L_B} \hat{e}_{ij}(\alpha_i^A, \alpha_j^B)\,. \tag{5}$$

Note that we only sum over inter-protein pairs of sites (i.e. involving one site in A and one in B) because we focus on interactions between A and B.

MI-based inference method [46]. The pointwise mutual information (PMI) of a pair of states (α_i, α_j) at a pair of sites (i, j) is defined from the empirical one and two-body frequencies of the training set as [68–70]:

$$\mathrm{PMI}_{ij}(\alpha_i, \alpha_j) = \log\left[\frac{f_{ij}(\alpha_i, \alpha_j)}{f_i(\alpha_i) f_j(\alpha_j)}\right]\,. \tag{6}$$

A pairing score S_{AB} for each possible pair AB in the testing set can then be defined as the sum of the PMIs of the inter-protein pairs of sites of this concatenated chain AB (i.e. those involving one site in chain A and one site in chain B):

$$S_{\mathrm{AB}} = \sum_{i=1}^{L_A} \sum_{j=L_A+1}^{L_A+L_B} \mathrm{PMI}_{ij}(\alpha_i^A, \alpha_j^B)\,. \tag{7}$$

From scores to partner prediction. Our goal is to find the best 1-to-1 mapping of putative partners A-B in each species of the testing set. We assign a score to each possible partner of interaction using Eq. 5 or Eq. 7, and select the one-to-one assignment which optimizes the sum of scores for all chosen pairs by solving the corresponding linear assignment problem [71–74].

Code availability

The code associated to this paper is available at: `https://github.com/Bitbol-Lab/Phylogeny-Partners`.

Results

Correlations from structural contacts and from phylogeny both contribute to the performance of partner inference

In order to understand the origin of the performance of partner inference from protein sequences observed on real data using DCA [24] and MI [46], we construct synthetic datasets from a minimal model where the contributions of structural contacts and phylogeny can be tuned via the number μ of mutations per branch of the tree and the sampling temperature T (see Methods). How do these two parameters impact the performance of partner inference? Addressing this question will provide insight into the interplay of correlations from structural contacts and from phylogeny in partner inference.

Impact of the number of mutations per branch. Fig 2 shows the impact of varying the number μ of mutations per branch of the tree on the performance of partner inference, measured via the fraction of correctly predicted partner pairs (recall that each protein A in the testing set is paired with one partner B within its species, see Methods). In Fig 2, when $\mu \lesssim 15$, the performance of partner inference in our dataset that incorporates both structural contacts and phylogeny approaches that of a dataset that only involves phylogeny (see Methods and Ref. [48]). Indeed, similarities between related sequences are large for small μ, yielding dominant phylogenetic correlations. Conversely, when $\mu \gtrsim 70$, Fig 2 shows that the performance of partner inference in our data set including both ingredients approaches the one obtained with only structural contacts (see Methods and Ref. [47]). Indeed, when μ becomes large enough, similarities due to phylogeny vanish even between closest relatives, and all sequences become effectively independent. More precisely, the number of differences between two sister sequences AB arising from the last duplication events is about 2μ (exactly 2μ if all mutations affect different sites), and if it is of the same order as the total sequence length $2L$, or larger, i.e. if $\mu \gtrsim L = 100$ here, then even sister sequences lose all phylogenetic correlations. Accordingly, Fig 2 shows that for $\mu \gtrsim 90$, performance in the pure-phylogeny dataset drops to the chance expectation ("null model"), which corresponds to making random one-to-one pairings of sequences A and B within each species. Fig 2 also demonstrates that DCA and MI yield similar performance for partner inference, with MI becoming slightly better when phylogeny is not too strong, consistently with Refs. [46, 48].

Importantly, Fig 2 shows that partner inference performance in our data set including contacts and phylogeny is better than for both limiting data sets. Therefore, partner inference is made more robust by the fact that correlations from contacts and from phylogeny both contribute. Depending on how strong phylogeny is (i.e., here, on how small μ is), the dominant ingredient is either contacts or phylogeny, but in the generic case, these two signals add constructively to increase performance.

Impact of sampling temperature. Fig 3 shows the impact of varying the sampling temperature T on the performance of partner inference for two different values of μ, one where phylogeny dominates, $\mu = 15$, and one where contacts and phylogeny both have an important contribution, $\mu = 30$ (see Fig 2). The sampling temperature T impacts inference because $1/T$ is a proxy for the strength of selection on structural contacts (see Methods). In addition, in our minimal model, a phase transition between a ferromagnetic (ordered) phase where all spins tend to align and a paramagnetic (disordered) phase occurs at an intermediate critical temperature $T_c \approx 4.2$ (found by examining the absolute magnetization of sequences generated with only contacts, see Fig. S3, top panels). In the dataset with only contacts, this phase transition strongly impacts partner inference performance, with a peak around T_c apparent in Fig. 3 (see also Refs. [47, 75]). Qualitatively, at very low T, deep into the ferromagnetic phase, sequences are very similar to one another, as spins tend to all align, which makes inference difficult. At very high T, sequences become fully disordered and no longer reflect constraints from contacts, making inference difficult again. An optimum is thus expected at intermediate temperatures. The increased performance close to T_c has been studied in detail in Ref. [75] for contact prediction.

What is the impact of temperature when data are generated with both phylogeny and contacts? Fig 3 shows that phylogeny substantially increases performance for $T > T_c$. More precisely, when phylogeny

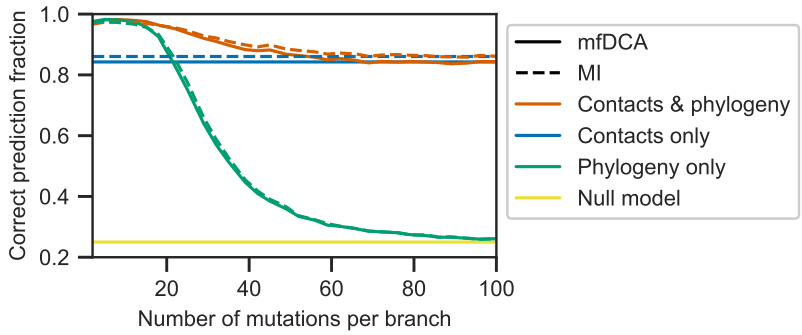

Fig 2. Impact of mutation number on partner inference performance. The fraction of correctly predicted partner pairs is shown versus the number μ of mutations per branch of the tree for the minimal model incorporating constraints both from contacts and from phylogeny. Specifically, an ancestral chain AB of $2L = 200$ spins is evolved along a binary branching tree (see Fig. 1) with $n = 10$ generations, and μ mutations per branch, yielding $2^{10} = 1024$ pairs AB, at sampling temperature $T = 5$ under the Hamiltonian 2 on an Erdős-Rényi graph with $p = 0.02$. The limiting cases with only contacts and only phylogeny are also shown for comparison. The first one corresponds to independent equilibrium sequences at $T = 5$ under the Hamiltonian 2 on the same graph. The second one corresponds to neutral evolution on the same binary branching tree. Each of these three datasets is randomly split into a training set of 400 chains AB and a testing set of 624 chains AB. The latter are randomly divided in 156 species of 4 chains AB each, and pairings between A and B chains are blinded in each species. Partnerships are then predicted using the scores in Eq. 5 (mfDCA) or Eq. 7 (MI). Each point is averaged over 20 generated data sets, and 20 random choices of the training set for each of them. The null model shows the expectation of the correct prediction fraction if pairs are made randomly within each species.

dominates ($\mu = 15$, Fig 3A), partner inference performance is modest at low temperatures, but it improves as T reaches T_c, and does not suffer the high-temperature decay observed in contact-only data when T increases above T_c. Performance remains very good at large T, tending to the phylogeny-only performance value, which is high in this regime, consistently with Ref. [48]. We note that, in the low-temperature regime, performance is worse than with contact-only data. In the ordered phase, about half of the contact-only sequences include mainly 1, while others include mainly -1 (see Fig. S1). In a typical species with four pairs AB comprising two sequences with mainly 1 and two with mainly -1 (or three of one type and one of the other), pairing the A and B chains of the same overall sign is easy, but degeneracy makes distinguishing among them very hard, making the baseline expectation of partner performance about 50% (more precisely, about 47%, as there is a 1/8 probability to get all four sequences of the same type). By contrast, in data generated with contacts and phylogeny, evolution starts from an equilibrium ancestral chain AB, which is already mainly composed of either 1 or -1. At low temperatures, switching the overall sign by successive mutations is difficult, and the whole phylogeny tends to retain the magnetization sign of its ancestor (see Fig. S1). Thus, the baseline expectation is only 25% (random within-species matching), thereby explaining the lower performance of partner inference for data with contacts and phylogeny at low T. However, this "freezing" of states occurring in the low-T limit in the present ferromagnetic Ising model is not expected to be the most relevant regime in real proteins. Thus, most other figures presented here correspond to $T = 5 > T_c$.

Let us now turn to the regime where phylogeny is less dominant ($\mu = 30$, Fig 3B). While the low-temperature results are similar to those for $\mu = 15$, performance now substantially decays in the high-temperature regime (this decay is very minor for $\mu = 15$). Indeed, contacts become less informative as disorder increases, as for contact-only data. However, phylogeny makes this decay less strong, and at very high temperature, performance tends to the phylogeny-only value. Fig 3 also confirms that DCA and MI yield similar results, with MI becoming slightly better at high temperatures.

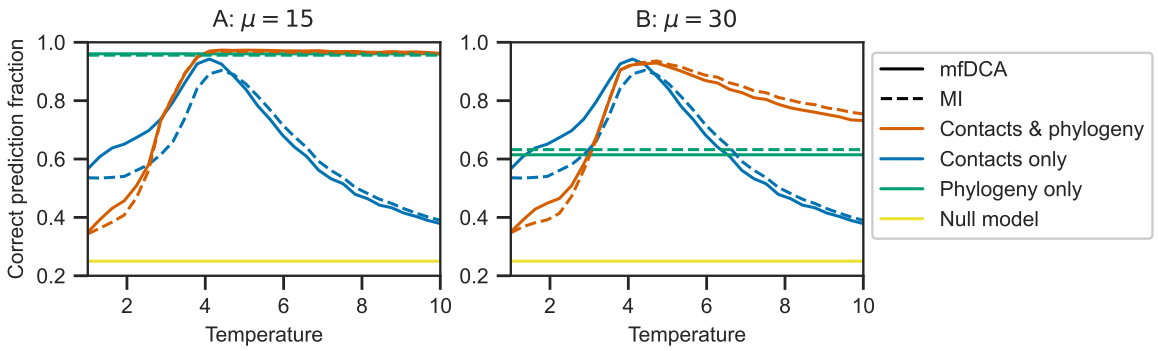

Fig 3. Impact of sampling temperature on partner inference performance. The fraction of correctly predicted partner pairs is shown versus the sampling temperature T for the minimal model incorporating constraints both from contacts and from phylogeny. Data generation and inference are performed exactly as in Fig. 2, using the same parameters and the same graph for contacts. However, here, T is varied and $\mu = 15$ (panel A) or $\mu = 30$ (panel B). The limiting cases with only contacts and only phylogeny are also shown for comparison. The null model shows the expectation of the correct prediction fraction if pairs are made randomly within each species.

Impact of the training set size and of the number of pairs per species. In addition to the parameters μ and T that allow us to tune the relative importance of phylogeny and contacts, other parameters strongly impact inference. First, Fig. S4 shows that a sufficiently large training set is required to accurately identify partners within each species. This holds both for DCA and for MI, but MI yields better performance for relatively small training sets, as for real data [46]. While trends are similar for $\mu = 15$ (Fig. S4A) and for $\mu = 30$ (Fig. S4B), larger training sets are required to obtain the same performance in the latter case, confirming the positive impact of phylogeny on partner inference. The need for sufficient training sets also holds for data sets including only contacts and only phylogeny (see also Refs. [47, 48]). These results are in line with previous ones obtained for DCA- [24, 25] and MI-based [46] predictions of protein-protein interactions from natural protein sequence data. Second, the pairing task becomes more difficult when the number of pairs per species increases. Accordingly, Fig. S5 shows that the performance of partner inference decays as species contain more pairs AB. This decay is slowest for our dataset including both contacts and phylogeny, highlighting that these two signals add constructively.

Impact of the graph of contacts. How does the set of structural contacts impact the performance of partner inference? Because inter-protein contacts are generally sparser than intra-protein ones, we now consider graphs of contacts that take into account this constraint [47], contrary to our minimal Erdős-Rényi graph. Apart from the graph defining contacts, the data generation process and the inference procedures are exactly the same as before. Fig. S6 shows that the same overall behavior is observed for the performance of partner inference for all graphs considered. With phylogeny, the range of temperature values leading to high performance is larger than for contact-only datasets. Moreover, for graphs possessing a smaller interface region between the two partners (Fig. S6 C-D), the signal from contacts only does not suffice for good inference, and phylogeny then rescues inference.

Couplings between non-contacting sites improve partner inference in the presence of strong phylogeny

Contributions of contacting and non-contacting sites to partner inference. How does phylogeny improve the inference of partners? So far, we have shown that phylogeny often enhances the performance of partner inference. Indeed, partners share a common evolutionary history (here, they are

generated together along the phylogenetic tree), and therefore, phylogeny yields correlations between sites that are informative of partnership. These phylogenetic correlations [48] are captured both by MI and DCA scores. Indeed, MI quantifies statistical dependence of any origin between random variables. The fact that DCA incorporates phylogenetic signal might seem more surprising, since it yields a Boltzmann distribution (Eq. 4) with a Potts Hamiltonian (Eq. 1), thus formally resembling an equilibrium physical model. However, DCA approximately constructs the maximum-entropy distribution matching the one- and two-body frequencies measured in the training set. Its training objective is to match these empirical frequencies, whatever their origin. Thus, inferred DCA couplings incorporate phylogeny. In our minimal model, structural contacts only exist on the graph edges, and other (non-contact) pairs of sites have zero couplings in the Hamiltonian 2 used for data generation. Nonzero values of the inferred couplings between non-contact sites can arise due to phylogeny, but also due to finite-size effects or to the approximations made in the inference procedure. Furthermore, the values of couplings between contacting sites can also be impacted by phylogeny, finite-size effects and inference approximations. How do couplings between contacting and non-contacting sites contribute to the inference in this synthetic data?

To address these questions, we restrict either to contact pairs of sites, or to non-contact pairs of sites in the score used for partner inference by DCA. Specifically, instead of the score E_{AB} in Eq. 5, we use either

$$E^C_{AB} = -\sum_{i=1}^{L_A} \sum_{j=L_A+1}^{L_A+L_B} \hat{e}_{ij}(\alpha^A_i, \alpha^B_j) I_{\mathcal{E}}(i,j)\,, \tag{8}$$

when restricting to contacts, or

$$E^{NC}_{AB} = -\sum_{i=1}^{L_A} \sum_{j=L_A+1}^{L_A+L_B} \hat{e}_{ij}(\alpha^A_i, \alpha^B_j)\left[1 - I_{\mathcal{E}}(i,j)\right]\,, \tag{9}$$

when restricting to non-contacts, where $I_{\mathcal{E}}(i,j)$ is 1 if $(i,j) \in \mathcal{E}$ and 0 otherwise, $\mathcal{E}$ being the set of edges of the graph representing contacts. Because in our minimal model $e_{ij}(\alpha_i, \alpha_j) = \alpha_i \alpha_j$ if $(i,j) \in \mathcal{E}$ and 0 otherwise, we also use the score

$$E^R_{AB} = -\sum_{i=1}^{L_A} \sum_{j=L_A+1}^{L_A+L_B} \alpha^A_i \alpha^B_j I_{\mathcal{E}}(i,j)\,, \tag{10}$$

which would coincide with that in Eq. 8 if inference was perfect.

Fig. 4 shows the performance of inference when these different scores are used. When phylogeny dominates, i.e. for small μ, restricting to contacts by using the score in Eq. 8 strongly deteriorates the performance of inference. Thus, couplings between non-contacting sites then include phylogenetic information relevant to infer partners. Conversely, when the effect of phylogeny is smaller, i.e. for large μ, restricting to contacts improves inference performance compared to using the full score. This deleterious impact of non-contacting pairs is probably due to the fact that these couplings also arise from finite-size noise and inference approximations, in addition to phylogeny. Results obtained from the score based on the real Hamiltonian Eq. 10 are close to those obtained with Eq. 8, and even a little worse when phylogeny is very strong, because the real Hamiltonian cannot capture any phylogeny.

By contrast, restricting to non-contact pairs via the score in Eq. 9 yields an inference performance as good as with the complete score in Eq. 8 for very small μ, confirming that non-contact pairs incorporate most phylogenetic signal. As μ increases and phylogenetic signal weakens, performance using the score in Eq. 9 decays sharply. This decay resembles that observed in Fig. 2 for partner inference on purely phylogenetic data, confirming that non-contact couplings mainly arise from phylogeny. However, performance remains higher than the null model for large μ, while it tends to it with purely phylogenetic data. This residual performance might arise from the fact that non-contact couplings contain indirect correlations, stemming from contacts but mediated by intermediate sites, despite the fact that DCA reduces such contributions compared to covariance or mutual information [5, 75]. The importance of

indirect correlations could be impacted by the approximations made when inferring couplings. To explore this, we employed couplings inferred by bmDCA, which yields a much better inference quality than mfDCA in terms of generative properties [63], but is computationally much heavier. Fig. S7 shows that the results using bmDCA are extremely similar to those obtained by mfDCA, which shows the robustness of our conclusions.

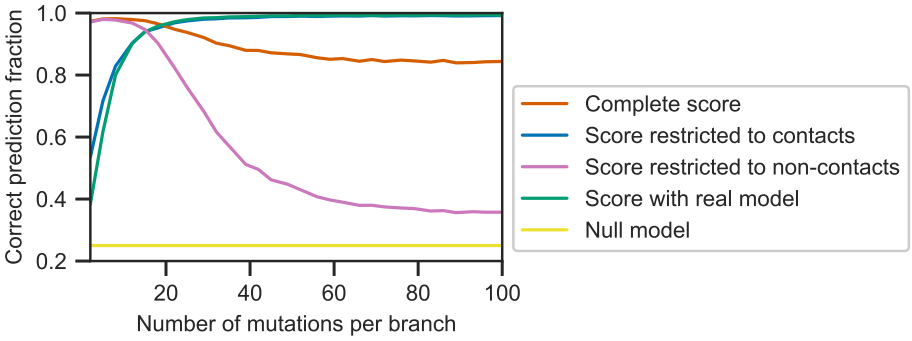

Fig 4. Impact of contact and non-contact pairs of sites on partner inference performance. The fraction of correctly predicted partner pairs is shown versus the number μ of mutations per branch of the tree for the minimal model incorporating constraints both from contacts and from phylogeny, either with the full score defined in Eq. 5 with couplings inferred using the training set, or with this score restricted to the pairs of sites that are actually in contact (Eq. 8), or to those that are not in contact (Eq. 9), or with a score computed with the real couplings used for data generation (Eq. 10). Data generation and inference (apart from the score definition) are performed exactly as in Fig. 2, using the same parameters and the same graph for contacts.

Impact of gradually removing non-contacting pairs of sites. To gain further insight into the impact of non-contact pairs of sites on partner inference, we next investigate the impact of removing them gradually. We remove them either randomly, or by decreasing rank order of the absolute value of the inferred couplings, or based on mutation timing. Indeed, non-contact pairs may have high impact because they have large absolute inferred coupling values, and/or because they feature strong phylogenetic correlations. Fig. 5 shows histograms of the values of inferred Ising couplings $\hat{J}_{ij}$, defined by $\hat{e}_{ij}(\alpha_i, \alpha_j) = \hat{J}_{ij}\alpha_i\alpha_j$, between contact and non-contact pairs of sites, for $\mu = 5$ (Fig. 5A) and for $\mu = 30$ (Fig. 5C). In the first case, where phylogeny dominates, the values of $\hat{J}_{ij}$ do not allow one to distinguish contacts from non-contacts. In the second case, contact pairs feature higher values of $\hat{J}_{ij}$, which would allow to infer them better, although there is still a strong overlap, partly due to the rather small training set (100 sequences) employed here, which was chosen for the partner inference task to be successful with the score in Eq. 5, but still improvable (see Fig. S4). Next, Fig. 5 B and D show the impact of progressively removing non-contact couplings on the performance of partner inference. For small μ, removing them decreases performance, while the opposite holds for large μ, consistent with Fig. 4. Moreover, for $\mu = 5$ (Fig. 5B), removing them in decreasing rank order of absolute $\hat{J}_{ij}$ yields a quicker and sharper decay of performance than removing them in random order. Thus, large absolute non-contact couplings contain relevant information for partner inference. Removing them in increasing order of the earliest time along the phylogeny when mutations have affected both sites i and j yields a similarly quick and sharp decay, corroborating the idea that phylogeny is the main relevant source of information in these non-contacting pairs. Indeed, early coupled mutations lead to large phylogenetic correlations. Conversely, for $\mu = 30$ (Fig. 5D), ranking-based removal of non-contacts leads to an earlier increase of performance than random removal, but earliest-mutation-based removal is almost equivalent to random removal. This illustrates the reduced importance of phylogeny in this case.

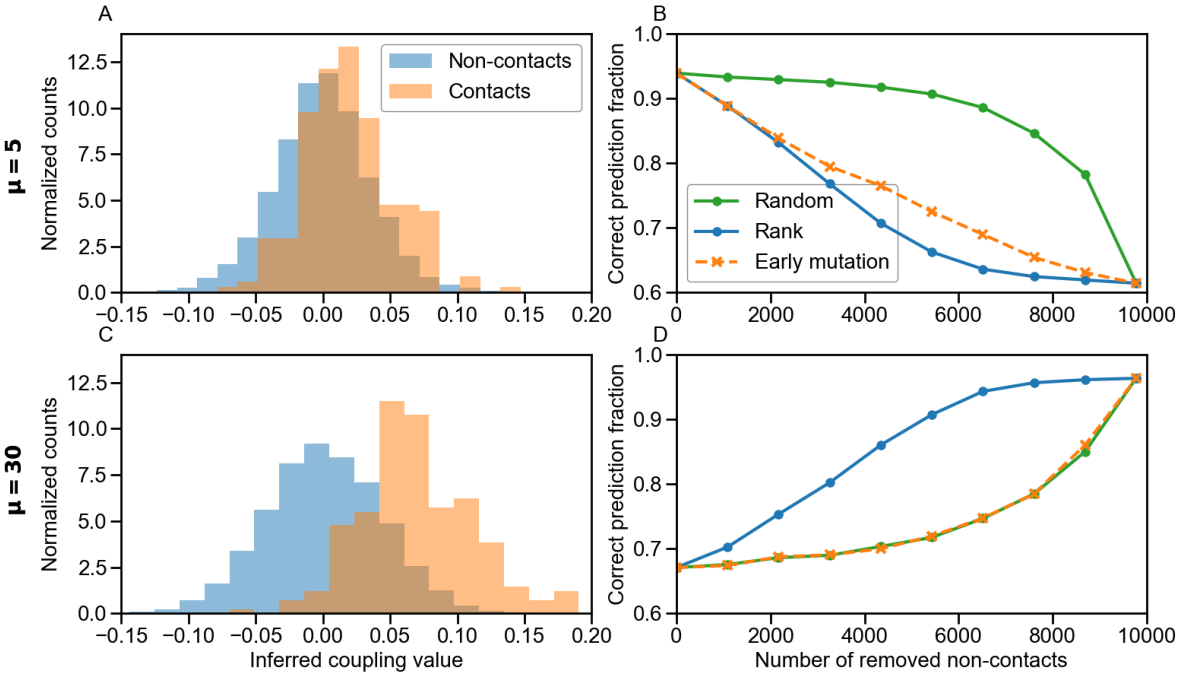

Fig 5. Impact of gradually removing non-contact pairs of sites on partner inference performance. Panels A and C show the normalized histogram of the inferred couplings $\hat{J}_{ij}$ for contact and non-contact pairs of sites. Panels B and D show the fraction of correctly predicted partner pairs versus the number of removed non-contact pairs of sites, when gradually removing their contribution from the score in Eq. 5, going toward the one in Eq. 8. The non-contact pairs are removed either randomly, or in decreasing rank order of $|\hat{J}_{ij}|$ values, or in increasing order of the earliest time (i.e., earliest generation in the tree) when mutations have affected both sites i and j. In A and B, $\mu = 5$, while in C and D, $\mu = 30$. All data is shown for the minimal model incorporating both constraints from contacts and phylogeny. Data generation and inference (apart from the score definition) are performed exactly as in Fig. 2, using the same graph for contacts and the same parameters, except that here, the training set comprises 100 paired chains AB. In panels A and C, data is generated just once. In panels B and D, each point is averaged over 100 generations of the data.

Interplay of contacts and phylogeny in natural data and in synthetic data generated from models inferred on natural data

Our minimal model allows us to tune the importance of contacts and phylogeny, but contains strong simplifications. Natural data comprises $q = 21$ possible states, which are the 20 natural amino acids and the alignment gap. Potts models inferred from natural data involve broad distributions of couplings, and include nonzero fields [5–7, 51]. Phylogenies inferred from data are also much more complex than our binary tree with a fixed number of mutations on each branch [49, 65], and the assignment of sequences to species results from speciation and horizontal gene transfer. Do the conclusions obtained with our minimal model hold for natural data? How important are signals from phylogeny and contacts in natural data? To address this question, we consider a data set composed of 23,633 interacting pairs of natural sequences of histidine kinases (HK) and response regulators (RR) from the P2CS database [61, 62]. To gain further insight into the importance of phylogenetic signal, we infer generative models of this paired sequence alignment using two state-of-the-art methods, bmDCA [63] and arDCA [64] (see Methods), and we generate data from them, either without phylogeny or along a phylogeny inferred from the natural alignment (see Methods, and Ref. [49]). We checked that two-body and one-body frequencies

of the original data set were well-reproduced by those of the dataset generated without phylogeny (Fig. S11). We also checked that the inference of contacts was possible on the generated data sets, although it deteriorates when generating with phylogeny (Fig. S12).

Fig. 6 shows the fraction of correct predicted pairs versus the size of the training set for the natural data set, as well as for the synthetic data sets generated from inferred models. Results employing the usual mfDCA-based score in Eq. 5 are qualitatively similar for the real and synthetic data sets, as well as to the results from our minimal model (Fig. S4). Furthermore, the performance of partner inference is similar for synthetic data sets generated from inferred models with and without phylogeny. This is *a priori* reminiscent of cases with relatively low phylogeny in the minimal model (Fig. S4B). However, DCA models inferred from natural data reproduce all empirical correlations, including those from phylogeny. Some inferred couplings are thus of phylogenetic origin, and give rise to correlations in the sequences generated from these models without phylogeny. These additional couplings can help partner inference, by extending the set of pairs of sites that can yield information relevant for pairing, and by increasing the diversity of covariance and coupling values. This may contribute to the performance of partner inference in the synthetic data set generated without phylogeny. This effect can be assessed in our minimal model, by generating data from models inferred by bmDCA or arDCA (the initial data employed for inference being generated using the Hamiltonian 2, with or without phylogeny, see Figs. S8 and S9). Fig. S10 shows that in our minimal model, partner inference performance is higher with data generated from inferred models, compared to similar data generated directly from the original Hamiltonian 2. This demonstrates that inferred models incorporating couplings from various sources, including but not restricted to phylogeny, yield better partner inference performance than contact-only models. This effect explains at least partly the performance of partner inference for synthetic datasets generated from more realistic inferred models but without phylogeny in Fig. 6.

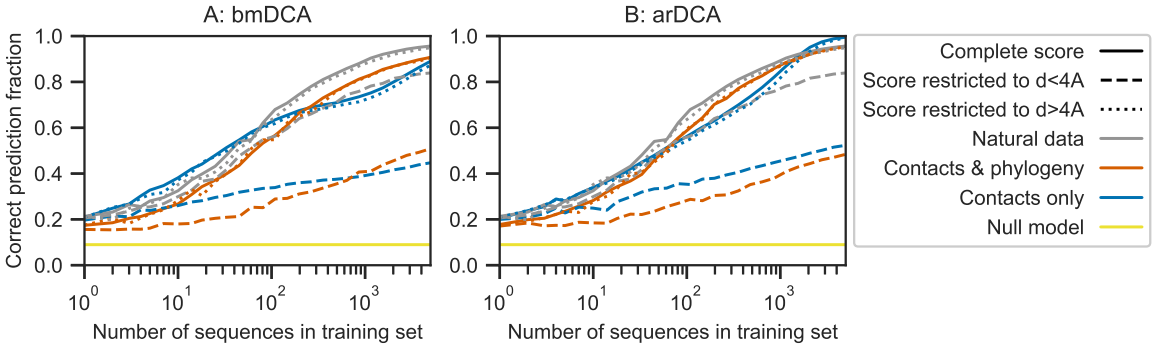

Fig 6. Partner inference in natural data and in synthetic data generated from models inferred on natural data. The fraction of correctly predicted partner pairs is shown versus the number of sequence pairs AB in the training set for a natural dataset of paired HK-RR sequences, and for synthetic datasets generated using models learned from this natural dataset by bmDCA (panel A) or arDCA (panel B), either with or without phylogeny. In each case, the full dataset is randomly split into a training set comprising a variable number of chains AB and a testing set. The latter is divided in species as defined in the natural dataset (for synthetic datasets, the species labels from the natural dataset are employed, and left associated to the leaves of the phylogenetic tree if data is generated with phylogeny). Partnerships are then predicted using the score in Eq. 5 based on mfDCA, or with its variants restricting to amino-acid pairs in contact (Eq. 8) or to those not in contact (Eq. 9). Contacts are defined using a threshold of 4 Å between closest atoms in the experimental HK-RR complex structure from Ref. [76]. The null model shows the expectation of the correct prediction fraction if pairs are made randomly within each species.

In this context, it is interesting to investigate the effect of restricting to contact pairs of amino-acids or to non-contact ones. Fig. 6 shows partner inference performance for the reduced scores in Eqs. 8 and 9.

It demonstrates that restricting to contact pairs (Eq. 8) deteriorates partner inference performance, both for natural data, and for data generated using DCA models inferred from natural data (even more in the latter case, consistent with the less good contact inference from generated data, see Fig. S12). Moreover, restricting to non-contact pairs (Eq. 9) only yields a very minor decrease of inference performance for all these data sets, compared to using the full score (Eq. 5). Thus, strikingly, the information contained in non-contact pairs is sufficient for inference. Overall, our results on natural data and on realistic synthetic data are consistent with what is observed for strong to relatively-strong phylogeny in our minimal model (Figs. 4, 5 and S10).

Discussion

While they obscure the identification of contacts by coevolution methods [5,6,30,51,53], correlations that arise in protein sequences due to phylogeny [28–32,38,41] become useful in order to identify interaction partners from protein sequences. Indeed, interaction partners tend to have similar evolutionary histories, which is directly exploited in some protein-protein interaction prediction methods [38–45]. In this context, the success of DCA- and MI-based approaches at predicting protein-protein interaction partners among paralogs from natural protein sequences [24,25,46] can potentially be due to correlations from structural contacts needing to maintain their complementarity, or to correlations from phylogeny, or both. Shedding light on the origin of the performance of these methods is an important step toward constructing better ones. However, disentangling the impact of different sources of signal in natural data is a difficult task.

In this study, we generated and analyzed synthetic data produced within a minimal model that allows us to control the amounts of structural constraints and phylogeny. We showed that these two signals add constructively to increase the performance of partner inference by DCA or MI. Furthermore, signal from phylogeny can rescue partner inference in cases where signal from structural contacts becomes less informative, including in the realistic case where inter-protein contacts are restricted to a small subset of sites. We also demonstrated that DCA-inferred couplings between non-contact pairs of sites improve partner inference in the presence of strong phylogeny, while deteriorating it otherwise. Furthermore, in the strongly phylogenetic regime, inference is almost as good when restricting to non-contact pairs of sites than when including all pairs. In a natural dataset, as well as in realistic synthetic data based on it, we found that non-contact pairs of sites contribute positively to partner inference performance. Moreover, restricting to non-contact pairs of amino acids yields inference performances that are very close to those obtained when all pairs are accounted for. These results are in line with what was observed in our minimal model with strong phylogeny, and evidence an important role of phylogeny in partner inference on natural data.

Several interesting extensions are possible. First, in this work, we considered the problem of partner inference starting from a training set of known partners, but our methods allow us to address the case where there is no training set via an Iterative Pairing Algorithm (IPA) [24,46]. It would be interesting to extend the present study to this case. Next, while our study of synthetic data generated using models inferred from natural data allowed us to bridge our minimal model and natural data, we were faced with the issue that inferred couplings include phylogeny. Thus, disentangling signals was much harder than in the minimal model, as the couplings from phylogeny make the model richer even in the absence of phylogeny in the data generation step. While this is a difficult problem, it could be partially addressed by applying phylogeny corrections to the inferred couplings [51,57]. This could also shed light on whether some of the useful signal from non-contact pairs is coming from collective functional constraints, similar to sectors in single proteins [29,57], an interesting possibility that was not explored here. Finally, understanding the relative impact of structural constraints and phylogeny in the inference of interaction partners from sequences opens the way to exploiting them together more efficiently.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 851173, to A.-F. B.).

References

1. Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, et al. The binary protein-protein interaction landscape of Escherichia coli. Nat Biotechnol. 2014;32(3):285–290.

2. Lapedes AS, Giraud BG, Liu L, Stormo GD. Correlated mutations in models of protein sequences: phylogenetic and structural effects. In: Statistics in molecular biology and genetics - IMS Lecture Notes - Monograph Series. vol. 33; 1999. p. 236–256.

3. Burger L, van Nimwegen E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol. 2008;4:165.

4. Jaynes ET. Information Theory and Statistical Mechanics. Phys Rev. 1957;106(4):620–630.

5. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA. 2009;106(1):67–72.

6. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS ONE. 2011;6(12):e28766.

7. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA. 2011;108(49):E1293–1301.

8. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proc Natl Acad Sci USA. 2012;109(26):10340–10345.

9. Dwyer RS, Ricci DP, Colwell LJ, Silhavy TJ, Wingreen NS. Predicting functionally informative mutations in Escherichia coli BamA using evolutionary covariance analysis. Genetics. 2013;195(2):443–455.

10. Cheng RR, Morcos F, Levine H, Onuchic JN. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proc Natl Acad Sci USA. 2014;111(5):E563–571.

11. Cheng RR, Nordesjo O, Hayes RL, Levine H, Flores SC, Onuchic JN, et al. Connecting the Sequence-Space of Bacterial Signaling Proteins to Phenotypes Using Coevolutionary Landscapes. Mol Biol Evol. 2016;33(12):3054–3064.

12. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Mol Biol Evol. 2016;33(1):268–280.

13. de la Paz JA, Nartey CM, Yuvaraj M, Morcos F. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. Proc Natl Acad Sci U S A. 2020;117(11):5873–5882.

14. Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci USA. 2013;110(51):20533–20538.

15. Malinverni D, Marsili S, Barducci A, De Los Rios P. Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. PLoS Comput Biol. 2015;11(6):e1004262.

16. Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, et al. An evolution-based model for designing chorismate mutase enzymes. Science. 2020;369(6502):440–445.

17. Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. PLoS ONE. 2011;6(5):e19729.

18. Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, et al. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. PLoS ONE. 2014;9(3):e92721.

19. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife. 2014;3:e02030.

20. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife. 2014;3:e03430.

21. Tamir S, Rotem-Bamberger S, Katz C, Morcos F, Hailey KL, Zuris JA, et al. Integrated strategy reveals the protein interface between cancer targets Bcl-2 and NAF-1. Proc Natl Acad Sci USA. 2014;111(14):5177–5182.

22. dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN. Dimeric interactions and complex formation using direct coevolutionary couplings. Sci Rep. 2015;5:13652.

23. Feinauer C, Szurmant H, Weigt M, Pagnani A. Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. PLoS ONE. 2016;11(2):e0149166.

24. Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. Proc Natl Acad Sci USA. 2016;113(43):12180–12185.

25. Gueudre T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc Natl Acad Sci USA. 2016;113(43):12186–12191.

26. Cong Q, Anishchenko I, Ovchinnikov S, Baker D. Protein interaction networks revealed by proteome coevolution. Science. 2019;365(6449):185–189.

27. Green AG, Elhabashy H, Brock KP, Maddamsetti R, Kohlbacher O, Marks DS. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. Nat Commun. 2021;12(1):1396.

28. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol. 1995;2(2):171–178.

29. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell. 2009;138(4):774–786.

30. Qin C, Colwell LJ. Power law tails in phylogenetic systems. Proc Natl Acad Sci USA. 2018;115(4):690–695.

31. Fryxell KJ. The coevolution of gene family trees. Trends Genet. 1996;12(9):364–369.

32. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co-evolution of proteins with their interaction partners. J Mol Biol. 2000;299(2):283–293.

33. Hakes L, Lovell SC, Oliver SG, Robertson DL. Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc Natl Acad Sci USA. 2007;104(19):7999–8004.

34. Juan D, Pazos F, Valencia A. Co-evolution and co-adaptation in protein networks. FEBS Lett. 2008;582(8):1225–1230.

35. Kann MG, Shoemaker BA, Panchenko AR, Przytycka TM. Correlated evolution of interacting proteins: looking behind the mirrortree. J Mol Biol. 2009;385(1):91–98.

36. Lovell SC, Robertson DL. An integrated view of molecular coevolution in protein-protein interactions. Mol Biol Evol. 2010;27(11):2567–2575.

37. Swapna LS, Srinivasan N, Robertson DL, Lovell SC. The origins of the evolutionary signal used to predict protein-protein interactions. BMC Evol Biol. 2012;12:238.

38. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng Des Sel. 2001;14(9):609–614.

39. Jothi R, Kann MG, Przytycka TM. Predicting protein-protein interaction by searching evolutionary tree automorphism space. Bioinformatics. 2005;21 Suppl 1:i241–250.

40. Bradde S, Braunstein A, Mahmoudi H, Tria F, Weigt M, Zecchina R. Aligning graphs and finding substructures by a cavity approach. EPL. 2010;89(3).

41. Ochoa D, Pazos F. Studying the co-evolution of protein families with the Mirrortree web server. Bioinformatics. 2010;26(10):1370–1371, `http://csbg.cnb.csic.es/mtserver`.

42. Ochoa D, Juan D, Valencia A, Pazos F. Detection of significant protein coevolution. Bioinformatics. 2015;31(13):2166–2173, `http://csbg.cnb.csic.es/pMT/`.

43. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999;96(8):4285–4288.

44. Croce G, Gueudré T, Ruiz Cuevas MV, Keidel V, Figliuzzi M, Szurmant H, et al. A multi-scale coevolutionary approach to predict interactions between protein domains. PLoS Comput Biol. 2019;15(10):e1006891.

45. Moi D, Kilchoer L, Aguilar PS, Dessimoz C. Scalable phylogenetic profiling using MinHash uncovers likely eukaryotic sexual reproduction genes. PLoS Comput Biol. 2020;16(7):e1007553.

46. Bitbol AF. Inferring interaction partners from protein sequences using mutual information. PLoS Comput Biol. 2018;14(11):e1006401.

47. Gandarilla-Pérez CA, Mergny P, Weigt M, Bitbol AF. Statistical physics of interacting proteins: Impact of dataset size and quality assessed in synthetic sequences. Phys Rev E. 2020;101:032413. doi:10.1103/PhysRevE.101.032413.

48. Marmier G, Weigt M, Bitbol AF. Phylogenetic correlations can suffice to infer protein partners from sequences. PLoS Comput Biol. 2019;15(10):e1007179.

49. Vorberg S, Seemayer S, Söding J. Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. PLOS Computational Biology. 2018;14(11):1–25. doi:10.1371/journal.pcbi.1006526.

50. Rodriguez Horta E, Barrat-Charlaix P, Weigt M. Toward Inferring Potts Models for Phylogenetically Correlated Sequence Data. Entropy. 2019;21(11). doi:10.3390/e21111090.

51. Rodriguez Horta E, Weigt M. On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins. PLoS Comput Biol. 2021;17(5).

52. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 1996;257(2):342–358.

53. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008;24(3):333–340.

54. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E. 2013;87(1):012707.

55. Hockenberry AJ, Wilke CO. Phylogenetic weighting does little to improve the accuracy of evolutionary coupling analyses. Entropy. 2019;21(10).

56. Malinverni D, Barducci A. Coevolutionary Analysis of Protein Subfamilies by Sequence Reweighting. Entropy. 2020;21(11):1127.

57. Colavin A, Atolia E, Bitbol AF, Huang KC. Extracting the phylogenetic dimension of coevolution reveals hidden functional signals. BioRxiv Preprint; p. DOI `10.1101/2020.09.23.310300`.

58. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys. 2018;81(3):032601.

59. Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, et al. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc Natl Acad Sci USA. 2011;108(28):11530–11535.

60. Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, et al. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Comput Biol. 2014;10(8):e1003776.

61. Barakat M, Ortet P, Jourlin-Castelli C, Ansaldi M, Mejean V, Whitworth DE. P2CS: a two-component system resource for prokaryotic signal transduction research. BMC Genomics. 2009;10:315.

62. Barakat M, Ortet P, Whitworth DE. P2CS: a database of prokaryotic two-component systems. Nucleic Acids Res. 2011;39(Database issue):D771–776.

63. Figliuzzi M, Barrat-Charlaix P, Weigt M. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? Mol Biol Evol. 2018;35(4):1018–1027.

64. Trinquier J, Uguzzoni G, Pagnani A, Zamponi F, Weigt M. Efficient generative modeling of protein sequences using simple autoregressive models. Nat Commun. 2021;12(1):5800.

65. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010;5(3):1–10. doi:10.1371/journal.pone.0009490.

66. Nguyen HC, Zecchina R, Berg J. Inverse statistical problems: from the inverse Ising problem to data science. Advances in Physics. 2017;66(3):197–261.

67. Plefka T. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. J Phys A: Math Gen. 1982;15(6):1971–1978.

68. Fano R. Transmission of Information: A Statistical Theory of Communications. Cambridge, MA: The MIT Press; 1961.

69. Church KW, Hanks P. Word association norms, mutual information, and lexicography. Comput Linguist. 1990;16(1):22–29.

70. Role F, Nadif M. Handling the impact of low frequency events on co-occurrence based measures of word similarity. In: Filipe J, Fred A, editors. Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011). Scitepress; 2011. p. 218–223.

71. Kuhn HW. The Hungarian Method for the assignment problem. Naval Research Logistics Quarterly. 1955;2:83–97.

72. Munkres J. Algorithms for the Assignment and Transportation Problems. Journal of the Society for Industrial and Applied Mathematics. 1957;5(1):32–38.

73. Buehren M. Functions for the rectangular assignment problem; 2011. Available from: `https://fr.mathworks.com/matlabcentral/fileexchange/6543-functions-for-the-rectangular-assignment-problem`.

74. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.

75. Ngampruetikorn V, Sachdeva V, Torrence J, Humplik J, Schwab DJ, Palmer SE. Inferring couplings in networks across order-disorder phase transitions. ArXiv Preprint; p. `https://arxiv.org/abs/2106.02349`.

76. Casino P, Rubio V, Marina A. Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell. 2009;139(2):325–336.

Supporting information

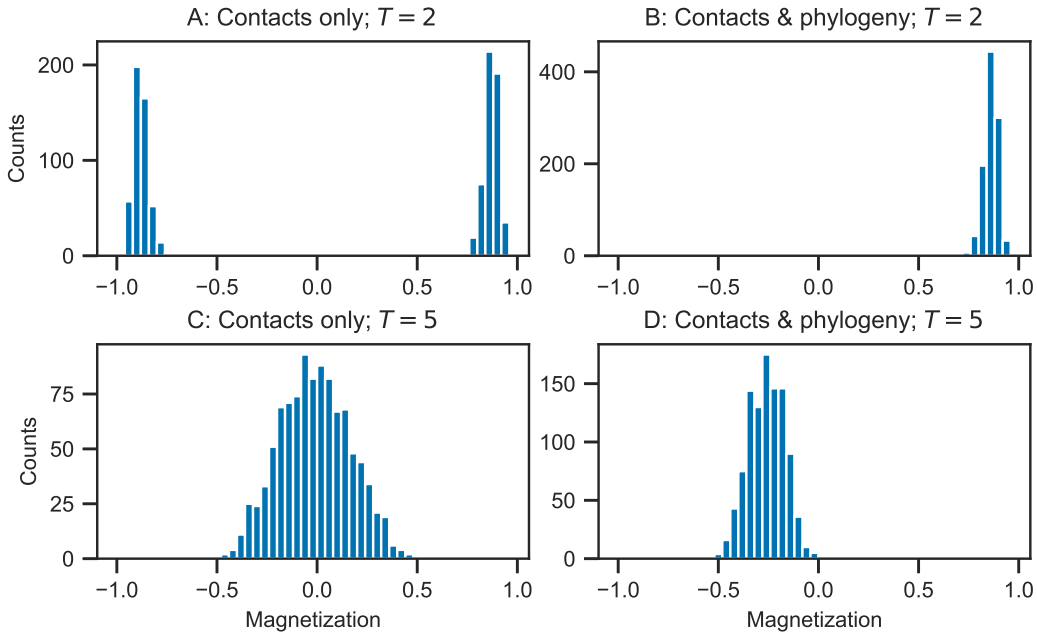

Fig S1. Magnetization for data generated without and with phylogeny. Histograms of the mean magnetization $m = (\sum_{i=1}^{2L} \alpha_i)/(2L)$ of a paired chain are shown at two different sampling temperatures T ($T = 2 < T_c$ in panels A and B; $T = 5 > T_c$ in panels C and D) for data generated without (contacts only, panels A and C) or with phylogeny (contacts & phylogeny, panels B and D). Contacts are defined by the same Erdős-Rényi graph with $p = 0.02$ as in Fig. 2. In the contact-only case, histograms are computed on an equilibrium dataset of 1024 independent sequences generated by Metropolis Monte Carlo sampling under the Hamiltonian 2 (without phylogeny). In practice, data is taken after 10,000 accepted mutations, see Fig. S2. In the contacts & phylogeny case, histograms are computed on a dataset of 1024 sequences generated on a single binary branching tree with $\mu = 15$ mutations per branch.

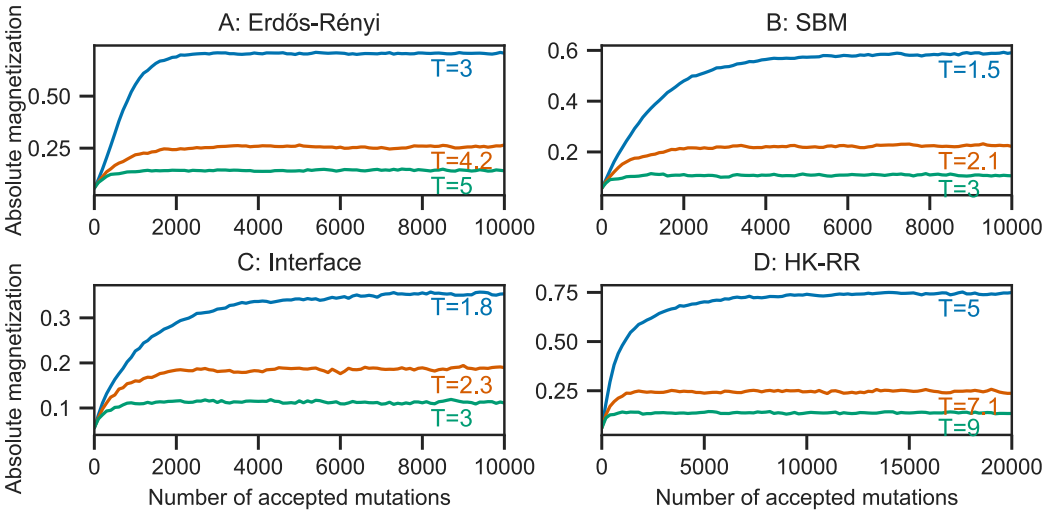

Fig S2. Convergence of absolute magnetization in the contact-only limit with different graphs. The mean of the absolute magnetization $|m| = |(\sum_{i=1}^{2L} \alpha_i)/(2L)|$ of a paired chain is shown versus the number of accepted mutations for each graph considered in Fig. S6. For each curve, datasets of 5000 independent sequences were generated by Metropolis Monte Carlo sampling under the Hamiltonian 2 (without phylogeny). The graphs considered are: (A) the same Erdős-Rényi graph with $p = 0.02$ as in Fig. 2; (B) a stochastic block model graph with two blocks of 100 nodes each, and $p = 0.02$ within each block and $p = 0.005$ between blocks; (C) an "interface" graph with two blocks of 100 nodes each, and $p = 0.02$ within each block, but where only 10 nodes in each block are allowed to be in contact with nodes of the other block, with $p = 0.25$; (D) a graph corresponding to the contact map from the experimental HK-RR complex structure in Ref. [76] with threshold at 4 Å between closest atoms.

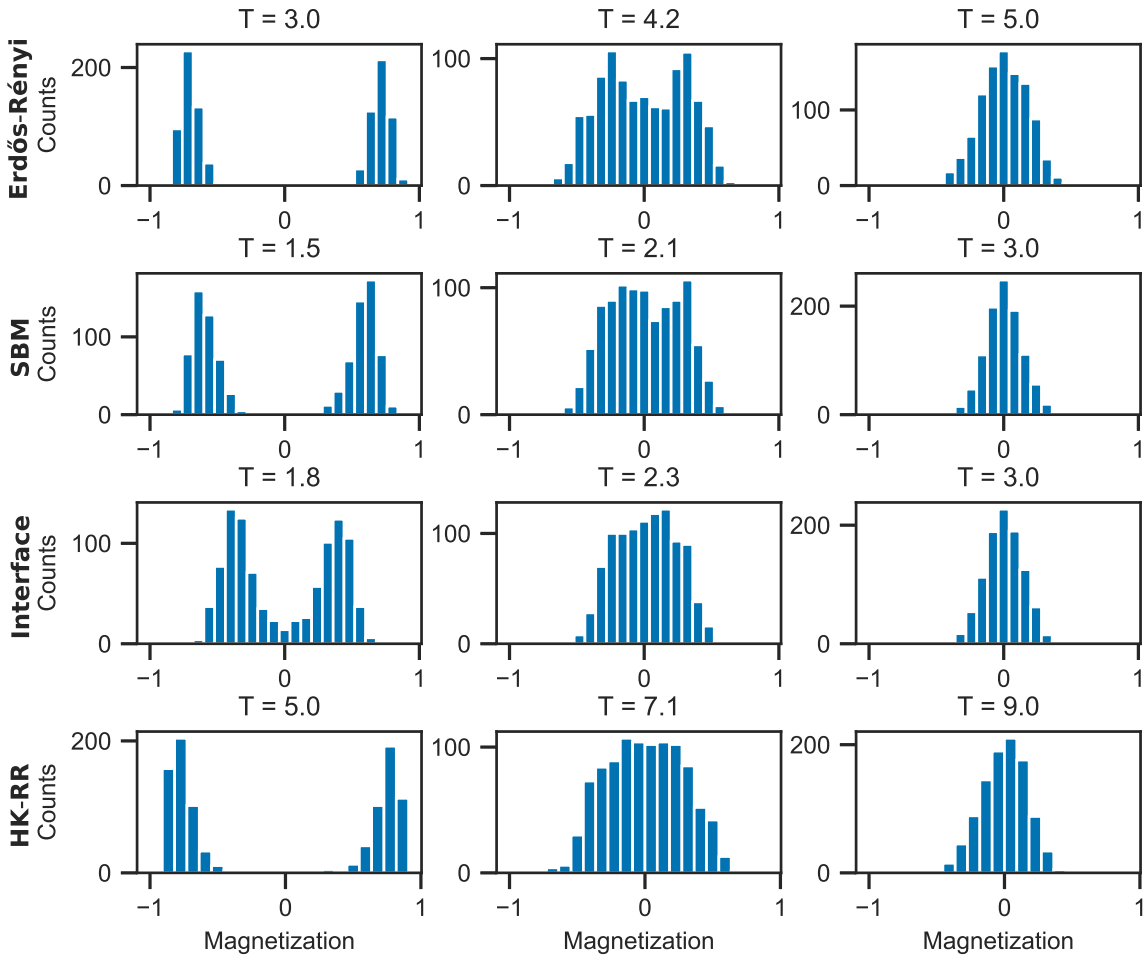

Fig S3. Ferromagnetic-paramagnetic phase transitions with different graphs. Histograms of the mean magnetization $m = (\sum_{i=1}^{2L} \alpha_i)/(2L)$ of a paired chain are shown at different sampling temperatures T for each graph considered in Fig. S6. Histograms are computed on 5000 independent sequences generated by Metropolis Monte Carlo sampling under the Hamiltonian 2 (without phylogeny). In practice, data is taken after 10,000 accepted mutations (20,000 for the HK-RR graph), see Fig. S2. The graphs considered are: (A) the same Erdős-Rényi graph with $p = 0.02$ as in Fig. 2; (B) a stochastic block model graph with two blocks of 100 nodes each, and $p = 0.02$ within each block and $p = 0.005$ between blocks; (C) an "interface" graph with two blocks of 100 nodes each, and $p = 0.02$ within each block, but where only 10 nodes in each block are allowed to be in contact with nodes of the other block, with $p = 0.25$; (D) a graph corresponding to the contact map from the experimental HK-RR complex structure in Ref. [76] with threshold at 4 Å between closest atoms. These graphs are exactly the same as in Fig. S2.

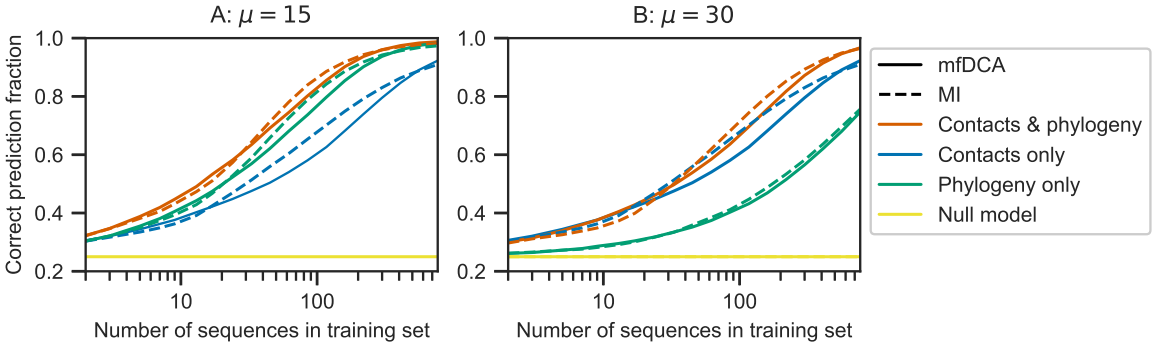

Fig S4. Impact of training set size on partner inference performance. The fraction of correctly predicted partner pairs is shown versus the number of sequence pairs AB in the training set for the minimal model incorporating both constraints from contacts and phylogeny, exactly as in Fig. 2. Specifically, an ancestral chain AB of $2L = 200$ spins is evolved along a binary branching tree (see Fig. 1) with $n = 10$ generations, and $\mu = 15$ (panel A) or $\mu = 30$ (panel B) mutations per branch, yielding $2^{10} = 1024$ pairs AB, at sampling temperature $T = 5$ under the Hamiltonian 2 on the same graph as in Fig. 2. The limiting cases with only contacts and only phylogeny are also shown for comparison. The first one corresponds to independent equilibrium sequences at $T = 5$ under the Hamiltonian 2 on the same graph. The second one corresponds to neutral evolution on the same binary branching tree. Each of these three datasets is randomly split into a training set comprising a variable number of chains AB and a testing set. The latter is randomly divided in species of 4 chains AB each, and pairings between A and B chains are blinded in each species. Partnerships are then predicted using the scores in Eq. 5 (mfDCA) or Eq. 7 (MI). Each point is averaged over 20 generated data sets, and 20 random choices of the training set for each of them. The null model shows the expectation of the correct prediction fraction if pairs are made randomly within each species.

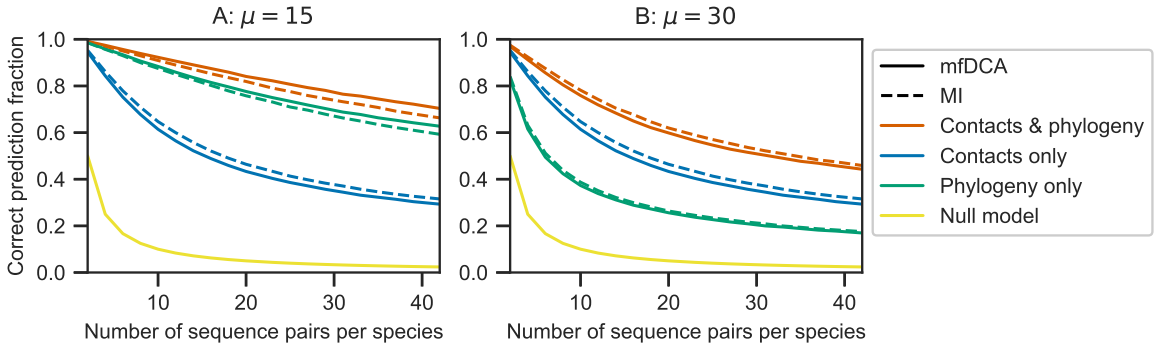

Fig S5. Impact of the number of pairs per species on partner inference performance. The fraction of correctly predicted partner pairs is shown versus the number m of sequence pairs AB per species in the testing set for the minimal model incorporating both constraints from contacts and phylogeny, exactly as in Fig. 2. Specifically, an ancestral chain AB of $2L = 200$ spins is evolved along a binary branching tree (see Fig. 1) with $n = 10$ generations, and $\mu = 15$ (panel A) or $\mu = 30$ (panel B) mutations per branch, yielding $2^{10} = 1024$ pairs AB, at sampling temperature $T = 5$ under the Hamiltonian 2 on the same graph as in Fig. 2. The limiting cases with only contacts and only phylogeny are also shown for comparison. The first one corresponds to independent equilibrium sequences at $T = 5$ under the Hamiltonian 2 on the same graph. The second one corresponds to neutral evolution on the same binary branching tree. Each of these three datasets is randomly split into a training set of 400 chains AB and a testing set of 624 chains AB. The latter are randomly divided in species of m chains AB each (here m is varied) and pairings between A and B chains are blinded in each species. Partnerships are then predicted using the scores in Eq. 5 (mfDCA) or Eq. 7 (MI). Each point is averaged over 20 generated data sets, and 20 random choices of the training set for each of them. The null model shows the expectation of the correct prediction fraction if pairs are made randomly within each species.

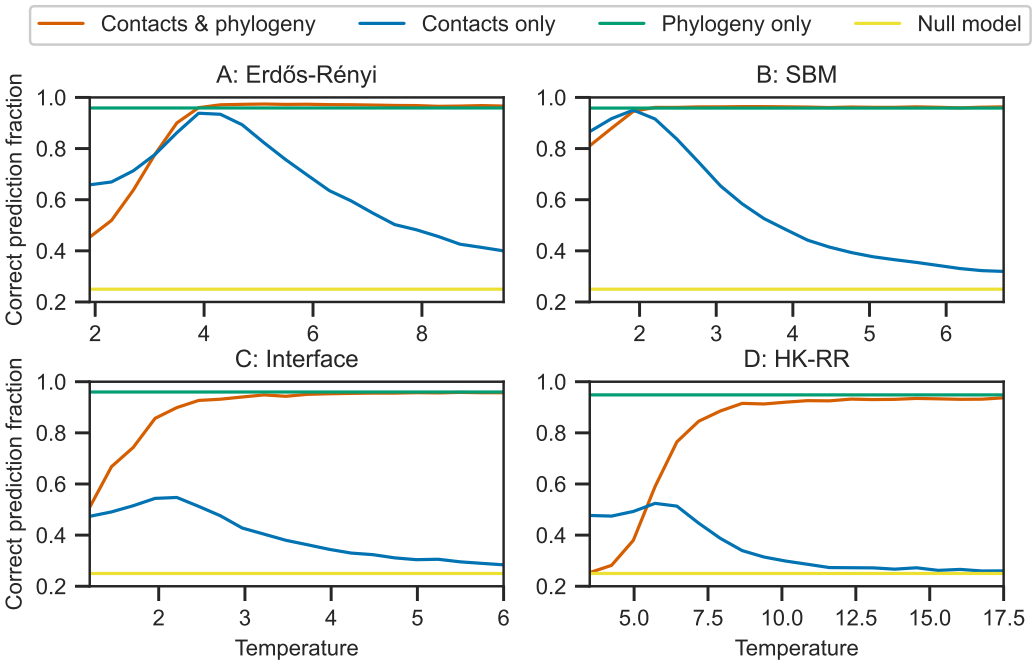

Fig S6. Partner inference performance for different graphs defining contacts. For four different graphs (panels A to D), the fraction of correctly predicted partner pairs is shown versus the sampling temperature T for the minimal model incorporating both constraints from contacts and phylogeny. Specifically, an ancestral chain AB of $2L = 200$ spins is evolved along a binary branching tree (see Fig. 1) with $n = 10$ generations, yielding $2^{10} = 1024$ pairs AB, and $\mu = 15$ mutations per branch, under the Hamiltonian 2 on (A) the same Erdős-Rényi graph with $p = 0.02$ as in Figs. 2 and 3; (B) a stochastic block model graph with two blocks of 100 nodes each, and $p = 0.02$ within each block and $p = 0.005$ between blocks; (C) an "interface" graph with two blocks of 100 nodes each, and $p = 0.02$ within each block, but where only 10 nodes in each block are allowed to be in contact with nodes of the other block, with $p = 0.25$; (D) a graph corresponding to the contact map from the experimental HK-RR complex structure in Ref. [76] with threshold at 4 Å between closest atoms. These graphs are exactly the same as in Figs. S2 and S3. For each graph, the limiting cases with only contacts and only phylogeny are also shown for comparison. The first one corresponds to independent equilibrium sequences under the Hamiltonian 2 on the same graph. The second one corresponds to neutral evolution on the same binary branching tree. For each graph, each of these three datasets is randomly split into a training set of 400 chains AB and a testing set of 624 chains AB. The latter are randomly divided in 156 species of 4 chains AB each, and pairings between A and B chains are blinded in each species. Partnerships are then predicted using the score in Eq. 5. Each point is averaged over 10 data sets. The null model shows the expectation of the correct prediction fraction if pairs are made randomly within each species. Note that, for sampling, equilibrium is considered reached for 10,000 accepted mutations, except for HK-RR graph where we perform 20,000 accepted mutations (see Fig. S1).

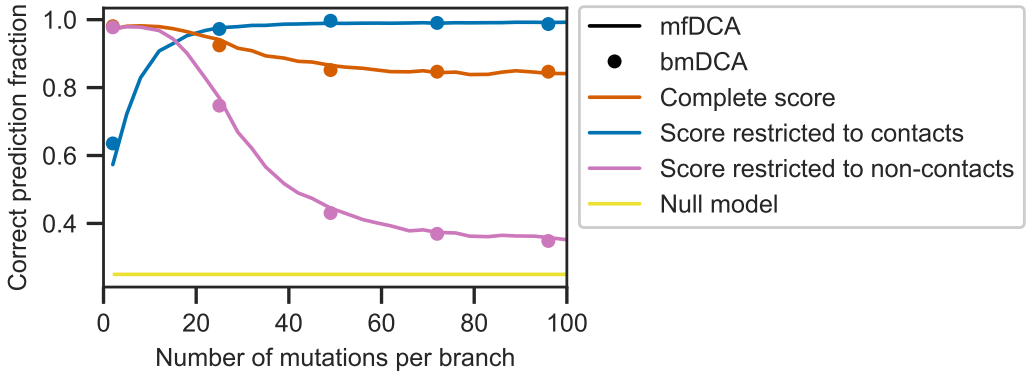

Fig S7. Impact of contact and non-contact pairs of sites on partner inference performance using mfDCA or bmDCA. The fraction of correctly predicted partner pairs is shown versus the number μ of mutations per branch of the tree for the minimal model incorporating constraints both from contacts and from phylogeny, either with the full score defined in Eq. 5 with couplings inferred using the training set, or with this score restricted to the pairs of sites that are actually in contact (Eq. 8), or to those that are not in contact (Eq. 9). Scores are computed using couplings inferred either by mfDCA (as in the rest of this work) or by bmDCA. Note that bmDCA is employed with reweigthing $\theta = 0$ here, as is done throughout for mfDCA. Apart from these points, data generation and inference are performed as in Fig. 2, using the same parameters and the same graph for contacts.

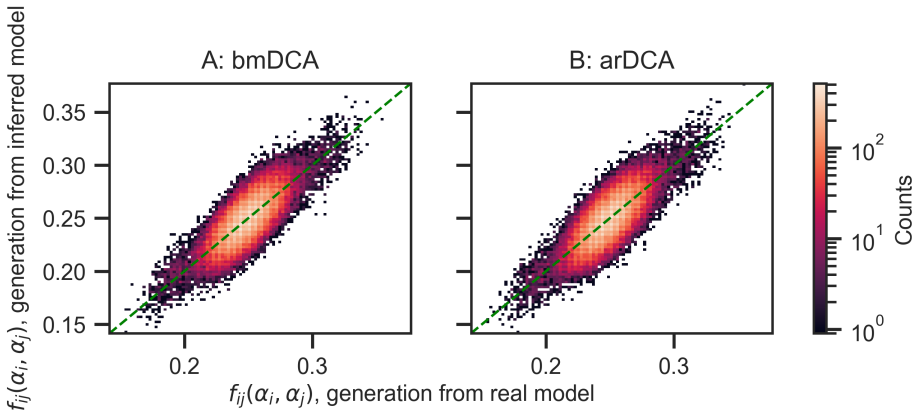

Fig S8. Reproduction of two-body frequencies by generative models, for the contact-only minimal model. An initial data set ("data set 1") of 1024 sequences was generated from the contact-only minimal model (Hamiltonian in Eq. 2 on the same Erdős-Rényi graph as in Figs. 2 and 3) at sampling temperature $T = 5$. Generative models were inferred from data set 1 by bmDCA (panel A) or arDCA (panel B). A new data set ("data set 2") of 1024 sequences was generated, still without phylogeny, from the resulting inferred model. Two-body frequencies in data set 2 are shown versus two-body frequencies in data set 1. The color of each marker represents the number of times it is observed (counts). The green dashed line shows the $y = x$ diagonal. Pearson correlation coefficients are 0.88 and 0.87 for panels A and B, respectively, and linear fits yield intercepts of 0.01 and 0 and slopes of 0.98 and 0.98 for panels A and B, respectively. The two-body frequencies shown for data set 1 include a reweighting of close sequences with Hamming distances under 0.2 [6, 7], since bmDCA and arDCA aim to match these reweighted frequencies [63, 64].

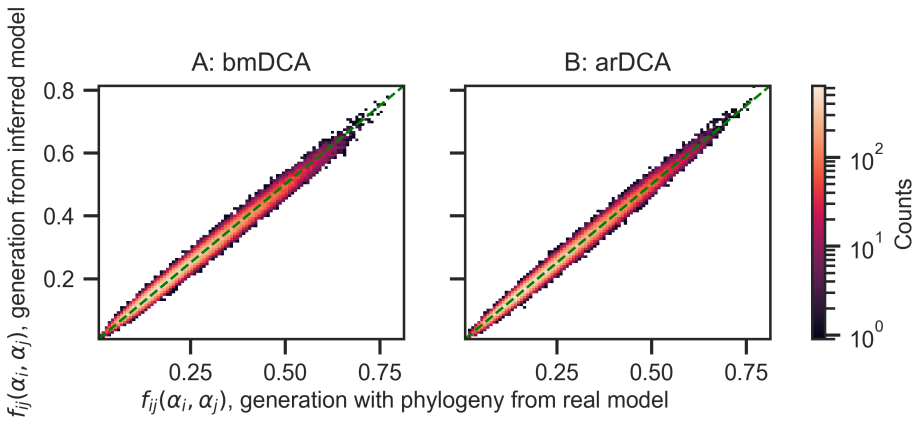

Fig S9. Reproduction of two-body frequencies by generative models, for the minimal model with contacts and phylogeny. An initial data set ("data set 3") of 1024 sequences was generated with phylogeny (binary branching tree with $\mu = 15$ mutations per branch, and sampling temperature $T = 5$) from the minimal model (Hamiltonian in Eq. 2 on the same Erdős-Rényi graph as in Figs. 2 and 3). Generative models were inferred from data set 3 by bmDCA [63] (panel A) or arDCA [64] (panel B). A new data set ("data set 4") of 1024 sequences was generated without phylogeny from the resulting inferred model. Two-body frequencies in data set 4 are shown versus two-body frequencies in data set 3. The color of each marker represents the number of times it is observed (counts). The green dashed line shows the $y = x$ diagonal. Pearson correlation coefficients are 0.994 and 0.995 for panels A and B, respectively, and linear fits yield intercepts of 0.01 and 0 and slopes of 0.97 and 1 for panels A and B, respectively. The two-body frequencies shown for data set 1 include a reweighting of close sequences with Hamming distances under 0.2 [6, 7], since bmDCA and arDCA aim to match these reweighted frequencies [63, 64].

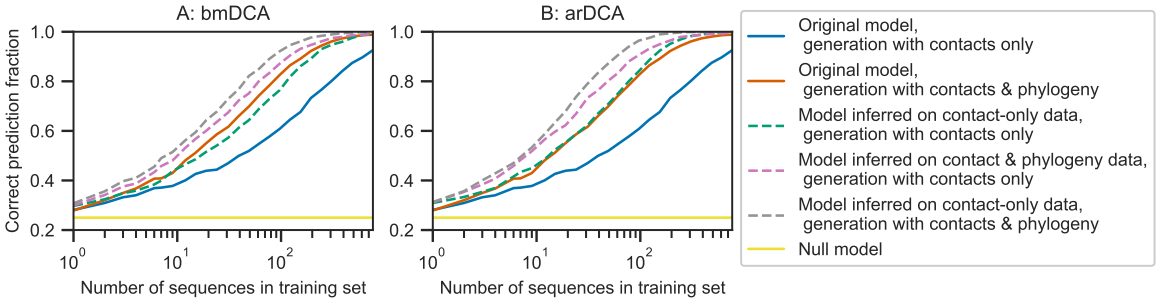

Fig S10. Partner inference performance for data generated with original and inferred models. The fraction of correctly predicted partner pairs is shown versus the number of sequence pairs AB in the training set for data generated under the minimal model, and for data generated using models inferred from this generated data, either by bmDCA [63] (panel A) or by arDCA [64] (panel B). Data generation is performed with or without phylogeny. Data generation and partner inference are performed in each case with a variable training set size, exactly as in Fig. S4. The data sets generated from the original minimal model correspond to data set 1 described in Fig. S8 (generated with contacts only) and data set 3 described in Fig. S9 (generated with contacts and phylogeny). The data sets generated from inferred models correspond to data sets 2 and 4 described in those figures, and to a data set generated with phylogeny from the model inferred from data set 1. Generation is performed at sampling temperature $T = 5$ and with $\mu = 15$ mutations per branch in cases with phylogeny. The training set is randomly divided in species of 4 chains AB each. Partnerships are predicted using the score in Eq. 5 based on mfDCA. For each point, the inference is performed over 20 generated data sets, and 20 random choices of the training set for each of them.

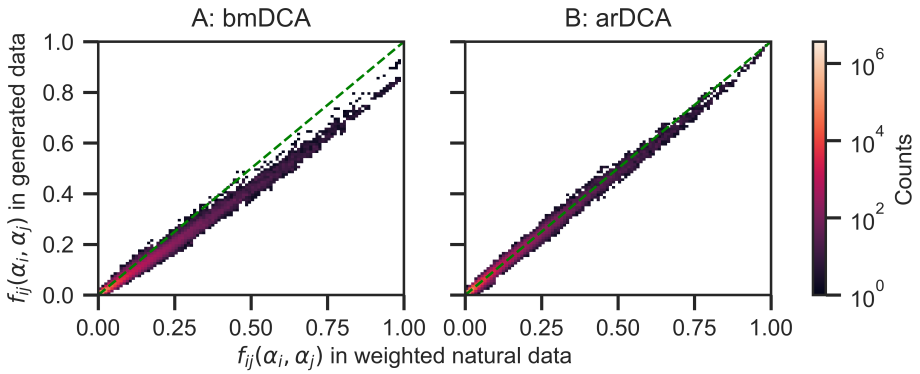

Fig S11. Reproduction of two-body frequencies by generative models. Two body frequencies in a data set of 23,633 pairs of sequences generated without phylogeny are shown versus the two body frequencies of the natural data set of 23,633 HK-RR pairs. The color of each marker represents the number of times it is observed (counts). The green dashed line shows the $y = x$ diagonal. Data is generated by bmDCA (panel A) or arDCA (panel B). The associated Pearson correlation coefficients are 0.993 and 0.996 for bmDCA and arDCA, respectively.

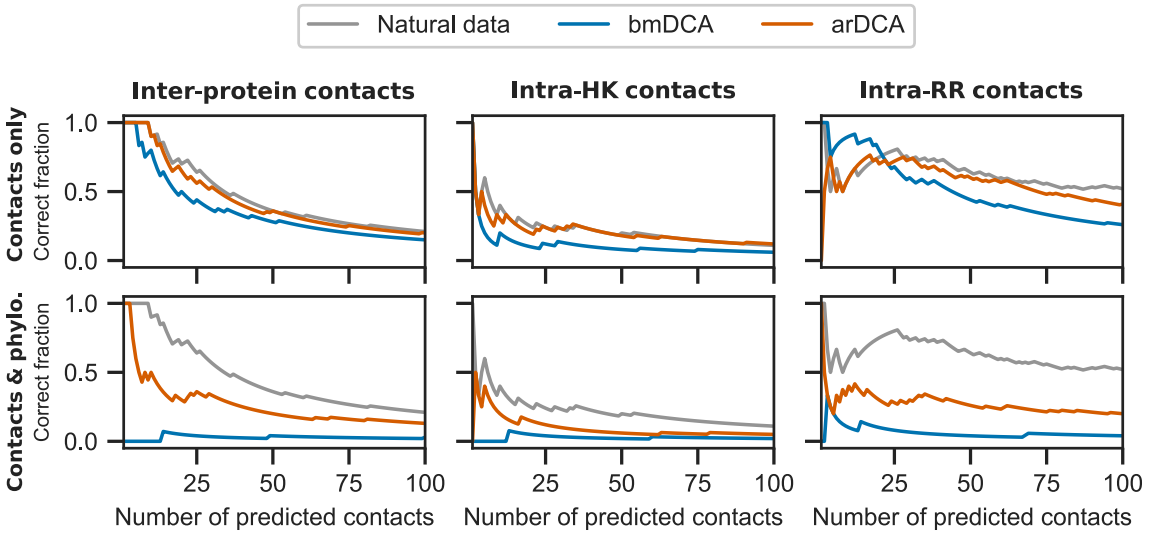

Fig S12. Inference of contacts with data generated from models inferred from natural data. The fraction of correctly predicted contacts is shown versus the number of predictions made. A contact between two amino acids is predicted based on the APC-corrected Frobenius norm [53, 54] of the couplings between amino acids inferred by mfDCA: these scores are ranked from highest to lowest, and the experimental HK-RR complex structure from Ref. [76] is employed to assess whether the contact exists or not in the real protein complex, using a threshold of 4 Å between closest atoms. Inter-protein contacts (panels A and D) are distinguished from intra-protein contacts (HK, panels B and E; RR, panels C and F). Inference of contacts is performed on the natural HK-RR dataset and on data generated from the model inferred on this natural dataset, either by bmDCA or by arDCA. Data is generated either with contacts only or with contacts and phylogeny.