Inferring interaction partners from protein sequences - Supporting Information Appendix

Anne-Florence Bitbol[a,b,c,1], **Robert S. Dwyer**[d], **Lucy J. Colwell**[e,1,2], **and Ned S. Wingreen**[a,d,1,2]

[a]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA; [b]Department of Physics, Princeton University, Princeton, NJ 08544, USA; [c]Sorbonne Universités, Université Pierre et Marie Curie - Paris 6, CNRS, Laboratoire Jean Perrin (UMR 8237), F-75005, Paris, France; [d]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA; [e]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

EXTENDED MATERIALS AND METHODS

1. Dataset construction

Complete HK-RR dataset. Our dataset is built using the online database P2CS (`http://www.p2cs.org/`) [1, 2], which includes two-component-system proteins from all fully-sequenced prokaryotic genomes. In the construction of P2CS, these proteins were identified by searching genomes for two-component system domains from the Pfam (`http://pfam.xfam.org/`) and SMART (`http://smart.embl-heidelberg.de/`) libraries. We kept only chromosome-encoded proteins, due to strong variability in plasmid presence. We also excluded hybrid and unorthodox proteins, which involve both HK and RR domains in the same protein, since the energetics of partnering is different and often less constraining for such proteins [3]. In HKs, there are different domain variants in the vicinity of the N-terminal Histidine-containing phosphoacceptor site, including the region that interacts with RRs. These variants are classified into several different Pfam domain families, which are all members of the His_Kinase_A domain clan (CL0025). In order to reliably align all HK sequences, we chose to focus on only one of these Pfam domain families, HisKA (PF00512). Proteins containing a HisKA domain account for the majority (64%) of all chromosome-encoded, non-hybrid, orthodox HKs in P2CS.

Proteins in P2CS are annotated based on genetic organization [2]. As our aim was to benchmark our method on known, specific interaction partners, we only considered HKs and RRs that are encoded by adjacent genes. Note that 67% of all chromosome-encoded, non-hybrid, orthodox HKs in P2CS are from such pairs. Suppressing the (rare) HKs with multiple HisKA domains and RRs with multiple Response_reg domains for which the pairing of domains is ambiguous, this yields 23,632 distinct pairs that differ in either sequence or species. Discarding the 208 pairs from species with only one such pair (see discussion below) yields a dataset of 23,424 HK-RR pairs. Grouping together sequences with mean Hamming distance per site < 0.3 (i.e. with 70% sequence identity or more) to estimate sequence diversity yields an effective number of HK-RR pairs $M_{\text{eff}} = 5391$ in the complete dataset.

These 23,424 HK-RR pairs are from 2102 different species, with numbers of pairs per species ranging from 2 to 41, with mean $\langle m_p \rangle = 11.1$. The distribution of the number of pairs per species in our complete dataset is shown in Fig. S6A.

Standard HK-RR dataset. In most of our work, we focused on a smaller "standard dataset" extracted from this complete dataset, both because protein families that possess as many members as the HKs and RRs are atypical, and in view of computational time constraints. Note, however, that our IPA was used to make predictions on the complete dataset, yielding a striking 0.93 final TP fraction (Fig. S7).

Our standard dataset was constructed by picking species randomly. Once 43 species with one single pair are suppressed (see discussion below), it comprises 5064 pairs from 459 species, with an average number of pairs per species $\langle m_p \rangle = 11.0$, which is very close to that of the complete dataset (see Fig. S6A for the distributions of the number of pairs per species). Grouping together sequences with mean Hamming distance per site < 0.3 to estimate sequence diversity yields an effective number of HK-RR pairs $M_{\text{eff}} = 2091$ in the standard dataset.

Suppressing species with a single pair. In our datasets, we discarded sequences from species that contain only one known pair, for which pairing is therefore unambiguous. This allowed us to quantitatively assess the impact of training set size (N_{start}) without the inclusion of an implicit training set via these pairs. More importantly, this enabled us to address prediction in the absence of any known pairs (no training set), which is crucial for predicting unknown protein-protein interactions between protein families, since no training set is then available. For other purposes, pairs from species with only one known pair might be included as a training set (but then one would need to be sure that they are actually interacting, because any error in the training set would be detrimental for the model). In our standard HK-RR dataset, if the 43 pairs from species with a single pair are treated as a training set instead of being discarded, the IPA yields a final TP fraction of 0.88 (vs. 0.84 starting from random pairings, i.e. in the absence of any training set). This value is the same as the one obtained for $N_{\text{start}} = 50$ (0.88, value averaged over 50 different random choices of the 50 training pairs, see Fig. 2). Interestingly, by exploiting multiple random initializations, a TP fraction of 0.89 is reached starting from random pairings (Fig. S8).

Multiple sequence alignment of HKs and RRs. All HKs in our dataset were aligned to the profile hidden Markov model (HMM) representing the Pfam HisKA domain (PF00512) using the `hmmalign` tool from the HMMER suite (`http://hmmer.org/`). Similarly, all RRs were aligned to the profile HMM representing the Pfam Response_reg domain (PF00072). The aligned sequences of each HK were then concatenated to those of their RR partner, yielding a concatenated multiple sequence alignment. The length of each concatenated sequence is $L = 176$ amino acids, among which

[1] To whom correspondence should be addressed. E-mails: anne-florence.bitbol@upmc.fr (A.F.B.), ljc37@cam.ac.uk (L.J.C.), wingreen@princeton.edu (N.S.W.)

[2] L.J.C. and N.S.W. contributed equally to this work.

the $L_{\mathrm{HK}} = 64$ first amino acids are from the HK, and the remaining 112 amino acids are from the RR. The full length of these sequences was kept throughout.

Dataset construction for ABC transporter proteins. While we used HK-RRs as the main benchmark for the IPA, we also applied it to several pairs of protein families involved in ABC (ATP-binding cassette) transporter complexes. These ubiquitous complexes enable ATP-powered translocation of various substances through membranes [4]. As in the case of HK-RRs, bacterial genomes typically contain multiple paralogs of these transporters, and actual pairings are known from genome proximity, enabling us to assess the success of the IPA.

We built paired alignments of homologs of the *Escherichia coli* interacting protein pairs MALG-MALK, FBPB-FBPC, and GSIC-GSID, all involved in ABC transporter complexes, using a method adapted from Ref. [5] and `http://gremlin.bakerlab.org/`. First, the homologs of each protein were retrieved from Uniprot (`http://www.uniprot.org/`) using `hhblits` from the HH-suite (`https://github.com/soedinglab/hh-suite`) with main options `-n 8 -e 1E-20`. Then `hhfilter` from the HH-suite was run with options `-id 100 -cov 75` to only retain the homologs that have at least 75% coverage. In order to focus on the relevant conserved domains involved in binding, as we did for HK-RRs, we then used `hmmsearch` from the HMMER suite to align a subsequence of each homolog to the profile HMM of the appropriate domain from Pfam. These domains are ABC_tran (PF00005) for MALK, and BPD_transp_1 (PF00528) for all other ABC transporter proteins considered here. For each pair of interacting protein families, sequences from the same species (found via the OX/OS field in the Uniprot headers) were then paired to their interacting partner by genome proximity (assessed via the Uniprot accession numbers, and using a maximum allowed difference of 20 between these IDs). These pairings enabled us to evaluate IPA performance (Fig. 5), as in the HK-RR case. Note that the paired alignment of HK-RRs homologous to BASS-BASR was constructed in the same way as the alignments of these ABC-transporter protein pairs.

We also considered a pair of protein families with no known interactions: BASR homologs (Response_reg domain) and MALK homologs (ABC_tran domain). These two protein families have very different biological functions, and no interaction between BASR and MALK has been reported in the STRING database (`http://string-db.org/`).

As in the case of HK-RRs, for each pair of protein families, we worked on subsets of ~ 5000 protein pairs extracted from the complete dataset by randomly picking species, and we discarded species with a single pair.

2. Statistics of the concatenated alignment (CA)

Henceforth, as in the main text, we will present our general method in the specific case of HK-RRs. Note that we applied it in the exact same way to ABC transporter protein pairs.

Let us consider a CA of paired HK-RR sequences. At each site $i \in \{1, .., L\}$, where L is the number of amino-acid sites, a given concatenated sequence can feature any amino acid (denoted by α with $\alpha \in \{1, .., 20\}$), or a gap (denoted by $\alpha = 21$), yielding 21 possible states α for each site i.

To describe the statistics of the alignment, we only employ the single-site frequencies of occurrence of each state α at each site i, denoted by $f_i^e(\alpha)$, and the two-site frequencies of occurrence of each ordered pair of states (α, β) at each ordered pair of sites (i, j), denoted by $f_{ij}^e(\alpha, \beta)$ [6]. The raw empirical frequencies, obtained by counting the sequences where given residues occur at given sites and dividing by the number M of sequences in the CA, are subject to sampling bias, due to phylogeny and to the choice of species that are sequenced [7, 8]. Hence, to define f_i^e and f_{ij}^e, we use a standard correction that re-weights "neighboring" concatenated sequences with mean Hamming distance per site < 0.3. The value of this similarity threshold is arbitrary, but our results depend very weakly on this choice, even when taking the threshold down to zero. The weight associated to a given concatenated sequence a is $1/m_a$, where m_a is the number of neighbors of a within the threshold [7–9]. This allows one to define an effective sequence number M_{eff} via

$$M_{\mathrm{eff}} = \sum_{a=1}^{M} \frac{1}{m_a} \,. \qquad \text{[S1]}$$

To avoid issues such as amino acids that never appear at some sites, which would present mathematical difficulties, e.g. a non-invertible correlation matrix and diverging couplings [7], we introduce pseudocounts via a parameter Λ [6–9]. The one-site frequencies f_i become

$$f_i(\alpha) = \frac{\Lambda}{q} + (1 - \Lambda) f_i^e(\alpha) \,, \qquad \text{[S2]}$$

where $q = 21$ is the number of states (i.e. of amino acids, including gaps) per site. Similarly, the two-site frequencies f_{ij} become

$$f_{ij}(\alpha, \beta) = \frac{\Lambda}{q^2} + (1 - \Lambda) f_{ij}^e(\alpha, \beta) \text{ if } i \neq j \,, \qquad \text{[S3]}$$

$$f_{ii}(\alpha, \beta) = \frac{\Lambda}{q} \delta_{\alpha\beta} + (1 - \Lambda) f_{ii}^e(\alpha, \beta) = f_i(\alpha) \delta_{\alpha\beta} \,, \qquad \text{[S4]}$$

where $\delta_{\alpha\beta} = 1$ if $\alpha = \beta$ and 0 otherwise. These pseudocount corrections are uniform (i.e. they have the same weight $1/q$ on all amino-acid states), and their importance relative to the raw empirical frequencies can be tuned through the parameter Λ. In practice, we take $\Lambda = 0.5$, which has been shown to be a satisfactory choice [7, 8]. Note that the correspondence of Λ with the parameter λ in Refs. [7–9] is obtained by setting $\Lambda = \lambda / (\lambda + M_{\mathrm{eff}})$.

From these quantities, we define the two-point correlations

$$C_{ij}(\alpha, \beta) = f_{ij}(\alpha, \beta) - f_i(\alpha) f_j(\beta) \,. \qquad \text{[S5]}$$

3. Maximum entropy model

Formulation. The maximum entropy principle [10] yields the following form for the least-structured global (L-point) probability distribution P of sequences consistent with the empirical one- and two-point statistics of the CA:

$$P(\alpha_1, ..., \alpha_L) = \frac{1}{Z} \exp\left\{ - \left[\sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j) \right] \right\} \,, \qquad \text{[S6]}$$

where Z is a normalization constant. Each one-body term h_i is known as the field at site i, and each two-body interaction term e_{ij} is known as the (direct) coupling between sites i and j. The fields h_i and the couplings e_{ij} are determined by imposing that the probability distribution P be consistent with the empirical one- and two-point frequencies f_i and f_{ij}:

$$\sum_{\alpha_k, k \neq i} P(\alpha_1, ..., \alpha_L) = f_i(\alpha_i) \,, \qquad \text{[S7]}$$

$$\sum_{\alpha_k, k \notin \{i,j\}} P(\alpha_1, ..., \alpha_L) = f_{ij}(\alpha_i, \alpha_j) \,. \qquad \text{[S8]}$$

Such pairwise interaction maximum entropy models have proved very successful in various fields (see e.g. Refs. [11–19]), including the prediction of protein structures and inter-protein contacts from multiple sequence alignments (see e.g. Refs. [6–8]). In particular, high couplings e_{ij} are better predictors of real contacts in proteins than high correlations C_{ij}, because the e_{ij} represent minimal direct couplings between amino acids, while high C_{ij} can arise from indirect effects [6–8].

Inference of the parameters. Eqs. S7 and S8 alone do not uniquely define all the fields $h_i(\alpha)$ and couplings $e_{ij}(\alpha, \beta)$ with $1 \leq i < j \leq L$ involved in Eq. S6, which amount to $Lq + L(L-1)q^2/2$ parameters, where $q = 21$ is the number of amino-acid states α. Indeed, while the number of equations in Eqs. S7 and S8 is the same as that of the empirical frequencies, the latter are not all independent. The two-site frequencies are symmetric ($f_{ij}(\alpha, \beta) = f_{ji}(\beta, \alpha)$) and consistent with the one-site frequencies ($f_{ii}(\alpha, \beta) = f_i(\alpha)\delta_{\alpha\beta}$; $\sum_\beta f_{ij}(\alpha, \beta) = f_i(\alpha)$; and $\sum_\alpha f_{ij}(\alpha, \beta) = f_j(\beta)$), which sum to one ($\sum_\alpha f_i(\alpha) = 1$). All these constraints reduce the number of independent variables among the one- and two-site frequencies, and thus of independent equations, to $L(q-1) + L(L-1)(q-1)^2/2$ [6, 7]. This yields a degree of freedom in the determination of the fields and couplings from Eqs. S7 and S8. Given the number of independent equations, one possible gauge choice is to set to zero the fields and couplings for one given state, e.g. state q (the gap) [7, 8]: $h_i(q) = 0$ and, for all α,

$$e_{ij}(\alpha, q) = e_{ij}(q, \alpha) = 0 \,. \qquad \text{[S9]}$$

Determining the remaining fields h_i and the couplings e_{ij} from Eqs. S7 and S8 is difficult, and various approximations have been developed to solve this problem. Following Refs [7, 8], we use the mean-field or small-coupling approximation, which was introduced in Ref. [20] for the Ising spin-glass model. In this approximation, for $i \neq j$ and $\alpha, \beta < q$, the couplings are given by $e_{ij}(\alpha, \beta) = A^{-1}_{kl}$, where A is a $(q-1)L \times (q-1)L$ correlation matrix: $A_{kl} = C_{ij}(\alpha, \beta)$, where $k = (q-1)(i-1)+\alpha$ and $l = (q-1)(j-1) + \beta$ [21]. This can be summarized as

$$e_{ij}(\alpha, \beta) = C^{-1}_{ij}(\alpha, \beta) \,. \qquad \text{[S10]}$$

Together, Eqs. S9 and S10 yield all the couplings. Note that the couplings are symmetric ($e_{ij}(\alpha, \beta) = e_{ji}(\beta, \alpha)$) since the correlations are.

This simple mean-field approximation has been used with success for protein structure prediction [7, 8]. (More sophisticated approximations typically improve performance by less than ten percent [21, 22].) Moreover, this approximation is computationally fast, since it only requires the inversion of a $(20L) \times (20L)$ correlation matrix. Computational rapidity is a considerable asset for our purpose, given that the IPA performs better with smaller increment step size $N_{\text{increment}}$ (see Fig. 3), i.e. with more iterations, and that the couplings e_{ij} are computed at each iteration. This approximation also enabled us to use the full-length sequences of domains to infer couplings, without needing to restrict to a subset of amino-acid sites as in some other works using more sophisticated approximations [6, 9]. We find that using full-length sequences increases the resulting TP fraction.

Gauge choice. Qualitatively, the gauge degree of freedom means that contributions to the effective energy of the system

$$H = \sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j) \qquad \text{[S11]}$$

can be shifted between the fields and the couplings [6]. Since our focus is on interactions, we do not want the couplings to include contributions that can be accounted for by the (one-body) fields [23]. The zero-sum (or Ising) gauge, where the couplings satisfy

$$\sum_\alpha e_{ij}(\alpha, \beta) = \sum_\beta e_{ij}(\alpha, \beta) = 0 \,, \qquad \text{[S12]}$$

minimizes the Frobenius norms of the couplings

$$\|e_{ij}\| = \sqrt{\sum_{\alpha,\beta=1}^{q} [e_{ij}(\alpha, \beta)]^2} \,. \qquad \text{[S13]}$$

Hence, the zero-sum gauge attributes the smallest possible fraction of the energy in Eq. S6 to the couplings, and the largest possible fraction to the fields [6, 21]. Furthermore, when employing this gauge, the Frobenius norm has proved to be a successful predictor of contacts in proteins [21, 22]. In particular, within the mean-field approximation Eq. S10, the use of the Frobenius norm (with an average-product correction) improves over the results obtained using direct information [21].

Thus, after calculating the couplings as described above, we change the gauge from the one defined in Eq. S9 to the one defined in Eq. S12, by replacing each coupling $e_{ij}(\alpha, \beta)$ by

$$e_{ij}(\alpha, \beta) - \langle e_{ij}(\gamma, \beta)\rangle_\gamma - \langle e_{ij}(\alpha, \delta)\rangle_\delta + \langle e_{ij}(\gamma, \delta)\rangle_{\gamma,\delta} \,, \qquad \text{[S14]}$$

where $\langle . \rangle_\gamma$ denotes an average over $\gamma \in \{1, ..., q\}$ [21].

Note that in Fig. 4, we use the Frobenius norm without the average-product correction [21]. With this correction, implemented by averaging within single proteins [5], we obtained similar results (see Fig. S13). Overall, with the correction, final performance is slightly worse, but training is visible slightly earlier in the IPA.

4. Iterative pairing algorithm (IPA)

The main steps of the IPA are shown in Fig. 1C. Here, we describe each of these steps in detail, after explaining how the CA is constructed for the very first iteration.

Initialization of the CA.

Starting from a training set of HK-RR pairs. The CA for the first iteration of the IPA is built from the pairs in the training set, which are considered as known interaction partners. In subsequent iterations, the training set pairs are *always kept* in the CA, and additional pairs with the highest confidence scores (see below) are added to the CA.

Starting from random pairings. In the absence of a training set, each HK of the dataset is randomly paired with an RR from its species. All M pairs, where M represents the total number of HKs, or, equivalently, RRs, in the dataset, are included in the CA for the first iteration of the IPA. Hence, this initial CA contains a mixture of correct and incorrect pairs, with one correct pair per species on average. At the second iteration, the CA is built using only the $N_{\mathrm{increment}}$ HK-RR pairs with the highest confidence scores obtained from this first iteration.

There are other ways to initialize the CA in the absence of a training set. We varied the number of pairs included at the second iteration ($N_{\mathrm{increment}}$ in the above scheme), and we also tried constructing the first CA from all possible HK-RR pairs from the species with few pairs (as for these species, exhaustive pairing yields a larger proportion of true pairs). These variants did not significantly increase the final TP fraction. Moreover, the random initialization of the CA can be exploited to increase the TP fraction (Figs. S10 and S9), which would be impossible for exhaustive initializations.

Now that we have described the initial construction of the CA, we describe each step of an iteration of the IPA (Fig. 1C).

Step 1: Correlations. At each iteration, the empirical one- and two-body frequencies are computed for the CA, using the re-weighting of neighbor sequences and the pseudocount correction described above (see Eqs. S1-S4). The empirical correlations C_{ij} are then deduced using Eq. S5.

Step 2: Direct couplings. The direct couplings in the pairwise maximum entropy model of the CA are inferred from the empirical correlations using Eqs. S9 and S10. The gauge is then changed to the zero-sum gauge (Eq. S12) using Eq. S14.

Step 3: Interaction energies for all possible HK-RR pairs. The interaction energy E of each possible HK-RR pair within each species of the dataset is calculated by summing the appropriate direct couplings:

$$E\left(\alpha_1, ..., \alpha_{L_{\mathrm{HK}}}, \alpha_{L_{\mathrm{HK}}+1}, ..., \alpha_L\right) = \sum_{i=1}^{L_{\mathrm{HK}}} \sum_{j=L_{\mathrm{HK}}+1}^{L} e_{ij}(\alpha_i, \alpha_j)\,, \qquad \text{[S15]}$$

where L_{HK} denotes the length (i.e. the number of amino-acid sites) of the HK sequence and L that of concatenated HK-RR sequence. Note that this HK-RR interaction energy only involves the inter-molecular couplings ($i \leq L_{\mathrm{HK}}$ and $j > L_{\mathrm{HK}}$; the case $i > L_{\mathrm{HK}}$ and $j \leq L_{\mathrm{HK}}$ does not need to be considered as the couplings are symmetric).

Step 4: HK-RR pair assignments and ranking by energy gap.

HK-RR pair assignments. In each separate species, the pair with the lowest interaction energy is selected first, and the HK and RR from this pair are removed from further consideration, since we assume one-to-one HK-RR matches (see Fig. 1D). Then, the pair with the next lowest energy is chosen, and the process is repeated until all HKs and RRs are paired.

Scoring by gap. Each assigned HK-RR pair is scored at the time of assignment by $\Delta E/(n+1)$, where ΔE is the energy gap between the match with the lowest energy and the next best one (see Fig. 1E), and n is the number of lower-energy matches discarded in assignments made previously (within that species and at that iteration). Qualitatively, the larger the energy gap, and the smaller the number n of rejected better candidates, the more reliable we expect the assignment to be.

More precisely, $\Delta E_{\mathrm{RR}} = E_{\mathrm{RR},2} - E_{\mathrm{RR},1} > 0$ is computed for the RR involved as minus the difference of the interaction energy $E_{\mathrm{RR},1}$ of this RR with its assigned partner (i.e. the "best" HK, which yields the lowest interaction energy with this RR, *among the HKs that are still unpaired*) and that $E_{\mathrm{RR},2}$ with the second-best HK *among the HKs that are still unpaired.* Meanwhile, n_{RR} is the number of HKs of that species that had lower interaction energies with this RR than the assigned partner, but that have been eliminated previously in that iteration's pairing process, because they were paired to other RRs with a lower interaction energy. A schematic example is shown on Fig. S12A. Similarly, the value of ΔE_{HK} and of n_{HK} are calculated for the HK involved in the assigned pair. Finally, the lowest score among the two obtained is kept:

$$\frac{\Delta E}{n+1} = \min\left(\frac{\Delta E_{\mathrm{RR}}}{n_{\mathrm{RR}}+1}, \frac{\Delta E_{\mathrm{HK}}}{n_{\mathrm{HK}}+1}\right). \qquad \text{[S16]}$$

We have chosen to divide the energy gap ΔE by $n+1$ in order to penalize the HK-RR pairs made after better candidates were discarded, even if their current gap among remaining candidates appears large, as illustrated by the second assignment in Fig. S12A. However, one could consider other definitions of confidence scores, such as $\Delta E/(n+1)^{\alpha}$, where α is a parameter. In Fig. S12B, we show that our confidence score significantly improves TP fraction over the raw energy gap ΔE, and that $\alpha = 1$ yields an optimal TP fraction.

This definition of the confidence score leaves an ambiguity for the last assigned pair of each species, since there is no remaining second-best match to define the energy gap. We have chosen to assign to this pair a confidence score equal to the lowest other one within the species, given that this pair, made by default, should not be deemed more reliable than any other pair in the species.

Another ambiguity exists when several pairs have exactly the same interaction energy. This mostly occurs when the model is built from one single HK-RR concatenated sequence (this case is not singular thanks to the pseudocount correction, and the model then yields a lower energy contribution for each residue pair identical to the initial concatenated sequence, and a higher energy contribution for each residue pair comprising one same and one different residue compared to the initial concatenated sequence). It also occurs in the extremely rare case where two identical HK (or RR) sequences are found in the same genome. In this case, we chose to randomly make one pair assignment between the equivalent matches, and to leave the other equal energy HKs and/or RRs to be paired later. We checked that the impact of this choice on final results is very small.

Ranking of pairs. Once all the HK-RR pairs are assigned and scored, we rank them in order of decreasing confidence score.

Step 5: Incrementation of the CA. The ranking of the HK-RR pairs is used to pick those pairs that are included in the CA at the next iteration. Pairs with a high confidence score are more likely to be correct because there was less ambiguity in the assignment. The number of pairs in the CA is increased by $N_{\mathrm{increment}}$ at each iteration, and the IPA is run until all the HKs and RRs in the dataset have been paired and added to the CA. In the last iteration, all pairs assigned at the second to last iteration are included in the CA.

Starting from a training set of HK-RR pairs. The N_{start} training pairs remain in the CA throughout and the HKs and RRs involved in these pairs are not paired or scored by the IPA. The HKs and RRs from all the other pairs in the CA are re-paired and re-scored at each iteration, and only re-enter the CA if their confidence score is sufficiently high. In other words, at the first iteration, the CA only contains the N_{start} training pairs. Then, for any iteration number $n > 1$, it contains these exact same N_{start} training pairs, plus the $(n-1)N_{\text{increment}}$ assigned HK-RR pairs that had the highest confidence scores at iteration number $n-1$.

Starting from random pairings. In the absence of a training set, all M HKs and RRs in the dataset are paired and scored at each iteration, and all the pairs of the CA are fully re-picked at each iteration based on the confidence score. The first iteration is special, since the CA is made of M random within-species HK-RR pairs (see above, "Initialization of the CA"). Then, for any iteration number $n > 1$, the CA contains the $(n-1)N_{\text{increment}}$ assigned HK-RR pairs that had the highest confidence scores at iteration number $n-1$.

Once the new CA is constructed, the iteration is completed, and the next one can start with Step 1, the computation of the empirical correlations in this CA.

Run time. The run time of the IPA strongly depends on $N_{\text{increment}}$ and on dataset size (length of concatenated sequences, number of such sequences in the dataset). For our standard HK-RR dataset, all single-processor run times for a Matlab-coded version of the IPA were shorter than one day down to $N_{\text{increment}} = 6$.

SUPPORTING FIGURES

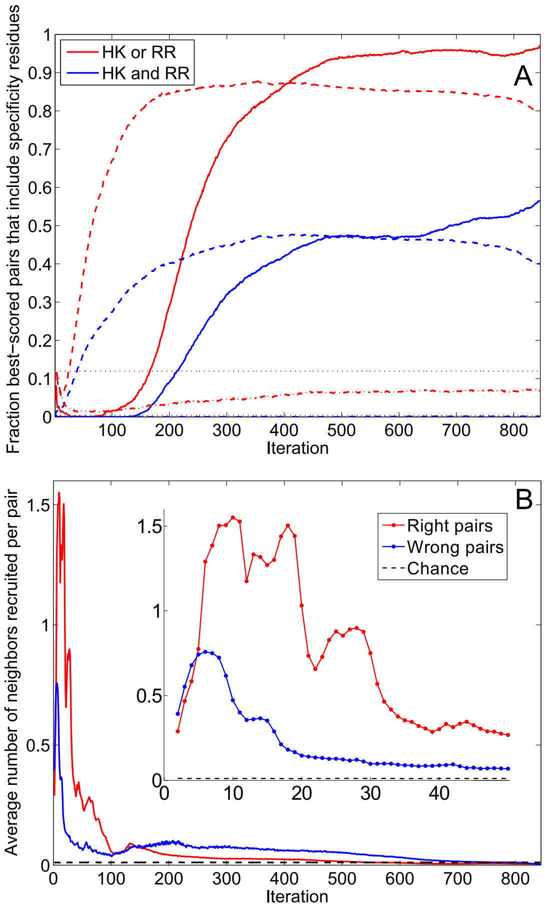

Fig. S1. Evolution of the coupling matrix and of the concatenated alignment (CA) during the IPA. (A) Training of the coupling matrix. As in Fig. 4A, pairs comprised of an HK residue site and an RR residue site are scored by the Frobenius norm (i.e. the square root of the summed squares) of the couplings involving all possible residue types at these two sites. The 10 best-scored pairs are compared to the main specificity residues determined experimentally in Refs. [24–27] (5 HK residues, T267, A268, A271, Y272, and T275 in the sequence of *T. maritima* HK853, and 5 RR residues, V13, L14, I17, N20, and F21 in the sequence of *T. maritima* RR468 [26]). Solid curves: Fraction of the 10 best-scored residue pairs that include HK and/or RR specificity residues versus the iteration number in the IPA. Dashed curves: Ideal case, where at each iteration $N_{\text{increment}}$ randomly-selected correct HK-RR pairs are added to the CA. Dash-dotted curves: Case where random HK-RR pairs are added to the CA. Dotted lines: Overall fraction of residue pairs that include specificity residues. (B) Neighbor recruitment. Average number of neighbors an HK-RR pair of the CA has among the new HK-RR pairs of the next CA versus iteration number. Two pairs are considered neighbors if the mean Hamming distance per site between the two HKs and between the two RRs are both < 0.3. Dashed line: Null model – at each iteration, $N_{\text{increment}}$ new correct HK-RR pairs are chosen at random and added to the CA. Inset: Expanded view of the first 50 iterations. In both panels, the IPA is performed on the standard dataset with $N_{\text{increment}} = 6$. In panel A (resp. B), data is averaged over 500 (resp. 5193) replicates that differ in their initial random pairings.

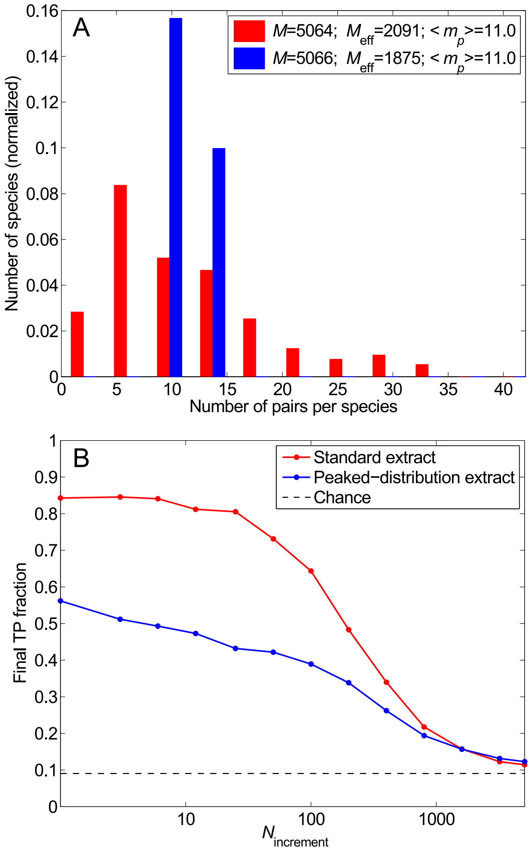

Fig. S2. Impact of the distribution of the number of HK-RR pairs per species. (A) Distribution of the number of pairs per species in two different datasets: the standard one (red) and one with the same total number of HK-RR pairs M and the same mean number of pairs per species $\langle m_p \rangle$, but with a more strongly peaked distribution (blue). (B) Final TP fraction versus $N_{\text{increment}}$ for the two datasets described in (A). All results are averaged over 50 replicates that differ in their initial random pairings. Dashed line: Average TP fraction obtained for random HK-RR pairings.

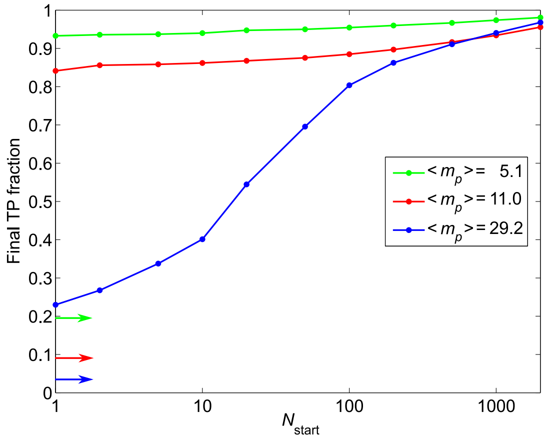

Fig. S3. Impact of the number of HK-RR pairs per species: starting from a training set. Final TP fraction versus N_{start} for the three datasets with different distributions of the number of pairs per species yielding different means $\langle m_p \rangle$ presented in Fig. 5. Colored arrows indicate the average TP fractions obtained for random HK-RR pairings in each dataset. The IPA is performed on the standard dataset with $N_{\text{increment}}$ = 6. All results are averaged over 50 replicates that differ by the random choice of pairs in the training set.

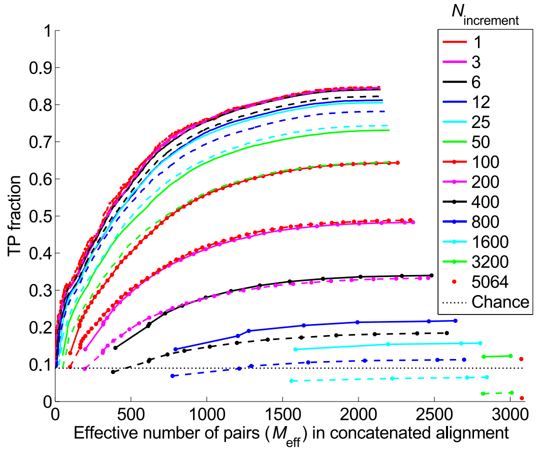

Fig. S4. Impact of the initial correct pairs. TP fraction versus effective number of HK-RR pairs (M_{eff}) in the concatenated alignment during iterations of the IPA, for different values of $N_{\text{increment}}$. Solid curves: Starting from random pairings (data also shown in Fig. 3). Dashed curves: Starting from random pairings with no initial correct pair (the color and symbol codes are the same as for the solid curves). The standard dataset is used. All results are averaged over 50 replicates that differ in their initial random pairings. Dotted line: Average TP fraction obtained for random HK-RR pairings.

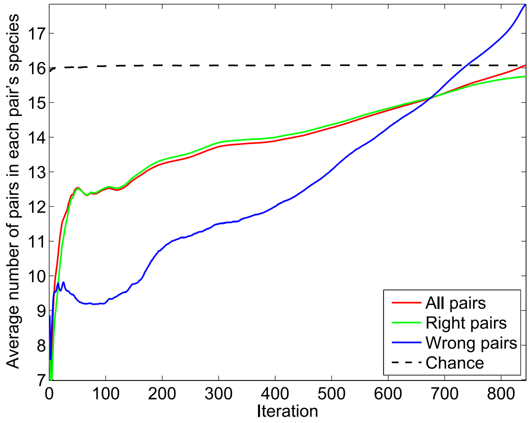

Fig. S5. Evolution of the concatenated alignment (CA) during the IPA. Average number of HK-RR pairs present in the species to which the pairs of the CA belong versus iteration number. The IPA is performed on the standard dataset, with $N_{\text{increment}}$ = 6, and all data is averaged over 5193 replicates that differ in their initial random pairings. Dashed line: At each iteration, 6 new correct HK-RR pairs are chosen at random and added to the CA. This "chance" result just matches the average number of pairs in a pair's species: 16.1. Note that this number is different from the above-discussed average number of pairs per species $\langle m_p \rangle$, which is 11.0 in the standard dataset (because the average over the pairs is not the same as the average over the species).

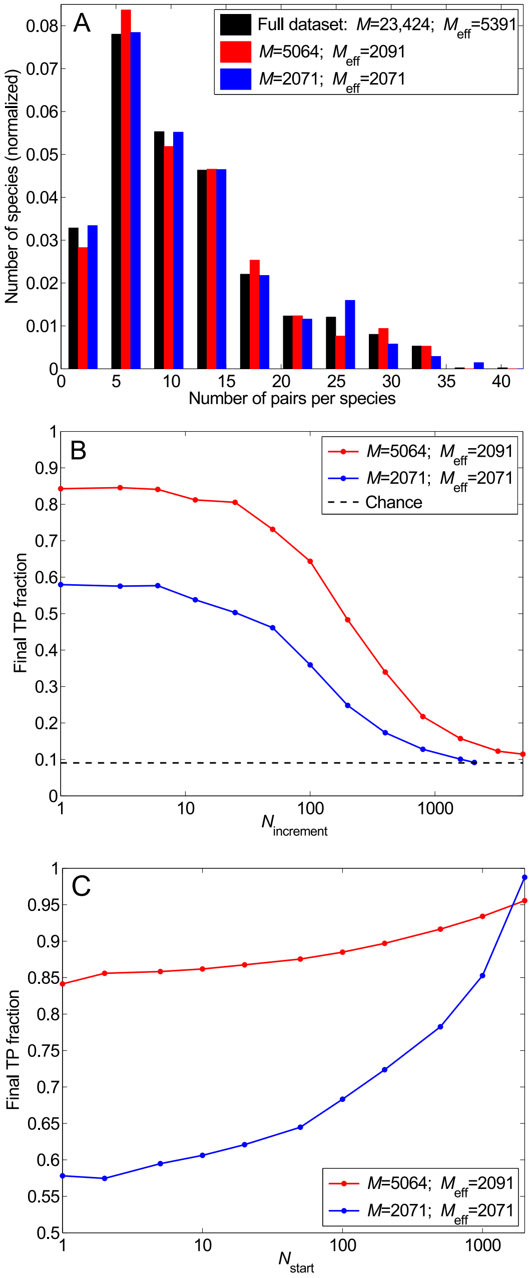

Fig. S6. Impact of sequence similarity in the dataset. (A) Distribution of the number of pairs per species in the complete dataset (black) and in two smaller selected datasets each with the same effective number of HK-RR pairs M_{eff}: the standard one (red) and one where similar sequences have been suppressed such that no two pairs have a mean Hamming distance per site < 0.3 (blue). (B) Final TP fraction versus $N_{\mathrm{increment}}$ for the two selected datasets described in (A), starting from random pairings. Dashed line: Average TP fraction obtained for random HK-RR pairings. (C) Starting from a training set. Final TP fraction versus N_{start} for the two selected datasets presented in (A), with $N_{\mathrm{increment}} = 6$. In (B) and (C), all results are averaged over 50 replicates.

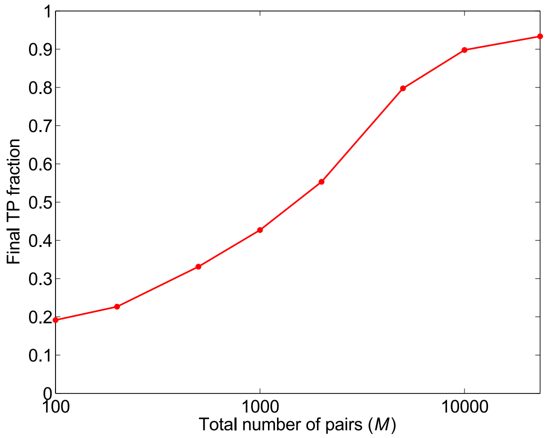

Fig. S7. Impact of the total number of HK-RR pairs in the dataset. Final TP fraction versus the total number M of HK-RR pairs in the dataset, starting from random pairings. For each M, datasets are constructed by picking species randomly from the full dataset, preserving the average distribution of the number of HK-RR pairs per species. For each M except the largest, results are averaged over multiple different such alignments (from 50 up to 500 for small M). For the largest M (full dataset), averaging is done on 50 different initial random pairings. All results correspond to the small-$N_{\text{increment}}$ limit.

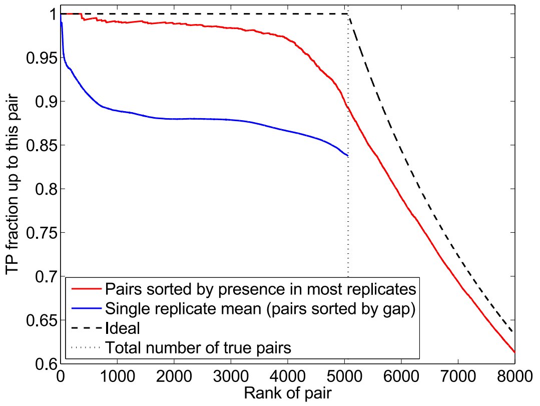

Fig. S8. Improved accuracy from multiple initial random pairings. Red curve: All possible HK-RR pairs (within each species) are ranked by the fraction f_r of replicates of the IPA in which they are predicted. The TP fraction up to each pair is plotted versus the rank of this pair. The standard dataset is used, with $N_{\text{increment}}$ = 6. 500 replicates that differ in their initial random pairings are considered. Blue curve: For each separate replicate, pairs are ranked by their confidence score, in decreasing order. The TP fraction up to each pair is computed, and the mean of these curves is shown. Dashed curve: Ideal classification, where the M = 5064 first pairs (dotted line) are correct, while all the others are incorrect. When ranking pairs by decreasing f_r (red curve), the TP fraction among the M = 5064 best-ranked pairs is 0.89, a significant improvement over the average of TP fractions from individual replicates, 0.84 (blue curve).

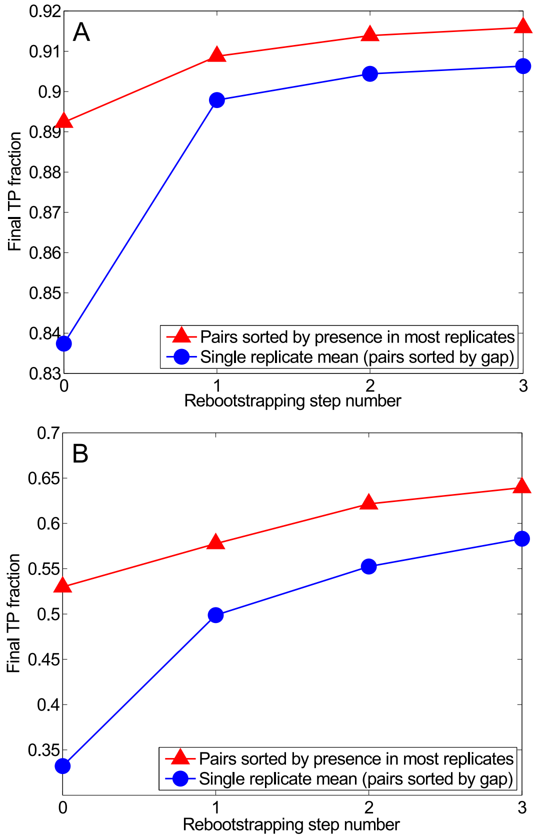

Fig. S9. Rebootstrapping: exploiting the high TP fraction of the HK-RR pairs predicted to be correct in most replicates of the IPA, which differ in their initial random pairings. (A) Rebootstrapping on the standard dataset (M = 5064 HK-RR pairs). The final TP fraction is plotted versus rebootstrapping step number. Step 0 corresponds to the IPA starting from random pairings (see main text and Fig. 6). 500 replicates are computed. We then take as a training set 1000 HK-RR pairs chosen randomly among those predicted to be correct in more than 50% of replicates. These pairs are chosen with probability equal to the fraction of replicates in which they are predicted to be true. The IPA is then performed again starting from such training sets. The process is then iterated. Here, 50 replicates were computed for steps 1, 2, and 3. The average final TP fraction is plotted (blue curve), as well as the TP fraction for the best M = 5064 pairs ranked by the fraction of replicates in which they are predicted to be true (red curve, see Fig. 6). Here, $N_{\text{increment}}$ = 6. (B) Rebootstrapping on a smaller dataset with M = 502 HK-RR pairs from 40 species (mean number of pairs per species $\langle m_p \rangle$ = 12.6). The process is the same as in (A), but here, at each rebootstrapping step, we take as a training set 200 HK-RR pairs chosen randomly among those predicted to be true in more than 25% of replicates at the previous step, and $N_{\text{increment}}$ = 1.

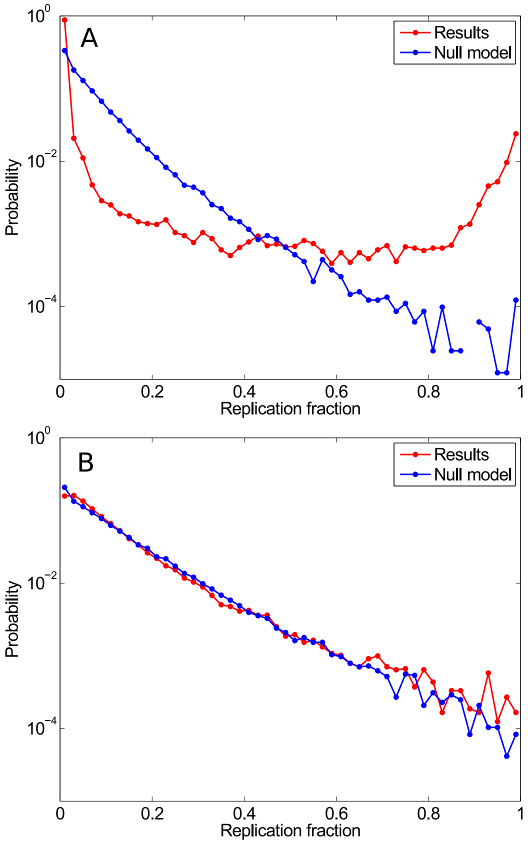

Fig. S10. Distribution of the fraction of replicates f_r of the IPA in which each possible within-species HK-RR pair is predicted as a pair. (A) Red curve: Distribution of f_r obtained by applying the IPA to the standard dataset (same data as in Fig. S8). Blue curve: Same dataset, but with each column of the alignment randomly scrambled. (B) HK-RR dataset with no correct pairs; a dataset of the same size as the standard one (M = 5062 in practice) that does not include any true HK-RR pairs was constructed. Red curve: Distribution of f_r obtained by applying the IPA to this dataset with no correct pairs. Blue curve: Same alignment, but with each column randomly scrambled. For each curve, 500 IPA replicates that differ in their initial random pairings were used, with $N_{\text{increment}}$ = 6. All data is binned into 50 equally-spaced bins between f_r = 0 and f_r = 1.

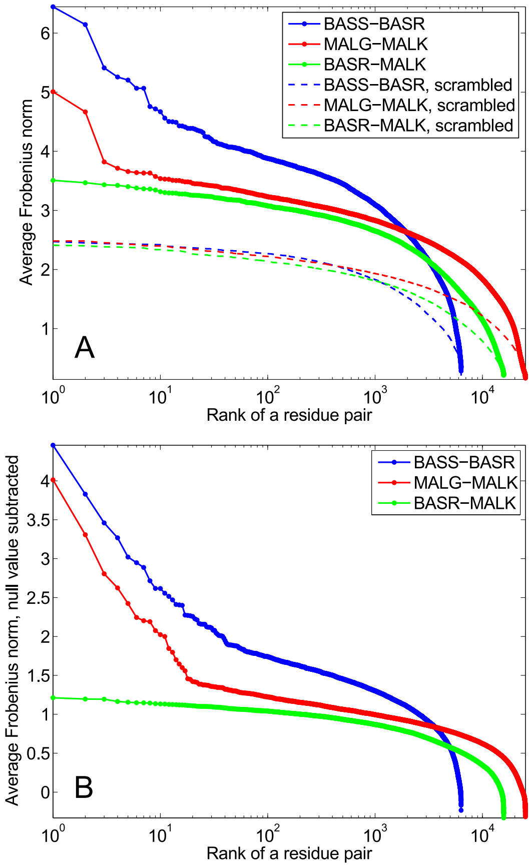

Fig. S11. Residue-based signature of protein-protein interactions. The Frobenius norm of the amino-acid couplings was evaluated for each pair of residue sites at the final iteration of the IPA, for datasets comprising ∼5000 homologs of the interacting pairs BASS-BASR and MALG-MALK, and of the non-interacting pair BASR-MALK. For each of these protein family pairs, the Frobenius norms were also calculated at the final iteration of the IPA on scrambled-within-column datasets (null model). (A) Frobenius norms averaged over 500 IPA replicates that differ in their initial random pairings, and then ranked by decreasing value. (B) Same average Frobenius norms, normalized by subtracting the average null value for each residue pair. For each curve, the IPA was run with $N_{\text{increment}} = 50$. The pairs with highest Frobenius norms, corresponding to the top predicted contacts, are outliers for both interacting family pairs, but not for the non-interacting pair BASR-MALK.

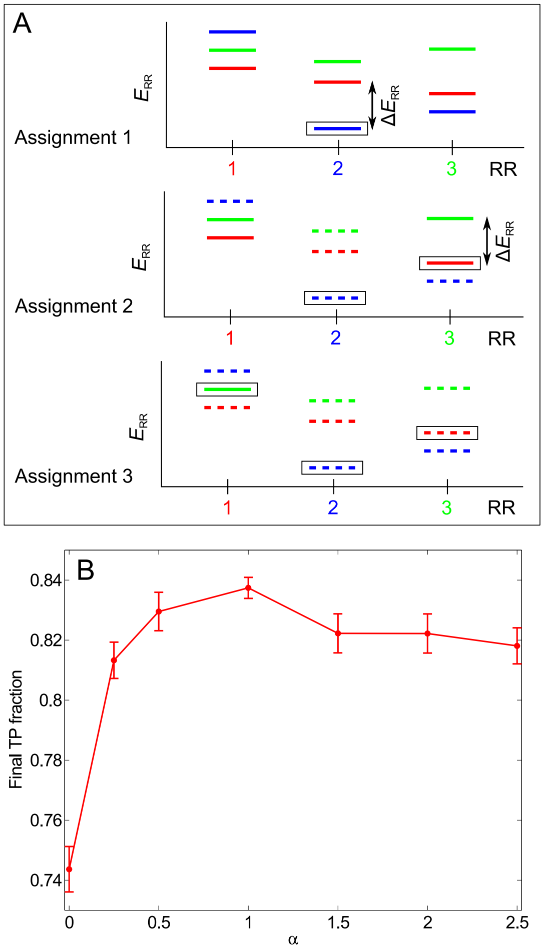

Fig. S12. Scoring by gap. (A) Determination of the confidence score of each assigned HK-RR pair in a given iteration of the IPA. In this schematic, we consider a species with three HKs and RRs. In the energy spectra showing the interaction energies for each RR with all three HKs, each color represents a given HK (red: HK 1, partner of RR 1; blue: 2; green: 3). Assignment 1: The pair with the lowest interaction energy (HK 2 - RR 2, boxed) is selected. The energy gap ΔE_{RR} is shown. Here $n_{\mathrm{RR}} = 0$ since no HK has been removed from consideration yet. Assignment 2: The HK and RR previously paired are removed from further consideration (dashed energy levels). The next pair with the lowest energy (HK 1 - RR 3, boxed) is chosen among the remaining ones. Here $n_{\mathrm{RR}} = 1$ since HK 2, which was paired previously, had a lower interaction energy with RR 3 than HK 1. Using the *ad hoc* confidence score $\Delta E_{\mathrm{RR}}/(n_{\mathrm{RR}} + 1)$, this (incorrect) pair is penalized with respect to the (correct) one made in the first assignment, even though their energy gaps are similar. Assignment 3: Only one possible pair remains. It is made, and its confidence score is taken to be equal to the lowest previously calculated confidence score for that species (the second one here). At each HK-RR pair assignment, symmetric confidence scores $\Delta E_{\mathrm{HK}}/(n_{\mathrm{HK}} + 1)$ are also calculated from the energy spectra showing the interaction energies for each HK with all three RRs. The final confidence score of a pair, denoted by $\Delta E/(n + 1)$, is the smallest of these two scores, i.e. $\min\{\Delta E_{\mathrm{RR}}/(n_{\mathrm{RR}} + 1),\ \Delta E_{\mathrm{HK}}/(n_{\mathrm{HK}} + 1)\}$. (B) More generally, in every iteration of the IPA, each predicted HK-RR pair can be scored by $\Delta E/(n + 1)^{\alpha}$, where α is a parameter. Red curve: Average final TP fraction obtained versus α; error bars: 95% confidence intervals around the mean. The IPA was performed on the standard dataset, with $N_{\mathrm{increment}} = 6$. Results are averaged over 200 replicates that differ in their initial random pairings for all α except $\alpha = 1$, for which 500 replicates were computed. As we found the highest TP fraction for $\alpha = 1$, all the results elsewhere in the paper were obtained using $\alpha = 1$.

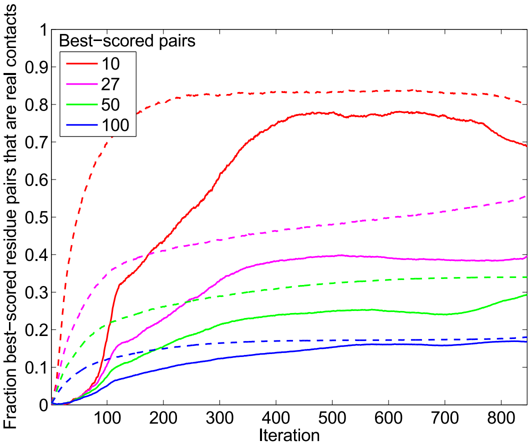

Fig. S13. Training of the couplings during the IPA: effect of the average product correction. Residue pairs comprised of an HK site and an RR site were scored by the average-product corrected Frobenius norm of the couplings involving all possible residue types at these two sites. The best-scored residue pairs were compared to the 27 HK-RR contacts found experimentally in Ref. [28]. Solid curves: Fraction of residue pairs that are real contacts (among the *k* best-scored pairs for four different values of *k*) versus the iteration number in the IPA. Dashed curves: Ideal case, where at each iteration $N_{\text{increment}}$ randomly-selected correct HK-RR pairs are added to the CA. The overall fraction of residue pairs that are real HK-RR contacts, yielding the chance expectation, is only 3.8×10^{-3}. As in Fig. 4, the IPA was performed on the standard dataset with $N_{\text{increment}} = 6$, and all data is averaged over 500 replicates that differ in their initial random pairings.

REFERENCES

1. Barakat M et al. (2009) P2CS: a two-component system resource for prokaryotic signal transduction research. *BMC Genomics* 10:315.
2. Ortet P, Whitworth DE, Santaella C, Achouak W, Barakat M (2015) P2CS: updates of the prokaryotic two-component systems database. *Nucleic Acids Res.* 43(Database issue):D536–541.
3. Cheng RR, Morcos F, Levine H, Onuchic JN (2014) Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.* 111(5):E563–571.
4. Rees DC, Johnson E, Lewinson O (2009) ABC transporters: the power to change. *Nat. Rev. Mol. Cell Biol.* 10(3):218–227.
5. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* 3:e02030.
6. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* 106(1):67–72.
7. Morcos F et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* 108(49):E1293–1301.
8. Marks DS et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12):e28766.
9. Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M (2011) Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. *PLoS ONE* 6(5):e19729.
10. Jaynes ET (1957) Information Theory and Statistical Mechanics. *Phys. Rev.* 106(4):620–630.
11. Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440(7087):1007–1012.
12. Lezon TR, Banavar JR, Cieplak M, Maritan A, Fedoroff NV (2006) Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 103(50):19033–19038.
13. Mora T, Walczak AM, Bialek W, Callan CG (2010) Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. U.S.A.* 107(12):5405–5410.
14. Bialek W et al. (2012) Statistical mechanics for natural flocks of birds. *Proc. Natl. Acad. Sci. U.S.A.* 109(13):4786–4791.
15. Wood K, Nishida S, Sontag ED, Cluzel P (2012) Mechanism-independent method for predicting response to multidrug combinations in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 109(30):12254–12259.
16. Ferguson AL et al. (2013) Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* 38(3):606–617.
17. Mann JK et al. (2014) The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput. Biol.* 10(8):e1003776.
18. Dwyer RS, Ricci DP, Colwell LJ, Silhavy TJ, Wingreen NS (2013) Predicting functionally informative mutations in Escherichia coli BamA using evolutionary covariance analysis. *Genetics* 195(2):443–455.
19. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M (2016) Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Mol. Biol. Evol.* 33(1):268–280.
20. Plefka T (1982) Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *J. Phys. A: Math. Gen.* 15(6):1971–1978.
21. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E* 87(1):012707.
22. Baldassi C et al. (2014) Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS ONE* 9(3):e92721.
23. Ekeberg M, Hartonen T, Aurell E (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.* 276:341–356.
24. Skerker JM et al. (2008) Rewiring the specificity of two-component signal transduction systems. *Cell* 133(6):1043–1054.
25. Capra EJ et al. (2010) Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. *PLoS Genet.* 6(11):e1001220.
26. Podgornaia AI, Casino P, Marina A, Laub MT (2013) Structural basis of a rationally rewired protein-protein interface critical to bacterial signaling. *Structure* 21(9):1636–1647.
27. Podgornaia AI, Laub MT (2015) Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* 347(6222):673–677.
28. Casino P, Rubio V, Marina A (2009) Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. *Cell* 139(2):325–336.

16 Bibol *et al.*