Supporting Information

Weigt *et al.* 10.1073/pnas.0805923106

SI Text

This supporting text exposes the technical details of the global inference approach to identify direct residue contacts in protein–protein interactions from genomic libraries.

Data and Frequency Counts. As described in the main text, data are extracted by scanning multispecies genomes with hidden Markov models for the HisKA and RR classes of sensor kinases (SK) and response regulators (RR). SK/RR pairs being adjacent on the chromosome are selected, resulting in $M = 2{,}546$ pairs. Each pair is concatenated into a single string of $N = N_{\mathrm{HK}} + N_{\mathrm{RR}} = 88 + 124 = 212$ letters from a $Q = 21$-letter alphabet. The alphabet contains 1 letter for each amino acid, and a supplementary letter for alignment gaps. In the following, the different nature of these letters is not distinguished. As a result, data are provided in the form of a $N \times M$ matrix $D = (A_i^a)$ with $i = 1, \ldots, N$ enumerating residue positions (sites) and $a = 1, \ldots, M$ protein pairs (samples).

Inference is *not* based directly on these sequences, but uses statistical sequence properties as given by the single- and 2-site occupancies. Frequency counts are introduced,

$$f_i(A_i) = \frac{1}{\lambda Q + M}\left[\lambda + \sum_{a=1}^{M} \delta(A_i, A_i^a)\right] \qquad [1]$$

$$f_{ij}(A_i, A_j) = \frac{1}{\lambda Q + M}\left[\frac{\lambda}{Q} + \sum_{a=1}^{M} \delta(A_i, A_i^a)\,\delta(A_j, A_j^a)\right]$$

for the occupancy of residue position i by amino acid type A_i and for the joint occupancy of position pair (i, j) by (A_i, A_j). The Kronecker symbol $\delta(\cdot,\cdot)$ equals one if arguments are equal, and zero else. The above formulae contain a pseudocount $\lambda > 0$, which can be considered as a prior needed to regularize frequency counts for finite-sample effects. It prevents counts from equaling zero, which in the inference task below would lead to divergent couplings. The pseudocount λ, which is set to 1 in our implementation, becomes less and less relevant the larger the dataset is, i.e., for $M \gg \lambda Q$. For the data used in this work, it has a very weak effect on the ranking of residue position pairs described below. Note that the pseudocount in Eqs. **1** is introduced in a way to ensure consistency,

$$\sum_{A_j=1}^{Q} f_{ij}(A_i, A_j) \equiv f_i(A_i) \qquad [2]$$

for all $i, j \in \{1, \ldots, N\}$ and all $A_i \in \{1, \ldots, Q\}$.

The quantities f_i and f_{ij} will be the actual inputs to both the calculation of mutual information and to the global model inference by message passing. The use of frequency counts instead of complete amino acid sequences has an important impact on the computational complexity of inference: Larger datasets for additional genomes will give a better estimate of the occupancies, but do not slow down the actual inference.

In principle one could also ask for statistical features beyond residue pairs, i.e., frequency counts including 3 or more residues. This is, however impossible with current data: A 3-residue distribution describes $21^3 = 9{,}261$ amino acid combinations, so the number M of interacting protein pairs should be increased by at least one order of magnitude compared with current data.

Mutual Information (MI). A simple local method to detect correlation between residue positions is given by mutual information. Raw MI is calculated as

$$MI_{ij}^{(\mathrm{raw})} = \sum_{A_i, A_j = 1}^{Q} f_{ij}(A_i, A_j) \ln \frac{f_{ij}(A_i, A_j)}{f_i(A_i) f_j(A_j)}. \qquad [3]$$

MI measures the Kullback–Leibler divergence of the joint distribution $f_{ij}(A_i, A_j)$ from its factorized counter part $f_i(A_i) f_j(A_j)$, MI equals zero if and only if f_{ij} is actually factorized, and it is positive else.

In our case, amino acid distributions are estimated from a finite sample. This will lead, even for a priori factorized cases, to a spurious nonzero contribution to the mutual information. This contribution actually depends on the single-site distributions (i.e., on the residue conservation in single sites), and is therefore expected to be different for different position pairs. To correct for this effect, the average MI of a null model is estimated and subtracted from the raw MI,

$$MI_{ij} = MI_{ij}^{(\mathrm{raw})} - \bar{M}_{ij}^{(0)} \qquad [4]$$

with the overbar denoting the average over 400 random realizations of the null model. The null model itself is very simple. We introduce a random permutation $\pi \in \mathcal{S}_M$ of the M samples, and estimate MI for pairs $(A_i^a, A_j^{\pi(a)})$. In the following we refer to this sample-size-corrected MI simply as mutual information.

Maximum Entropy Model. High mutual information can result from a strong direct coupling of sites i and j (which we would consider an indicator for spatial vicinity of residues in the folded protein/dimerized protein pair). It may also result from indirect coupling of i and j via intermediate sites. In this case, one would not expect residues to be necessarily close in 3-dimensional structure.

MI is a local measure; it encounters only 1 residue pair at a time. MI is intrinsically unable to disentangle direct from indirect coupling. Consequently, prediction of spatial vicinity of interacting residues by MI is restricted. Therefore a global approach is proposed that will lead to the notation of *direct information* (DI). DI measures the part of MI that results from the direct coupling of residue pairs.

Model derivation. To reach this aim, a global residue distribution $P(A_1, \ldots, A_N)$ for the joint behavior of *all* residues shall be constructed (actually, this task is computationally too expensive; we restrict it, therefore, to selections of up to 60 positions from both protein domains). It is not practical to construct this distribution directly as a histogram from the data since $O(21^N)$ samples are required. Direct estimates of joint frequencies from our $M = 2{,}546$ examples are possible only up to position pairs. Consistency of the statistical model $P(A_1, \ldots, A_N)$ to data is therefore restricted to coinciding marginal distributions up to the 2-site level,

$$\begin{aligned} f_i(A_i) &\equiv P_i(A_i) \\ &= \sum_{\{A_k | k \neq i\}} P(A_1, \cdots, A_N) \\ f_{ij}(A_i, A_j) &\equiv P_{ij}(A_i, A_j) \\ &= \sum_{\{A_k | k \neq i, j\}} P(A_1, \cdots, A_N). \end{aligned} \qquad [5]$$

Besides this condition, the minimally constrained model P will be reconstructed. The latter is given by the maximum entropy principle (1–3), i.e., we maximize the entropy

$$S = -\sum_{\{A_i\}} P(A_1, \cdots, A_N) \ln P(A_1, \cdots, A_N) \qquad \textbf{[6]}$$

under the constraint that all of Eqs. **5** is satisfied. Introduction of Lagrange multipliers for constraints leads to the form

$$P(A_1, \cdots, A_N) = \frac{1}{Z} \prod_{i<j} \exp\{-e_{ij}(A_i, A_j)\} \prod_i \exp\{h_i(A_i)\} \qquad \textbf{[7]}$$

which includes both purely local terms $h_i(A_i)$ and 2-site couplings $e_{ij}(A_i, A_j)$. The latter will finally serve to estimated the strength of direct coupling of residue positions i and j. The partition function

$$Z = \sum_{\{A_i\}} \prod_{i<j} \exp\{-e_{ij}(A_i, A_j)\} \prod_i \exp\{h_i(A_i)\} \qquad \textbf{[8]}$$

guarantees correct normalization of Eq. **7**. Note that in statistical physics this model is known as a disordered Q-state Potts model. It can be described equivalently by the *Hamiltonian*

$$\mathcal{H}(A_1, \cdots, A_N) = \sum_{i<j} e_{ij}(A_i, A_j) - \sum_i h_i(A_i) \qquad \textbf{[9]}$$

with pair-interaction energies $e_{ij}(A_i, A_j)$ and local fields $h_i(A_i)$. The global probability distribution P is given by the Boltzmann–Gibbs distribution (temperature is set to unity without loss of generality)

$$P(A_1, \cdots, A_N) = \frac{1}{Z} \exp\{-\mathcal{H}(A_1, \cdots, A_N)\}. \qquad \textbf{[10]}$$

Gauge fixing. Note that the conditions [**5**] are not independent: they are coupled by the consistency constraints [**2**] and by normalization of all P_i and P_{ij}. This is reflected by the possibility of changing parameters of the Hamiltonian in Eq. **9** without changing the value of $\mathcal{H}$; contributions can be shifted between local fields and interaction energies (gauge invariance). The number of parameters in Eq. **9** is $\binom{N}{2}q^2 + Nq$, but the number of free parameters can be determined to be only $\binom{N}{2}(q-1)^2 + N(q-1)$. To have a reasonable and reproducible inference result, one has to get rid of this degeneracy. Since the ultimate goal is estimating direct interaction strengths, it seems reasonable to put as much as possible of the Hamiltonian into local fields, and as few as possible into interaction energies. This intuitive idea can be formalized by asking to minimize the square norm $\Sigma_{A_i, A_j} [e_{ij}(A_i, A_j)]^2$ of all interaction terms under the condition of constant value of $\mathcal{H}$. This task can readily be fulfilled by the gauge-fixing conditions

$$\sum_A e_{ij}(A_i, A) = \sum_A e_{ij}(A, A_j) = 0 \qquad \textbf{[11]}$$

for all i, j, A_i, A_j. Furthermore, the invariance of model P with respect to additive constants in $\mathcal{H}$ allows to set

$$\sum_A h_i(A) = 0 \qquad \textbf{[12]}$$

for all i. These fixations reduce the number of free parameters in the model to the number of independent conditions in Eqs. **5**. We therefore expect that under this choice the inference task of P has a uniquely determined solution.

Note that this gauge fixing reproduces the Ising model Hamiltonian for $Q = 2$. For $Q \geq 3$ and interactions invariant w.r.t. permutations of the Q Potts values, the gauge-fixing condition leads to a vector representation of Potts spins in $Q - 1$ dimensions.

Model Inference by Gradient Descent. Model parameters can be estimated by iterating the following 2-step procedure:

1. For a given Hamiltonian, marginal distributions for single positions and position pairs have to be calculated.
2. By using gradient descent, parameters in the Hamiltonian are updated according to the difference of the estimated marginals and the frequencies of amino acid occurrences calculated from data.

A reasonable starting point for this procedure seems to be to set the interaction terms initially to zero, $e_{ij}^{(0)}(A_i, A_j) = 0$, and the fields to get the correct single-site marginals, $h_i^{(0)}(A_i) = c_i + \ln f_i(A_i)$ with c_i chosen according to the gauge condition [**12**]. Steps 1 and 2 are iterated until the difference between the model and the data-given frequency counts reaches some predefined precision.

Assuming that step 1 can be obtained (how to do this is the major concern in *Message Passing*), we have to define a correct update scheme for the model parameters. We choose

$$\begin{aligned} \Delta e_{ij}(A_i, A_j) &= \epsilon \Bigg[f_{ij}(A_i, A_j) - P_{ij}(A_i, A_j) \\ &\qquad - \frac{f_i(A_i) + f_j(A_j) - P_i(A_i) - P_j(A_j)}{Q} \Bigg] \\ \Delta h_i(A_i) &= \epsilon [f_i(A_i) - P_i(A_i)] \end{aligned} \qquad \textbf{[13]}$$

with step size ϵ. This choice preserves the gauge conditions [**11**, **12**].

A Bayesian Perspective. Why should Eqs. **13** converge to a unique solution? The answer can be given via a Bayesian approach. Given a model, i.e., given the interactions $e_{ij}(A_i, A_j)$ and the local fields $h_i(A_i)$, and assuming statistical independence of the data points, the probability of the data set $D = (A_i^a)$ is given as

$$\begin{aligned} &P(D | \{e_{ij}(A_i, A_j), h_i(A_i)\}) \\ &= \prod_{a=1}^{M} \left[\frac{1}{Z} \exp\left\{ -\sum_{i<j} e_{ij}(A_i^a, A_j^a) + \sum_i h_i(A_i^a) \right\} \right]. \end{aligned} \qquad \textbf{[14]}$$

This quantity is a convex function of the model parameters, so the probability of obtaining a dataset can be maximized by using simple gradient descent. Eqs. **6** follow immediately from the variation of the log-likelihood $\ln P(D \mid \{e_{ij}(A_i, A_j), h_i(A_i)\})$ with respect to $e_{ij}(A_i, A_j)$ and $h_i(A_i)$, where the gauge conditions **11** and **12** are imposed via Lagrange multipliers (which can be eliminated analytically).

Message Passing. Step 1 of the above iterative procedure is, however, computationally hard. The direct calculation of the marginal distributions of model $P(A_1, \ldots, A_N)$ requires the summation over all but 1 resp. 2 amino acids; the algorithmic complexity is of order $O(21^{N-1})$. System sizes would be restricted to 7–8 aa at most, and global inference would be hard to obtain.

A popular alternative is random sampling from $P(A_1, \ldots, A_N)$ by Monte Carlo Markov chains (MCMC). For a 21-states model, it is to be expected that fluctuations remain large even for long sampling times, and convergence can again be guaranteed only after exponential times. In this work, it is therefore proposed to use message passing (4, 5) as an approximate tool. Since message passing is based on the solution of deterministic equations, it will (if converging to a single solution) provide a unique solution for P. The approximate character of message passing is not well-controlled, but we expect deviations not to exceed those which come from imperfect sampling of the space of all possible amino acid configurations by (evolutionary related) biological input data.

For the model inference both single-site and 2-site marginals are needed. The first can be calculated by using message passing in the form of standard belief propagation (BP). To calculate 2-point distributions, we apply a generalization of BP recently called susceptibility propagation by Mézard and Mora (6).

Belief propagation and single-residue quantities. The marginals for single-residue positions can be evaluated by using standard BP. In this algorithm, messages $P_{i\to j}(A_i)$ (beliefs) are self-consistently exchanged between sites. BP equations for model **7** read

$$P_{i\to j}(A_i) \sim e^{h_i(A_i)} \prod_{k\neq i,j}\left[\sum_{A_k} e^{-e_{ki}(A_k,A_i)} P_{k\to i}(A_k)\right]. \qquad [15]$$

The message sent from i to j thus results from the local field in i and all messages sent from other positions $k \neq i, j$ to i. The proportionality in Eq. **15** signals that messages are normalized. The message $P_{i\to j}(A_i)$ can be understood as the marginal amino acid distribution in position i in a system, where position j has been removed. The method is therefore known under the name *cavity method* in the statistical physics of disordered systems.

BP equations can be solved iteratively by first initializing all messages arbitrarily, and then using Eq. **15** for updates. Our experience shows that random sequential updates are most efficient. Iteration has to be repeated until no message is updated by more than a predefined precision value, here 10^{-5} was chosen. Having calculated all messages, we can estimate the true marginal by including all messages sent to a residue,

$$P_i(A_i) \sim e^{h_i(A_i)} \prod_{k\neq i}\left[\sum_{A_k} e^{-e_{ki}(A_k,A_i)} P_{k\to i}(A_k)\right]. \qquad [16]$$

In this approach, we assume all $h_i(A_i)$ to be known and determine their conjugate quantities $P_i(A_i)$. In principle, our inference task is defined in an inverse way. We know from biological data the values that the marginal distributions will take at the end, $P_i(A_i) = f_i(A_i)$, and we want to infer the value of the fields. Eqs. **15, 16** allow us to explicitly achieve this inversion by eliminating the message sent from j to i in [**16**]. We write

$$P_{i\to j}(A_i) \sim \frac{f_i(A_i)}{\sum_{Aj} e^{-e_{ij}(A_i,A_j)} P_{j\to i}(A_j)} \qquad [17]$$

and determine fields after convergence as

$$e^{h_i(A_i)} \sim \frac{f_i(A_i)}{\prod_{j\neq i}\sum_{Aj} e^{-e_{ij}(A_i,A_j)} P_{j\to i}(A_j)} \qquad [18]$$

The proportionality constant in the last equation can be fixed by using the gauge condition [**12**]. This inversion guarantees that the single-residue marginals as estimated by BP always are identical to those given by protein data; therefore, the gradient descent update for fields becomes obsolete.

Note that in Eq. **17** the message update factorizes over the links. The message sent from site i to j depends only on the marginal in i, the interaction $e_{ij}(A_i, A_j)$ between the 2 sites, and the message sent from j to i. It can be understood as the exact message-passing approach for a hypothetical system where i and j together with their link are isolated from the full system, and the correct single site marginals are imposed on the 2 sites.

This trick largely improves the efficiency of inference by message passing. Potts models for $Q > 3$ tend to show first-order phase transitions, small changes in parameters may lead to major changes of the BP solution toward new (metastable) states. Such transitions can be avoided by fixing the marginal distribution instead of the local fields.

Susceptibility propagation and residue-pair distributions. If the underlying graphical structure would be a tree (i.e., a loop-free graph), BP would give exact results for single-site marginals. The latter could be easily extended to 2-site functions since there would be only 1 connecting path between any 2 variables i and j. In our situation, the model is, however, a priori completely connected. We do not know which couplings are relevant, so we have to start with a fully connected system. (In fact, the system will stay fully connected but some links will become weak. It would be interesting to design a dilution strategy that allows us to systematically delete links without losing too much in fitting precision.) Correlations between 2 sites can thus be induced by a large number of different paths. The crucial idea of Mézard and Mora (6) to resolve this problem consists in the application of the fluctuation dissipation theorem,

$$\frac{\partial P_i(A_i)}{\partial h_j(A_j)} = P_{ij}(A_i,A_j) - P_i(A_i)P_j(A_j), \qquad [19]$$

i.e., the response of the residue statistics in site i to a change in the local field in site j equals the connected part of the 2-point distribution. Now it is sufficient to derive Eqs. **15, 16** with respect to fields in arbitrary positions to get self-consistent equations called susceptibility propagation. Introducing messages

$$M_{i\to j;k}(A_i,A_k) = \frac{\partial P_{i\to j}(A_i)}{\partial h_k(A_k)} \qquad [20]$$

we can write

$$M_{i\to j;k}(A_i,A_k) = P_{i\to j}(A_i)\left[\delta(i,k)\delta(A_i,A_k) + \sum_{l\neq i,j}\frac{\sum_{A_l} e^{-e_{li}(A_l,A_i)} M_{l\to i;k}(A_l,A_k)}{\sum_{A_l} e^{-e_{li}(A_l,A_i)} P_{l\to i}(A_l)} - c_{i\to j;k}(A_k)\right]. \qquad [21]$$

The quantity $c_{i\to j;k}(A_k)$ results from the derivative of the normalization in Eq. **16** and is chosen such that

$$\sum_{A_i} M_{i\to j;k}(A_i,A_k) = 0 \qquad [22]$$

for all i, j, k and all A_k. This condition follows easily from the fact that $\Sigma_{A_i} P_{i\to j}(A_i) \equiv 1$, so its derivative with respect to $h_k(A_k)$ vanishes identically. In complete analogy we derive

$$\frac{\partial P_i(A_i)}{\partial h_j(A_j)} = P_i(A_i)\left[\delta(i,j)\delta(A_i,A_j) + \sum_{l\neq i}\frac{\sum_{A_l} e^{-e_{li}(A_l,A_i)} M_{l\to i;j}(A_l,A_j)}{\sum_{A_l} e^{-e_{li}(A_l,A_i)} P_{l\to i}(A_l)} - c_{ij}(A_j)\right]. \qquad [23]$$

Note that the calculation of the 2-site distribution is much more time consuming then its single-site counter part. In particular, the number of messages goes up to $O(N^3Q^2)$, and an efficient implementation of the above equations requires $O(N^4Q^2)$ elementary steps for a global system update. This time complexity is currently the bottleneck of our approach. By restricting the inference to the 60-residue positions involved in the highest MI values (for the SK/RR dataset we include the first 140 MIs), the task can be completed on a single CPU of Dell dual quad-core 2.33 GHz Xeon processor in 4–5 days. The full problem including all 212 residues present in the dataset would require several years of running time and therefore remains infeasible for global inference.

These equations complete the task of estimating single- and 2-site distributions from a general model with Q states variables and pair interactions. It can be used in the first of the 2 steps of *Model inference by gradient descent* to complete the task of inferring coupling constants.

Coupling Strength and Direct Information. The main motivation to infer a global statistical model was to divide pairs of high MI into 2 classes, the first one being characterized by a strong direct coupling, the second characterized by its absence. To maintain consistency it seems logical to estimate the contribution of the direct link to the mutual information, a quantity called here *direct information* (DI). To achieve this aim, we define 2-site distributions

$$P_{ij}^{(\mathrm{dir})}(A_i,A_j) = \frac{P_{i\to j}(A_i)e^{-e_{ij}(A_i,A_j)}P_{j\to i}(A_j)}{\sum_{A_1,A_2}P_{i\to j}(A_1)e^{-e_{ij}(A_1,A_2)}P_{j\to i}(A_2)}. \qquad \textbf{[24]}$$

This distribution describes 2 variables that are uniquely coupled by the direct link (i, j) of interaction $e_{ij}(A_i, A_j)$, but that have the correct marginal distributions f_i and f_j. No contributions coming from indirect paths connecting i and j via intermediate sites exist in this expression. Remember that, due to the factorization of Eq. **17** over the links, the direct pair distribution $P_{ij}^{(\mathrm{dir})}$ (A_i, A_j) can be calculated separately for each link. It allows measurement of the DI as

$$DI_{ij} = \sum_{A_i,A_j} P_{ij}^{(\mathrm{dir})}(A_i,A_j)\ln\frac{P_{ij}^{(\mathrm{dir})}(A_i,A_j)}{f_i(A_i)f_j(A_j)}. \qquad \textbf{[25]}$$

Compared with the mutual information $M_{ij}(A_i, A_j)$ this quantity measures only the contribution introduced by the direct link (i, j) to the correlations between the corresponding amino acids, so it is to be expected to be strictly smaller than $M_{ij}(A_i, A_j)$. We use this quantity as a measure for the direct coupling strength. Other measures, e.g., the norm $\|e_{ij}\|^2 = \Sigma_{A_i, A_j}$ $[e_{ij}(A_i, A_j)]^2$, can be introduced, but are found to contain essentially the same information. Even if some pairs change relative order in ranking according to these different measures, the distinction in weak and strong interactions given in the main part of the article remain invariant. Furthermore, DI is invariant under different gauges of Hamiltonian $\mathcal{H}$, whereas other measures might depend on the selected gauge.

Algorithmic Implementation. The above ideas are summarized in a description of the step-by-step implementation of the method.

1. Starting from the concatenated protein sequences of mated SK/RR pairs, for all residue positions resp. position pairs marginal counts $f_i(A_i)$ resp. $f_{ij}(A_i, A_j)$ according to Eqs. **1** are determined.
2. The mutual informations, MI$_{ij}$, are calculated by using Eqs. **3** and **4**; pairs are ordered according to MI values.
3. A MI cutoff is determined such that *a* given number of residue positions is contained in the network of supercutoff pairings (in our case, 60 positions). All pairs of these positions are considered during inference, including pairs inside the proteins and pairs of low MI. The statistical model is initialized via $e_{ij}^{(0)}$ $(A_i, A_j) = 0$, $h_i^{(0)}$ $(A_i) = c_i + \ln f_i(A_i)$.
4. BP Eq. **17** is solved iteratively for messages $P_{i\to j}(A_i)$. A random sequential update is used.
5. Eq. **21** is solved iteratively, using the previous solution of the BP equations. The solution is used to determine 2-site distributions by using Eqs. **23**.
6. If the 2-site distributions P_{ij} differ more than the desired precision from the data distributions f_{ij}, couplings are updated following the first of Eqs. **13**, and the previous 2 steps (belief and susceptibility propagation) are used to determine the updated 2-site distributions.
7. Once the desired precision is reached, Eqs. **24** and **25** are used to determine DI for all considered residue pairs. They are used for ranking of the direct interactions of interprotein pairs.

1. Jaynes ET (1949) Information theory and statistical mechanics. *Physiol Rev* 106:620–630.
2. Jaynes ET (1949) Information theory and statistical mechanics. *Physiol Rev* 108:171–190.
3. Bialek W, Ranganathan R (2007) Rediscovering the power of pairwise interactions, preprint: arXiv.org:0712.4397 [q–bio.QM].
4. Kschischang FR, Frey BJ, Loeliger H-A (2001) Factor graphs and the sum-product algorithm. *IEEE Trans Inf Theory* 47:498–519.
5. Yedidia JS, Freeman WT, Weiss Y (2001) Generalized belief propagation. *News Physiol Sci* 13:689–695.
6. Mézard M, Mora T (2008) Constraint satisfaction and neural networks: A statistical-physics perspective. preprint: arXiv.org:0803.3061 [q–bio.NC].

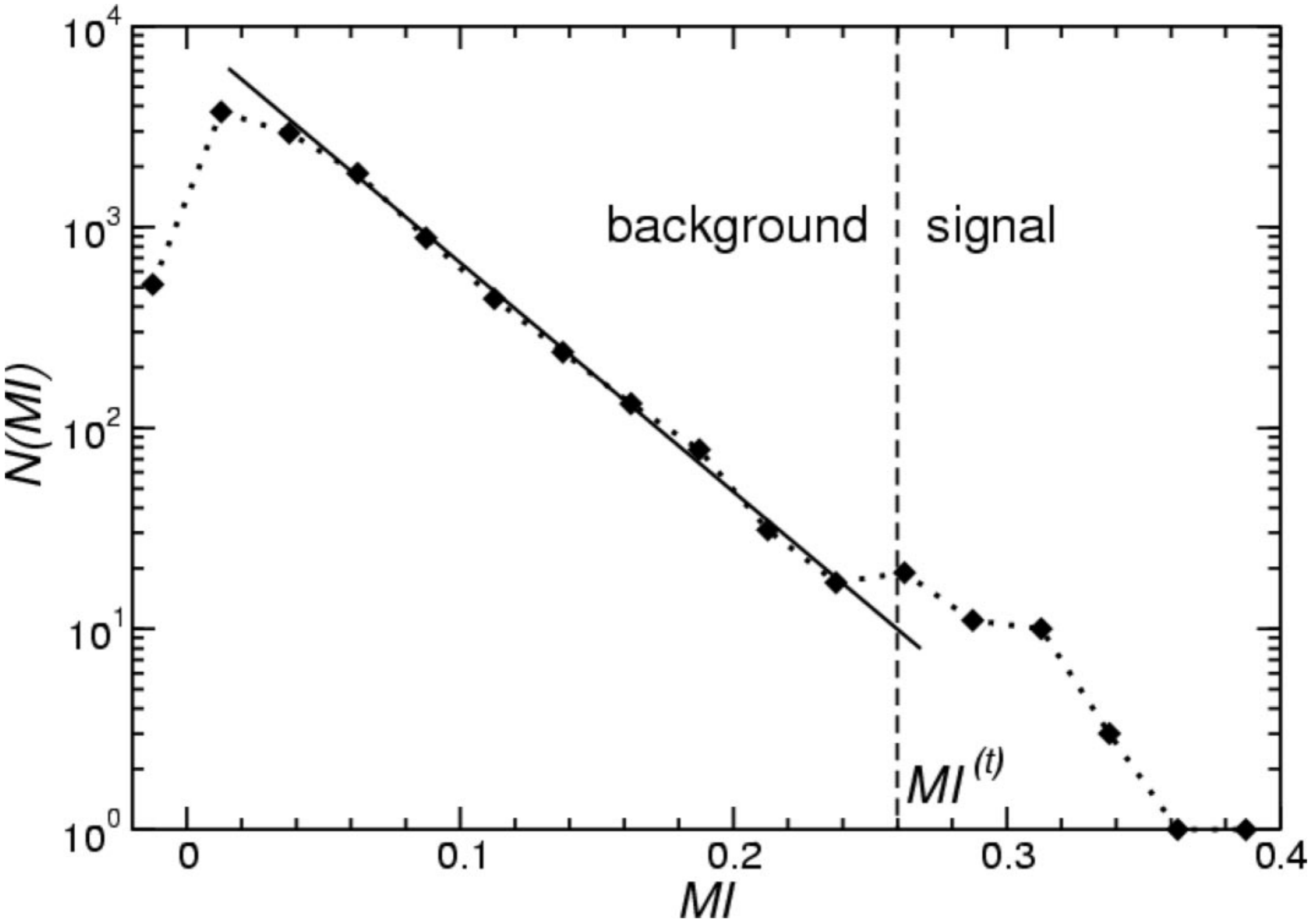


Fig. S1. The distribution of mutual information values MI reveals a threshold value $\mathrm{MI}^{(t)}$ for significant residue correlation. The sample size corrected mutual information values MI for the set of >10,000 histidine kinase/response regulator residue pairings were plotted against their number of occurrence *N*(MI). For $\mathrm{MI} < \mathrm{MI}^{(t)} \simeq 0.26$ (dashed line), the distribution is consistent with an exponential background over 2 decades (solid line). Deviation above the background begins in the region around $\mathrm{MI}^{(t)}$.

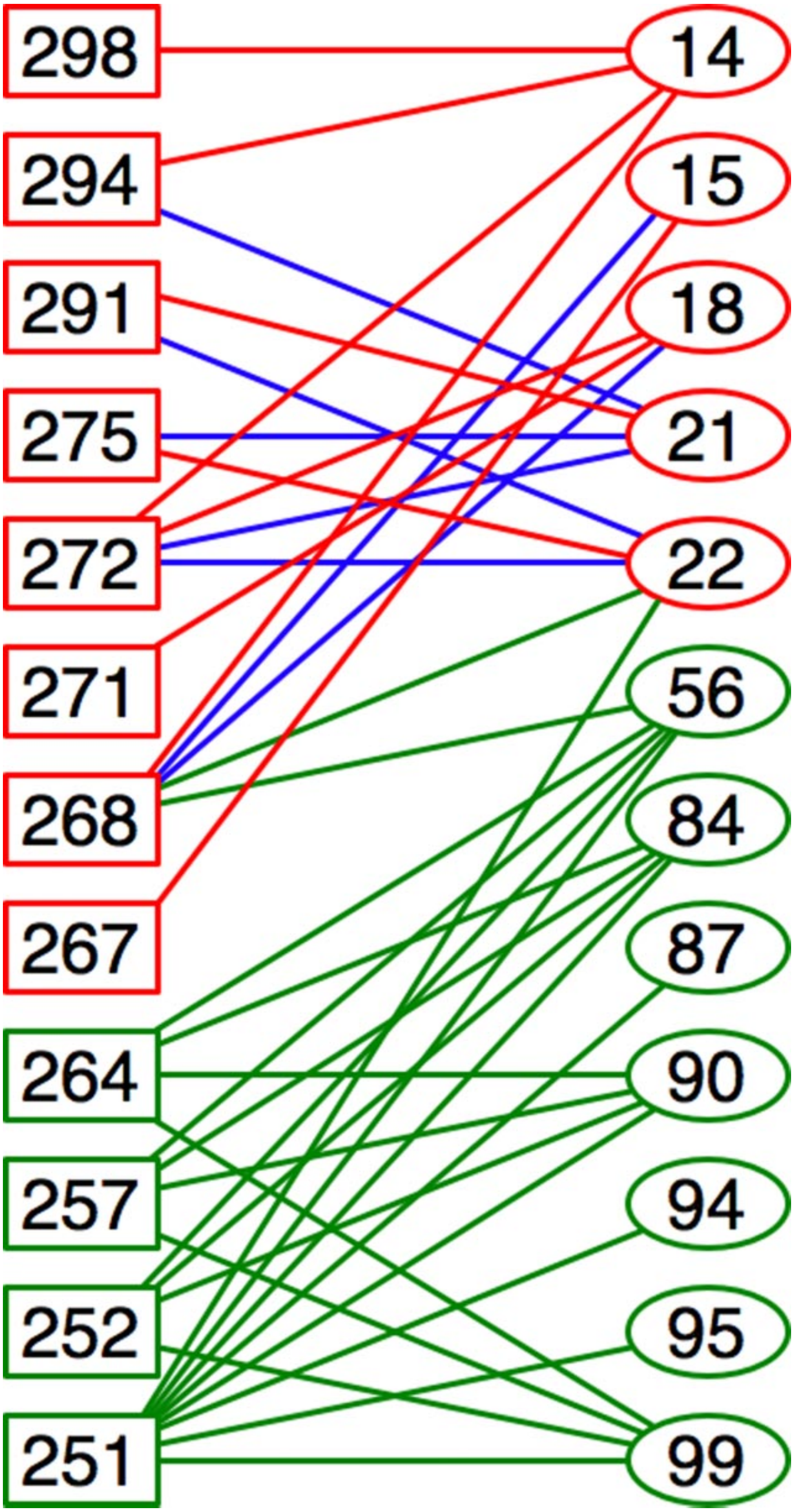


Fig. S2. MI and DI identify a network of coupled residues. Displayed is the network of coupled residue between SK (rectangular nodes) and RR (elliptic nodes). Elements of group I (displayed in red) are coupled via links showing both high MI and DI (red links), the links in group II (green links) have high MI but low DI. The high MI of the latter group results from the high connectivity among its group members. The links close to threshold (blue zone in Fig. 2*A*) are depicted in blue; they are seen to further connect residues in group I. The numbering refers to 2 specific proteins used in the further discussion, HK853 from *Thermotoga maritima* for SK and Spo0F from *Bacillus subtilis* for RR.

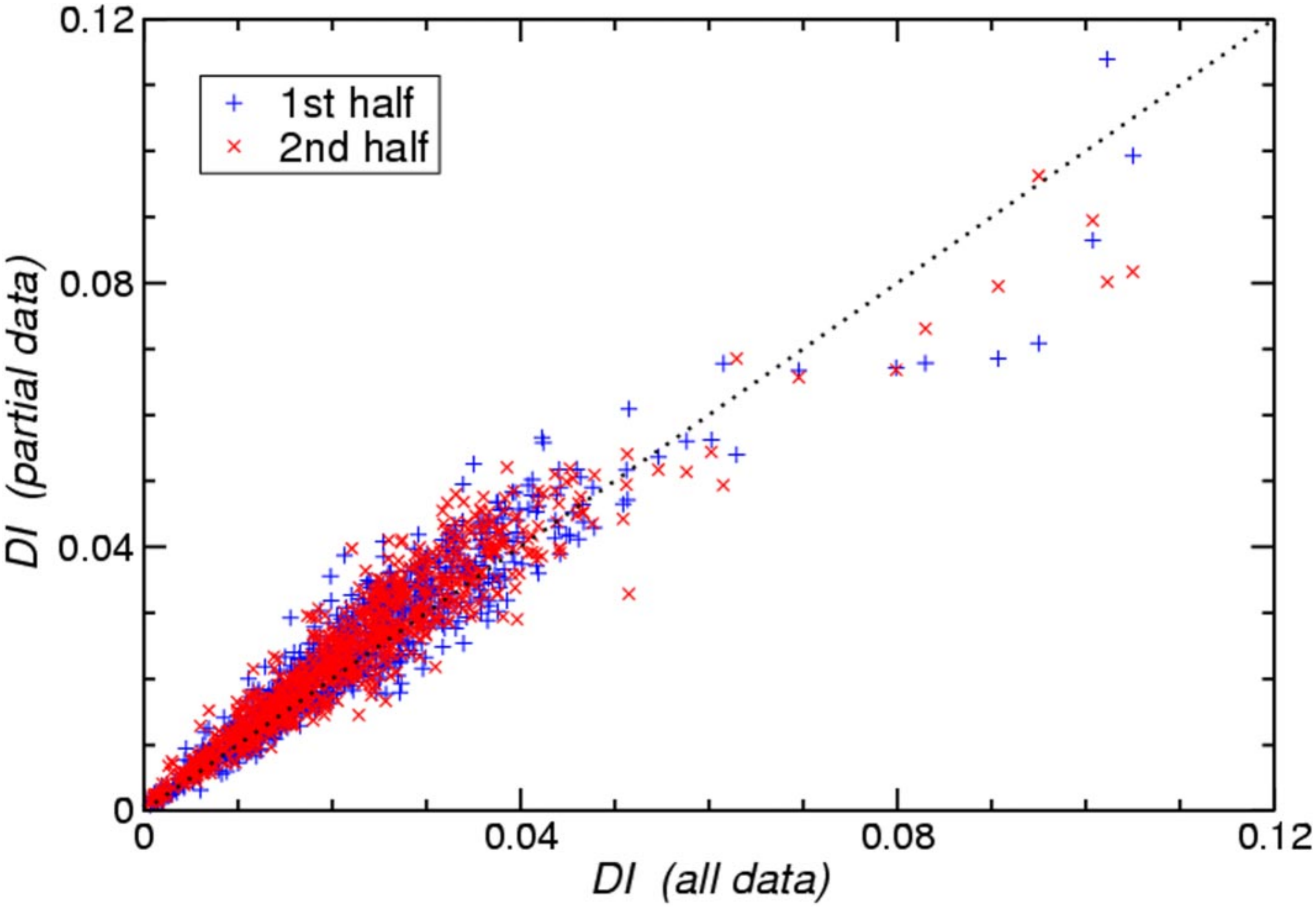


Fig. S3. The inferred DI values are robustly reproduced on disjoint data subsets. DI values as inferred by using only the first half (blue +) or the second half (red ×) of the mated sequences are plotted versus DI values inferred from all data. The set of top-ranking residue pairs remains almost unchanged. A small but systematic change from smaller to larger datasets can be observed. Small DI values become smaller, high DI values larger by extending the dataset. Therefore, larger datasets allow for a better separation of relevant DI values from the background.

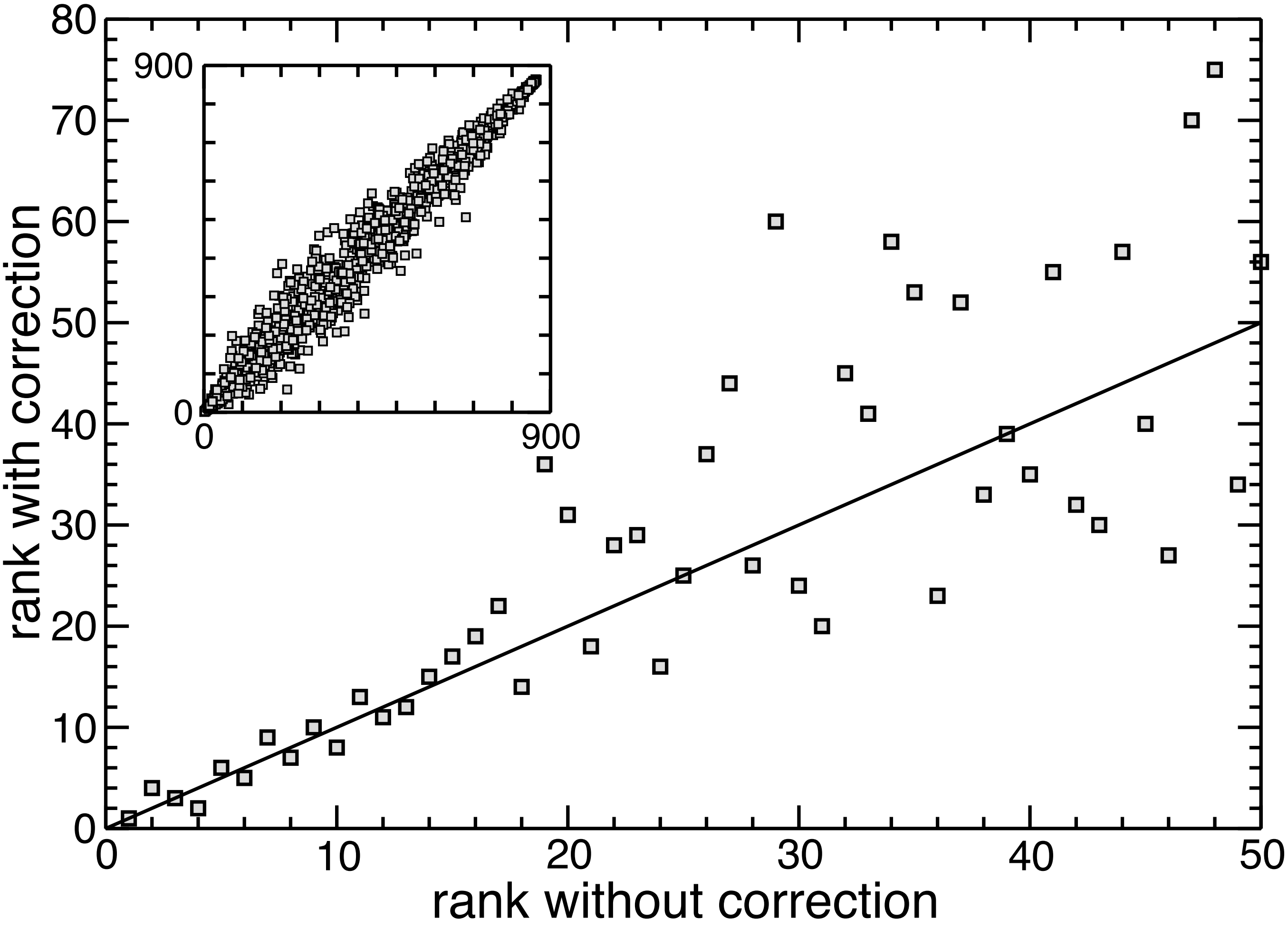
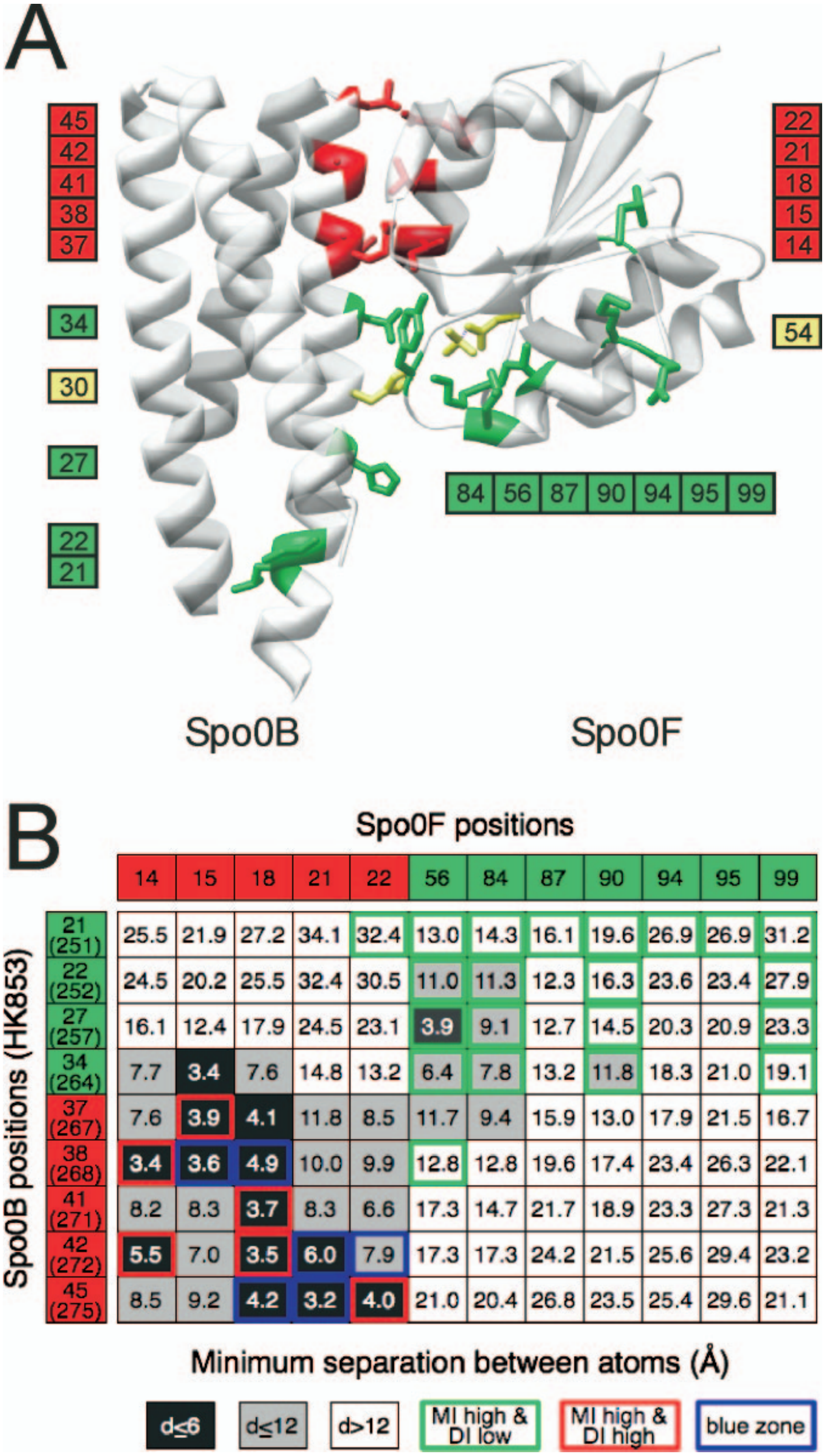


Fig. S4. DI ranking is robust with respect to sampling corrections. DI rank without and with reweighting to correct for sampling effects. For each interacting SK/RR pair $\overline{A^a}$, the number n^a of sequences having $>$80% sequence identity with $\overline{A^a}$ is determined, and $\overline{A^a}$ is assigned a weight $1/(n^a + 1)$ in frequency counts. Weights range from 1/14 to 1, the effective sample number calculated as the sum of all weights goes down to 1,792. DI is determined for reweighted frequency counts, and compared with DI without reweighting. The main figure shows the highest 50 ranks. In between the 10 highest-ranking position pairs, one finds 10 common pairs; in between the first 20 pairs, 17; in between the first 30 ranks, 25 common position pairs. (*Inset*) Full DI rankings; their Pearson correlation coefficient is 0.98.



	14	15	18	21	22	56	84	87	90	94	95	99
21 (251)	25.5	21.9	27.2	34.1	32.4	13.0	14.3	16.1	19.6	26.9	26.9	31.2
22 (252)	24.5	20.2	25.5	32.4	30.5	11.0	11.3	12.3	16.3	23.6	23.4	27.9
27 (257)	16.1	12.4	17.9	24.5	23.1	3.9	9.1	12.7	14.5	20.3	20.9	23.3
34 (264)	7.7	3.4	7.6	14.8	13.2	6.4	7.8	13.2	11.8	18.3	21.0	19.1
37 (267)	7.6	3.9	4.1	11.8	8.5	11.7	9.4	15.9	13.0	17.9	21.5	16.7
38 (268)	3.4	3.6	4.9	10.0	9.9	12.8	12.8	19.6	17.4	23.4	26.3	22.1
41 (271)	8.2	8.3	3.7	8.3	6.6	17.3	14.7	21.7	18.9	23.3	27.3	21.3
42 (272)	5.5	7.0	3.5	6.0	7.9	17.3	17.3	24.2	21.5	25.6	29.4	23.2
45 (275)	8.5	9.2	4.2	3.2	4.0	21.0	20.4	26.8	23.5	25.4	29.6	21.1

Fig. S5. Direct Information (DI) identifies contact residues in the Spo0B/Spo0F cocrystal structure. (*A*) All individual residue positions in HisKA and RR domain that show at least 1 pairing with MI $>$ MI$^{(t)}$ were mapped on the Spo0B/Spo0F structure, except for HisKA α2-helix positions (291, 294, and 298), which do not map well to the Spo0B structure. Phosphotransfer residues are in yellow. (*B*) The distances were measured for all possible combination of these residues. Distances up to 6 Å are highlighted in black, up to 12 Å are in gray, and distances $>$12 Å are in white. Residue pairings with DI $>$ 0.058 and MI $>$ 0.26 are framed in red (group I), with DI $<$ 0.058 and MI $>$ 0.26 are framed green (group II), and residues in the blue zone of Fig. 2*A* are framed in blue.

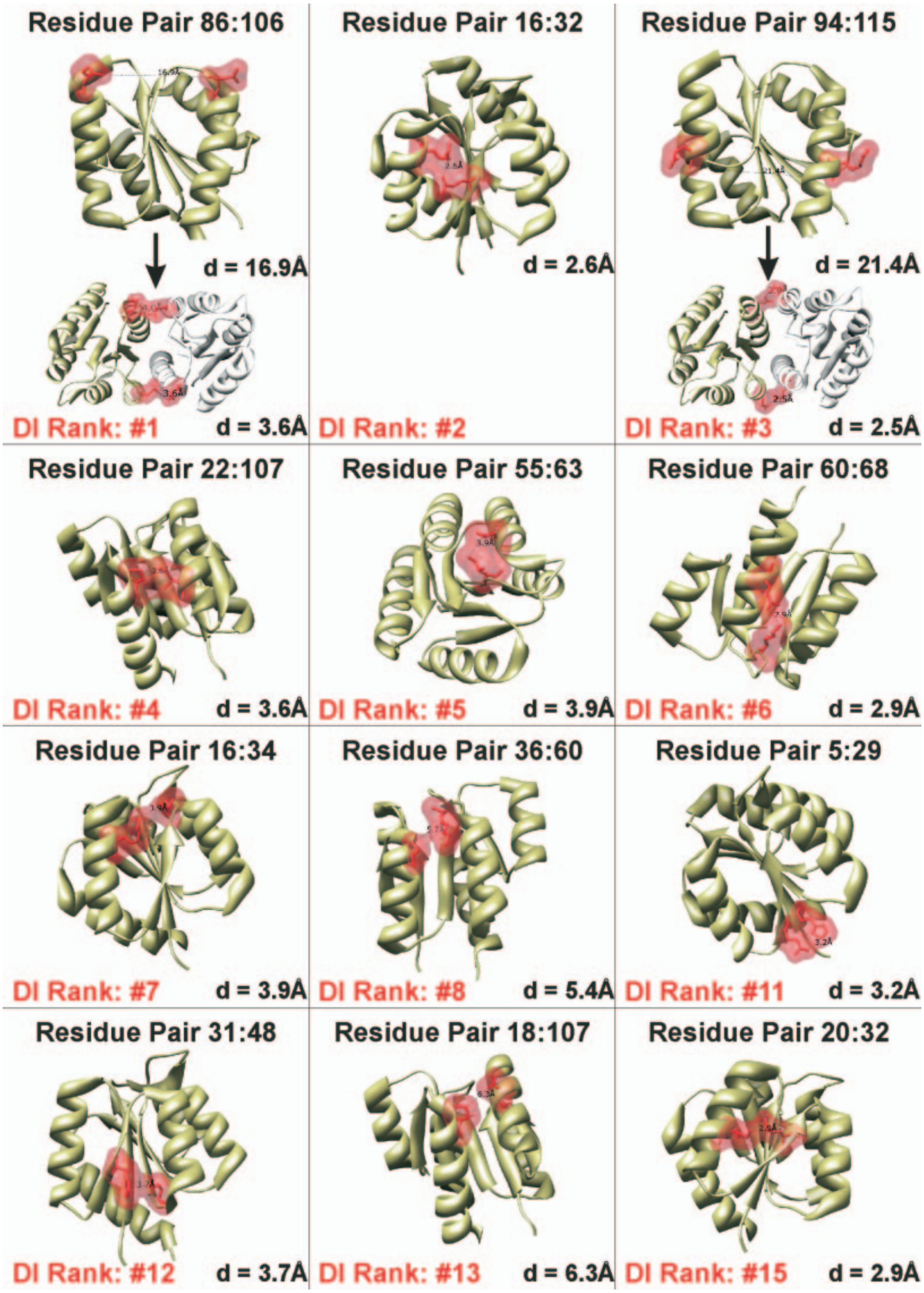


Fig. S6. Direct information detects inter- and intramonomer contact pairings. Positions of the 15 RR/RR residue pairings with the highest DI were mapped onto the ArcA structure, and the minimal distances (*d*) between these pairs were determined in Ångströms. Figures are depicted according to the DI ranking as indicated. High DI residue pairings, which are also within 2 residues in sequence space were omitted from this analysis because of obvious closeness of such pairs (rankings 9, 10, and 14). Two pairings (rankings 8 and 13) show a distance of $>$ 5Å. These, however, are in close proximity in *Escherichia coli* PhoP and *Streptococcus pneumoniae* MicA, indicating that not every contact is made in every RR structure.

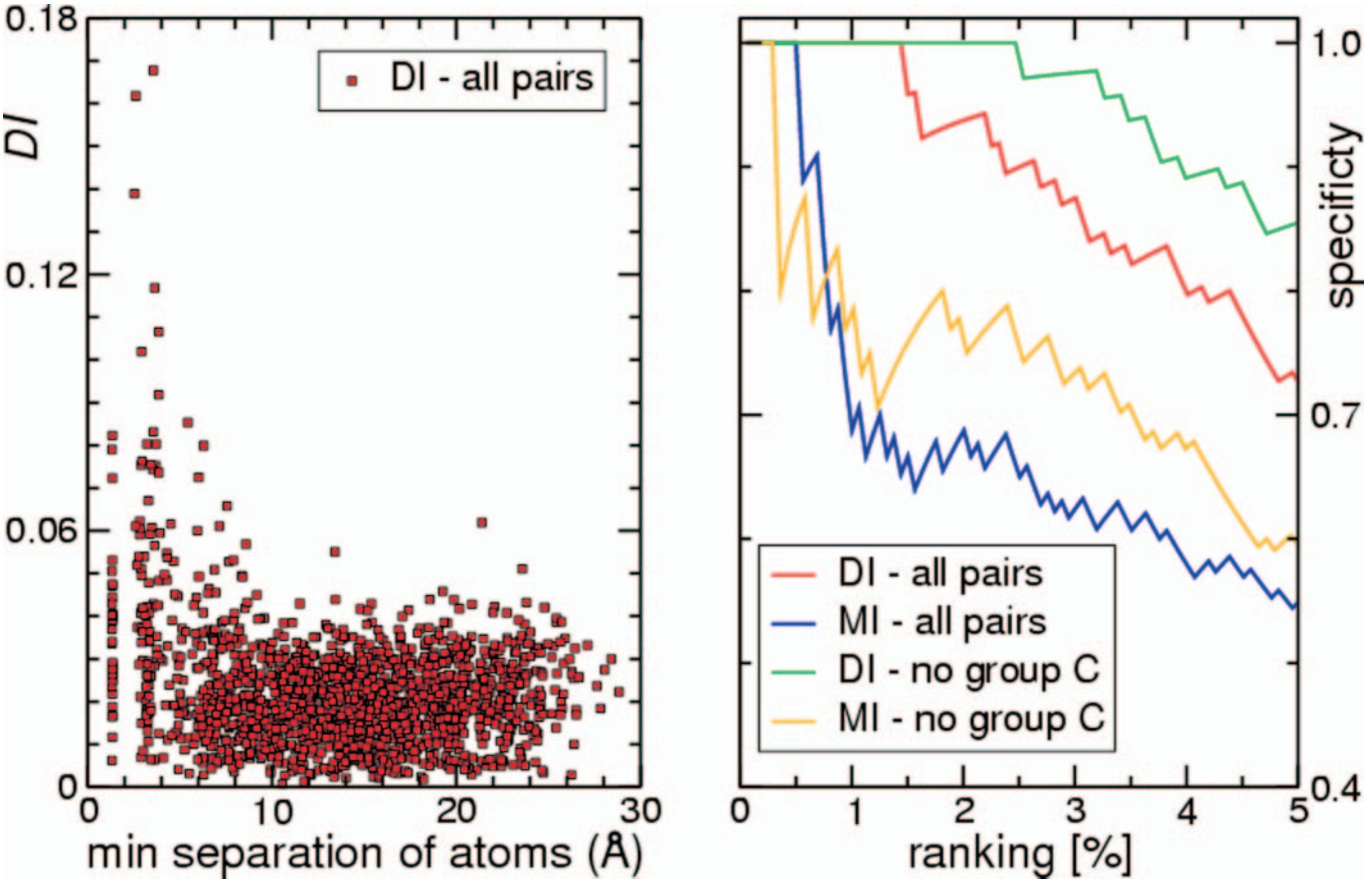


Fig. S7. Direct information (DI) is inversely correlated with residue distance of pairs in the ArcA/PhoP/MicA dimer structures. (*A*) Average minimal distances for all possible RR/RR parings in Ångströms derived from the ArcA, PhoP, and MicA dimer structures (PDB ID codes 1XHE, 2PKX, and 1NXW) were plotted against direct information DI (red diamonds). (*B*) Specificity vs. rank percentile for predicting contact pairs via DI (red curve) and MI (blue curve). Specificity is defined as the fraction of pairings at the given percentile that are within 6 Å in the various RR structures. DI identifies 23, MI only 8 contact pairs before including the first false positive. This simple definition counts group C pairings including positions on the α1-helix as false positives, even if they carry biologically sensible information, compare the caption for Table S1. In the green (DI) and orange curves (MI), positions 14, 18, 21, 22, and 25 of the α1-helix were therefore excluded from the dataset, and ranking was thus restricted to pairings showing correlations because of contacts inside the RR monomer or the RR/RR dimer alone. DI identifies 34 contact pairs at specificity 1.

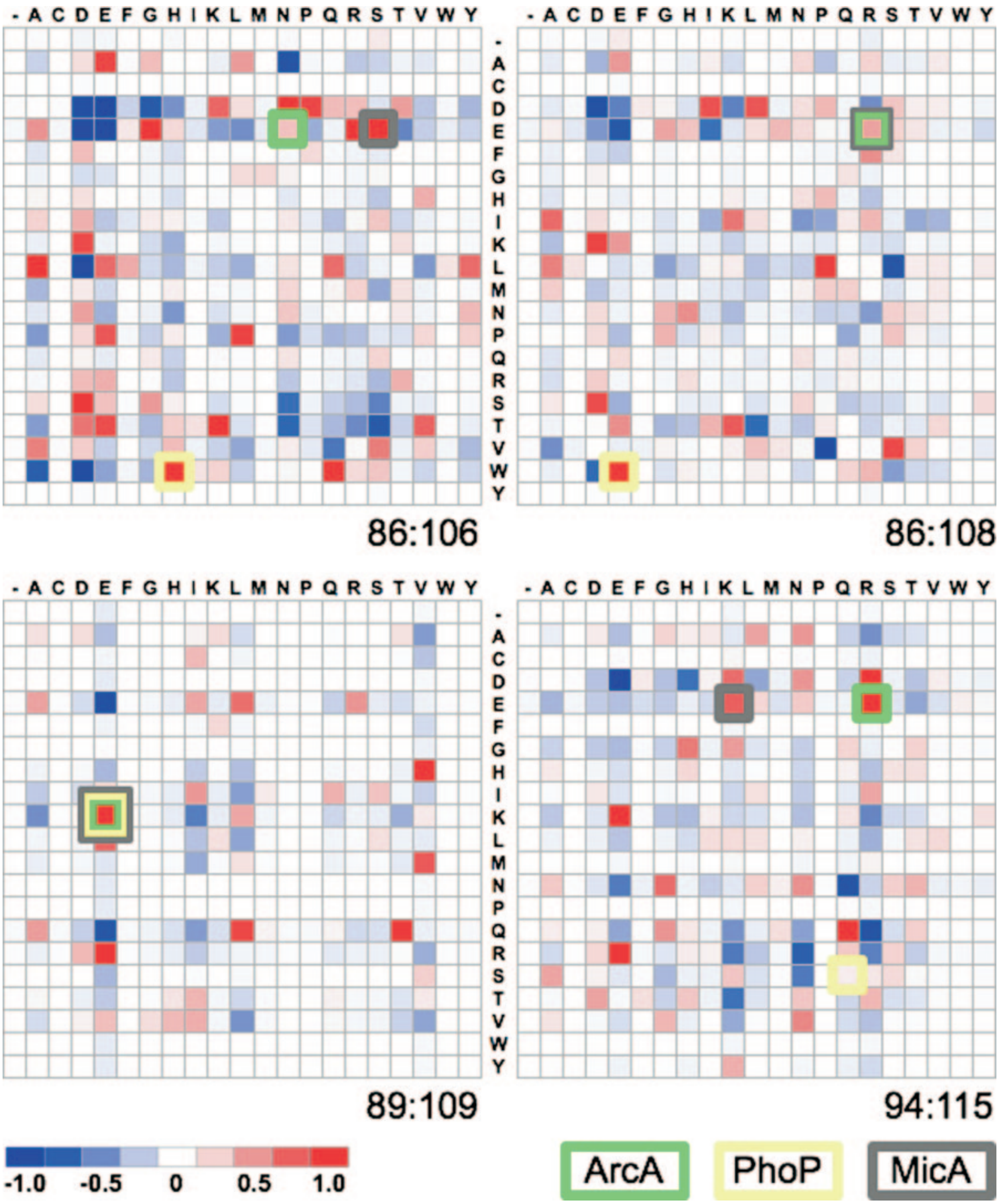


Fig. S8. Correlation matrices reveal dominant pairings for the 4 different dimer contacts. The residue-pair interaction matrices are shown for the 4 intermonomer pairings depicted in Fig. 3. The *x* axis depicts the possible residues in the first position, the *y* axis in the second position. The direct interaction strength $e_{ij}(A_i, A_j)$ is depicted on a scale from dark red for positively coupled pairs over white to dark blue for strongly negative couplings (see scale). Thus, red entries indicate the favored amino acid combinations in positions *i* and *j*, whereas the blue entries indicate the avoided combinations. The actual pairings found at those positions in ArcA, PhoP, and MicA structural examples are framed in green, yellow, and gray, respectively (see also Fig. 3 in the main article). Of note, the 89:109 salt bridge conserved between the 3 structural examples is only one of several common pairings observed at this position. Also, whereas most of the pairings correspond to the red entries, not all 4 contacts are made in all 3 structures. For instance, the Q:S combination observed for pairing 94:115 (ArcA numbering) in PhoP is too distant to support the dimer structure (Fig. 3, main article) and consequently is rarely found at these positions as evident from the very light pink color in the correlation blot.

Other Supporting Information Files

Table S1