Supplementary text: Direct-coupling analysis of residue co-evolution captures native contacts across many protein families

F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, and M. Weigt

I. INPUT DATA

Data are given as a multiple sequence alignment (MSA), i.e. a rectangular array with entries coming from a 21-letter alphabet (20 amino acids, 1 gap):

$$\mathbf{A} = (A_i^a)\,, \quad i = 1, ..., L, \quad a = 1, ..., M \tag{1}$$

with L being the number of residues in each MSA row (the protein length), and M the number of MSA rows (the number of proteins). For simplicity of notation we assume that the $q = 21$ amino acids are translated into consecutive numbers 1,...,q.

II. SEQUENCE STATISTICS

The aim of the analysis is to detect statistical coupling between the amino-acid occupancies of any two columns of the MSA $\mathbf{A}$. For doing so, we first introduce single site and pair frequency counts,

$$f_i(A) = \frac{1}{M}\sum_{a=1}^{M}\delta_{A,A_i^a}; \quad f_{ij}(A,B) = \frac{1}{M}\sum_{a=1}^{M}\delta_{A,A_i^a}\delta_{B,A_j^a}, \tag{2}$$

with $1 \le i,j \le L$, $1 \le A,B \le q$, and δ denoting the Kronecker symbol, which equals one if the two indices coincide, and zero else. The first count determines the fraction of proteins which show amino acid A in column i (residue position), the second one the fraction of MSA rows where amino acids A and B co-appear in positions i and j.

A. Reweighted frequency counts

These simple frequency counts represent faithfully the statistical properties of the MSA if and only if rows are drawn independently from the same distribution. Biological sequence data show a strong sampling bias due phylogenetic relations between species, due to the sequencing of different strains of the same species, and due to a bias in the selection of species which are currently sequenced. As a simple correction, we use a reweighting scheme, which we have introduced in [1, 2].

First, we define a similarity threshold $0 < x < 1$: Two sequences of identity (number of positions with coinciding amino acids) larger than xL are considered to carry almost the same information, smaller sequence identities are considered to carry substantially independent information. In practical tests we have found that values of x around 0.7-0.9 lead to very similar results, we use $x = 0.8$.

Second, for each sequence $A^a = (A_1^a, ..., A_L^a)$ we determine the number of similar sequences $A^b = (A_1^b, ..., A_L^b)$ via

$$m^a = \left|\left\{b \mid 1 \le b \le M,\ \mathrm{seqid}(A^a, A^b) \ge xL\right\}\right| \,. \tag{3}$$

Note that this count is always at least one, since sequence A^a is counted itself in m^a. For each sequence, we use the weight $1/m^a$ in the frequency counts, i.e., sequences without similar sequences take weight one, and sequences featuring similar sequences are down-weighted. We redefine the frequency counts as

$$f_i(A) = \frac{1}{\lambda + M_{eff}}\left(\frac{\lambda}{q} + \sum_{a=1}^{M}\frac{1}{m^a}\delta_{A,A_i^a}\right) \tag{4}$$

$$f_{ij}(A,B) = \frac{1}{\lambda + M_{eff}}\left(\frac{\lambda}{q^2} + \sum_{a=1}^{M}\frac{1}{m^a}\delta_{A,A_i^a}\delta_{B,A_j^a}\right) \,.$$

This equation also contains a pseudo-count λ, which is a standard tool in estimating probabilities from counts in biological sequence analysis [3]. It serves to regularize parameters in the case of insufficient data availability, and has an interpretation in terms of Bayesian inference. The total weight of all sequences, $M_{eff} = \sum_{a=1}^{M} 1/m^a$, can be understood as the effective number of independent sequences.

Note that using $x = 1$ would reweight each sequence by the number of times it appears in the MSA, removing thus simple repeats. Lower values for x aim at giving a smaller weight to regions which are more densely sampled, and a higher weight to regions which are less densely sampled.

B. Mutual information as a correlation measure

If two MSA columns i and j were statistically independent, the joint distribution $f_{ij}(A,B)$ would factorize into $f_i(A) \times f_j(B)$, any deviation from this factorization signals correlations between the columns. Such correlation can be quantified by the mutual information

$$MI_{ij} = \sum_{A,B} f_{ij}(A,B)\ln\frac{f_{ij}(A,B)}{f_i(A)f_j(B)} \,. \tag{5}$$

It equals zero if and only if $f_{ij}(A,B)$ factorizes into the single marginals, and it is positive whenever $f_{ij}(A,B)$ does not factorize.

III. MAXIMUM-ENTROPY MODELING

As discussed in the main text, inter-column correlation may be caused by direct statistical coupling, but

also by indirect correlation effects via intermediate MSA columns. As shown in [1], such direct and indirect effects may be disentangled: The idea is to infer a global statistical model $P(A_1, ..., A_L)$ for entire amino-acid sequences of the protein domain under study. This model has to be coherent to the empirical data, i.e. to generate the empirical single- and two-site frequency counts:

$$\begin{aligned} P_i(A_i) &= \sum_{\{A_k|k\neq i\}} P(A_1, ..., A_L) = f_i(A_i) \qquad (6) \\ P_{ij}(A_i, A_j) &= \sum_{\{A_k|k\neq i,j\}} P(A_1, ..., A_L) = f_{ij}(A_i, A_j) \ . \end{aligned}$$

Beyond these constraints, we aim at the most general, i.e. least constrained model $P(A_1, ..., A_L)$. It can be determined using the distribution maximizing the entropy

$$S = -\sum_{\{A_i|i=1,...,L\}} P(A_1, ..., A_L) \ln P(A_1, ..., A_L) \qquad (7)$$

while satisfying the constraints in Eqs. (6). The solution to this optimization problem is standard [4]: after introducing constraints via Lagrange multipliers, we find the analytical form of the distribution:

$$P(A_1, ..., A_L) = \frac{1}{Z} \exp\left\{\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i)\right\} \ . \qquad (8)$$

The Lagrange multipliers $h_i(A)$ and $e_{ij}(A, B)$ have a simple interpretation in terms of local amino-acid biases (local fields in statistical-physics language) and statistical residue couplings (coupling strength in statistical-physics language). Their numerical values have to be tuned such that the constraints given by Eqs. (6) are respected. The normalization constant

$$Z = \sum_{\{A_i|i=1,...,L\}} \exp\left\{\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i)\right\} \qquad (9)$$

is called *partition function* in statistical physics. For later convenience, we also introduce the *Hamiltonian*

$$\mathcal{H} = -\sum_{1\leq i<j\leq L} e_{ij}(A_i, A_j) - \sum_{i=1}^{L} h_i(A_i) \ , \qquad (10)$$

such that our probabilistic model reads $P(A_1, ..., A_L) = \exp\{-\mathcal{H}\}/Z$.

The major problem in this context is the determination of the marginal distributions $P_i(A)$ and $P_{ij}(A, B)$ from $P(A_1, ..., A_L)$. Doing this exactly by tracing over all other variables A_i as written in Eqs. (6) would require an exponential time, which grows like q^L with the length of the aligned proteins. Different strategies have already been suggested for tackling this problem (most of them for the restricted Ising model having $q = 2$): In [1] we used a message-passing algorithm originally proposed in [5], [6] uses improved Monte Carlo sampling, [7–9] suggest perturbative expansion schemes, whereas [10] uses pseudo-likelihoods decoupling inference for different sites. For an overview over the relative performance of these algorithms on artificial data see [11].

It is important to note that the partition function itself contains all necessary information on the marginals, in particular we have

$$\begin{aligned} \frac{\partial \ln Z}{\partial h_i(A)} &= -P_i(A) \\ \frac{\partial^2 \ln Z}{\partial h_i(A)\, \partial h_j(B)} &= -P_{ij}(A, B) + P_i(A)\, P_j(B) \ . \qquad (11) \end{aligned}$$

For later convenience we introduce the connected correlations

$$C_{ij}(A, B) = P_{ij}(A, B) - P_i(A)\, P_j(B) \ , \qquad (12)$$

where indices i, j run from $1, ..., L$, whereas A, B from $1, .., q-1$. The significance of excluding $A, B = q$ will become clear below. Note that we will consider $C_{ij}(A, B)$ as a $L(q-1)\times L(q-1)$-dimensional matrix, i.e. each pair (i, A) is interpreted as a parametrization of a single, joint index.

A. The number of independent parameters

The statistical model in Eq. (8) has $\binom{N}{2}q^2 + Nq$ parameters, but not all of them are independent. In fact, the consistency conditions in Eqs. (6) are also not independent, since the single-site marginals are implied by the two-site marginals, and all distributions are normalized. Careful inspections unveils $\binom{N}{2}(q-1)^2 + N(q-1)$ independent consistency conditions. We may therefore fix a part of the parameters in Eq. (8). Without loss of generality, we set

$$e_{ij}(A, q) = e_{ij}(q, A) = h_i(q) = 0 \qquad (13)$$

for all $i, j = 1, .., L$ and $A = 1, ...q$. Intuitively, this corresponds to a situation where all couplings and biases are measured with respect to the state q. The number of remaining parameters matches now the number of constraints, and the solution of the maximum-entropy model is unique.

B. Small-coupling expansion

The algorithmic approach is based on a systematic small-coupling expansion, i.e., on a Taylor expansion around zero coupling. This expansion was introduced in [12] by Plefka for disordered Ising models (Ising spin-glasses, corresponding to binary variables with $q = 2$). A more elegant derivation was proposed Georges and Yedidia [13], we generalize their approach to the case of Potts models with $q > 2$.

First we introduce the perturbed Hamiltonian

$$\mathcal{H}(\alpha) = -\alpha \sum_{1\leq i<j\leq L} e_{ij}(A_i, A_j) - \sum_{i=1}^{L} h_i(A_i) \ , \qquad (14)$$

depending on the additional parameter α. This parameter allows to interpolate between independent variables for $\alpha = 0$, and the original model for $\alpha = 1$. Furthermore we introduce the so-called *Gibbs potential*

$$-\mathcal{G}(\alpha) = \ln\left[\sum_{\{A_i|i=1,\dots,L\}} e^{-\mathcal{H}(\alpha)}\right] - \sum_{i=1}^{L}\sum_{B=1}^{q-1} h_i(B)P_i(B) \tag{15}$$

as the Legendre transform of the *free energy* $\mathcal{F} = -\ln Z$. Whereas the free energy depends canonically on the couplings and the fields, the Gibbs potential depends on the couplings and the marginal single-site distributions $P_i(A)$, i.e.

$$\mathcal{G}(\alpha) = \mathcal{G}\left(\{\alpha e_{ij}(A,B)\}_{1\le i<j\le L}^{A,B=1,..,q-1}, \{P_i(A)\}_{i=1,\dots,L}^{A=1,\dots,q-1}\right). \tag{16}$$

This choice is particularly practical for the following derivation, since it guarantees the first of Eqs. (6) to be valid at any α. Note that the Potts variables in this expression run only up to $q-1$. Due to the gauge of the couplings and the normalization of the marginals, values for $A, B = q$ are not independent variables.

The fields can be found via the standard expression for Legendre transforms, cf. Eq. (11),

$$h_i(A) = \frac{\partial \mathcal{G}(\alpha)}{\partial P_i(A)} \ , \tag{17}$$

and

$$\left(C^{-1}\right)_{ij}(A,B) = \frac{\partial h_i(A)}{\partial P_j(B)} = \frac{\partial^2 \mathcal{G}(\alpha)}{\partial P_i(A)\,\partial P_j(B)} \ . \tag{18}$$

It is worth pointing out that the previous relations hold at any value of α and are a consequence of the functional form of the Legendre transform defined in Eq. (15). We remind that the matrix C was defined in Eq. (12) to have dimension $L(q-1)$, i.e. Potts-state indices are constrained to values up to $q-1$. This restriction makes C an invertible matrix (at least for non-zero pseudo-count λ), removing trivial linear dependencies resulting from the normalization of P_{ij}. Using this last equation, we can calculate the two-point marginal distributions P_{ij} directly from the Gibbs potential by means of two partial derivations and one matrix inversion.

Our aim is to expand this Gibbs potential up to first order in α around the independent-site case $\alpha = 0$,

$$\mathcal{G}(\alpha) = \mathcal{G}(0) + \left.\frac{d\mathcal{G}(\alpha)}{d\alpha}\right|_{\alpha=0} \alpha + \mathcal{O}(\alpha^2) \ . \tag{19}$$

In the following subsections, we calculate the still unknown terms on the right-hand side of this equations, i.e. the Gibbs potential and its first derivative in $\alpha = 0$.

C. Independent-site approximation

To start with, let us consider the Gibbs potential in $\alpha = 0$. In this case, the Gibbs potential equals the negative entropy of an ensemble of L uncoupled Potts spins $A_1, ..., A_L$ of given marginals $P_i(A_i)$. This claim results from basic statistical mechanics: The free energy equals the average energy (average Hamiltonian) minus the entropy. For $\alpha = 0$, the Legendre transform removes the complete average energy.

However, the entropy of uncoupled spins of given distribution is known to be

$$\begin{aligned}
\mathcal{G}(0) &= \sum_{i=1}^{L}\sum_{A=1}^{q} P_i(A)\ln P_i(A) \\
&= \sum_{i=1}^{L}\sum_{A=1}^{q-1} P_i(A)\ln P_i(A) \\
&\quad + \sum_{i=1}^{L}\left[1 - \sum_{A=1}^{q-1} P_i(A)\right]\ln\left[1 - \sum_{A=1}^{q-1} P_i(A)\right] \ ;
\end{aligned} \tag{20}$$

the last line eliminates terms in $P_i(q)$ and reduces the expression to the independent variables.

D. Mean-field approximation

To get the first order in Eq. (19), we have to determine $d\mathcal{G}(\alpha)/d\alpha$ in $\alpha = 0$. Recalling the definition of the Gibbs potential in Eq. (15), we write

$$\begin{aligned}
\frac{d\mathcal{G}(\alpha)}{d\alpha} &= -\frac{d}{d\alpha}\ln Z(\alpha) - \sum_{i=1}^{L}\sum_{A=1}^{q-1}\frac{dh_i(A)}{d\alpha}P_i(A) \\
&= -\sum_{\{A_i\}}\left[\sum_{i<j} e_{ij}(A_i,A_j) + \sum_i \frac{dh_i(A)}{d\alpha}\right]\frac{e^{-\mathcal{H}(\alpha)}}{Z(\alpha)} \\
&\quad - \sum_{i=1}^{L}\sum_{A=1}^{q-1}\frac{dh_i(A)}{d\alpha}P_i(A) \\
&= -\left\langle \sum_{i<j} e_{ij}(A_i,A_j)\right\rangle_\alpha \ .
\end{aligned} \tag{21}$$

The first derivative of the Gibbs potential with respect to α equals thus the average of the coupling term in the Hamiltonian. At $\alpha = 0$, this average can be done easily, since the joint distribution of all variables becomes factorized over the single sites,

$$\left.\frac{d\mathcal{G}(\alpha)}{d\alpha}\right|_{\alpha=0} = -\sum_{i<j}\sum_{A,B} e_{ij}(A,B)\,P_i(A)\,P_j(B) \ . \tag{22}$$

Plugging this and Eq. (20) into Eq. (19), we find the first-order approximation of the Gibbs potential. First and second partial derivatives with respect to the marginal distributions $P_i(A)$ provide self-consistent equations for the local fields,

$$\frac{P_i(A)}{P_i(q)} = \exp\left\{h_i(A) + \sum_{\{j|j\neq i\}}\sum_{B=1}^{q-1} e_{ij}(A,B)\,P_j(B)\right\} \tag{23}$$

and the inverse of the connected correlation matrix,

$$\left(C^{-1}\right)_{ij}(A,B)\Big|_{\alpha=0} = \begin{cases} -e_{ij}(A,B) & \text{for } i \neq j \\ \frac{\delta_{A,B}}{P_i(A)} + \frac{1}{P_i(q)} & \text{for } i = j \end{cases} . \quad (24)$$

This last equation allows for solving the original inference problem in mean-field approximation in a single step, without resorting to iterative schemes like gradient decent. Since we want to fit one- and two-site marginal of $P(A_1, ..., A_L)$ to the empirical values $f_i(A)$ and $f_{ij}(A,B)$ derived from the original protein MSA, we just need to determine the empirical connected correlation matrix

$$C_{ij}^{(emp)}(A,B) = f_{ij}(A,B) - f_i(A)\, f_j(B) \quad (25)$$

and invert this matrix to get the couplings e_{ij}. Even if matrix inversion is of complexity $\mathcal{O}(L^3)$ and thus of the same complexity as susceptibility propagation, the mean-field approximation is found to be $10^3 - 10^4$ times faster. This results from the simple fact that $> 10^3$ iteration are needed in susceptibility propagation to reach sufficient precision in fitting the empirical data by the maximum-entropy model.

IV. DIRECT INFORMATION AS A DIRECT-COUPLING MEASURE

Given the estimate of the pair couplings $e_{ij}(A,B)$ we would like to rank residue pairs according to their interaction strength. To do so, we need a meaningful mapping from the $(q-1)\times(q-1)$-dimensional coupling matrices to a single scalar parameter. A way to do this which is independent of the selected gauge, was already proposed in [1]. The quantity introduced there was called *direct information* (DI) and measures the mutual information due to the direct coupling. To do so, we isolate a pair i, j of positions and introduce a two-site model

$$P_{ij}^{(dir)}(A,B) = \frac{1}{Z_{ij}} \exp\left\{ e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B) \right\} \quad (26)$$

with the coupling being the one inferred before. The new fields $\tilde{h}_{i/j}$ are determined by imposing the empirical single-site frequency counts as marginal distributions,

$$\begin{aligned} f_i(A) &= \sum_{B=1}^{q} P_{ij}^{(dir)}(A,B) \\ f_j(B) &= \sum_{A=1}^{q} P_{ij}^{(dir)}(A,B) \ , \end{aligned} \quad (27)$$

and Z_{ij} follows by normalization. The direct information is the mutual information associated to $P_{ij}^{(dir)}$:

$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{(dir)}(A,B) \ln \frac{P_{ij}^{(dir)}(A,B)}{f_i(A)\, f_j(B)} . \quad (28)$$

In this expression, any indirect effect is obviously removed, only the strength of the direct coupling $e_{ij}(A,B)$ is measured.

[1] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[2] A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, and M. Weigt. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: Orphans and crosstalks. *PLoS ONE*, 6(5):e19729, 05 2011.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[4] D.J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge Univ Pr, 2003.

[5] M. Mézard and T. Mora. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris*, 103(1-2):107 – 113, 2009. Neuromathematics of Vision.

[6] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[7] S. Cocco, S. Leibler, and R. Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences*, 106(33):14058–14062, 2009.

[8] V. Sessak and R. Monasson. Small-correlation expansions for the inverse Ising problem. *Journal of Physics A: Mathematical and Theoretical*, 42:055001, 2009.

[9] H. Kappen and F.B. Rodriguez. Efficient learning in boltzmann machines using linear response theory. *Neural Computation*, 10:1137, 1998.

[10] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *Annals of Statistics*, 38:1287, 2010.

[11] Y. Roudi, J.A. Hertz, and E. Aurell. Statistical physics of pairwise probability models. *Front. Comput. Neurosci.*, 3:22, 2009.

[12] T. Plefka. Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971, 1982.

[13] A. Georges and J.S. Yedidia. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9):2173, 1991.

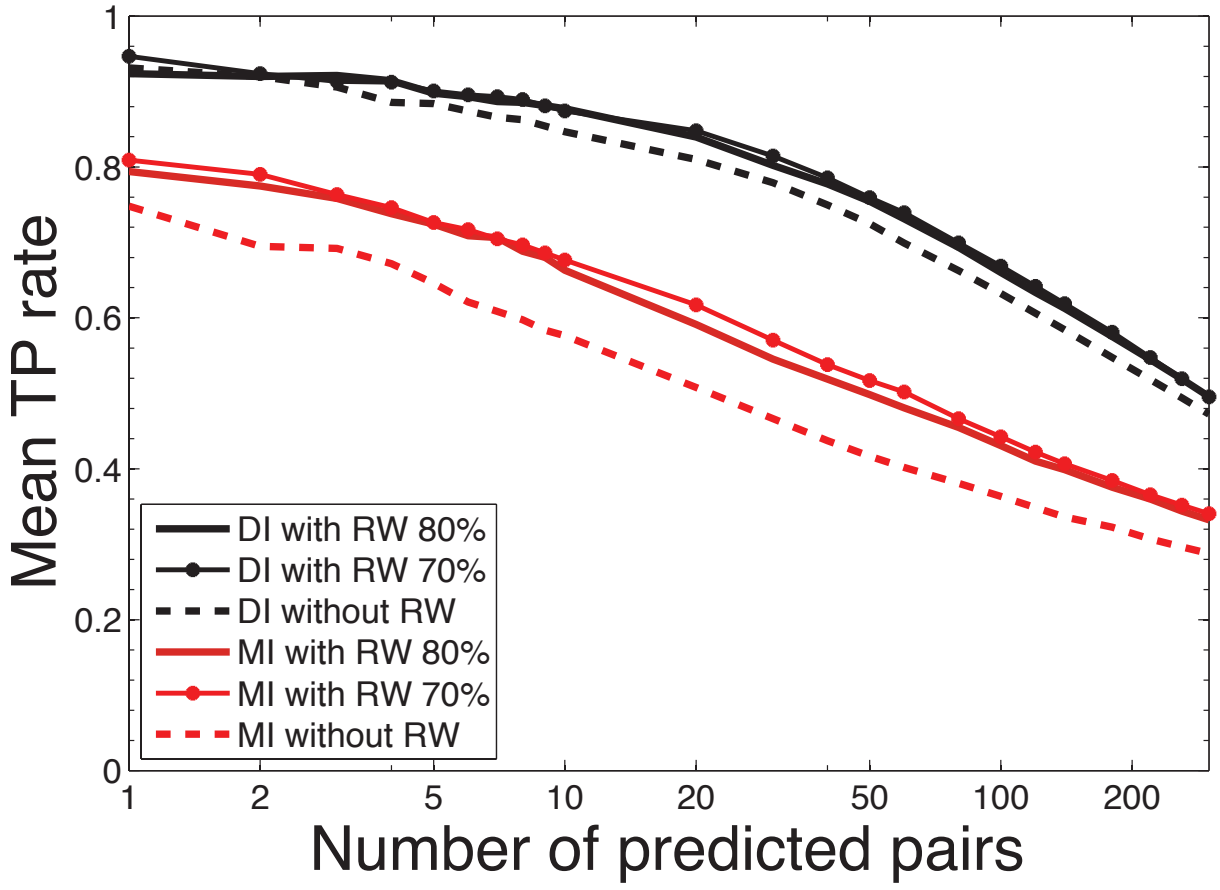

Figure S1. Mean prediction performance for 131 domain families with respect to the top number of ranked contacts. The effect of sampling correction by re-weighting (RW), i.e. clustering redundant sequences for > 80% identity is beneficial for both MI and DI methods. Results with sampling correction (solid lines) are always better than their counterparts without re-weighting (dashed lines). Using a different threshold e.g, from 80% to 70% does not have a significant influence on the mean TP performance.

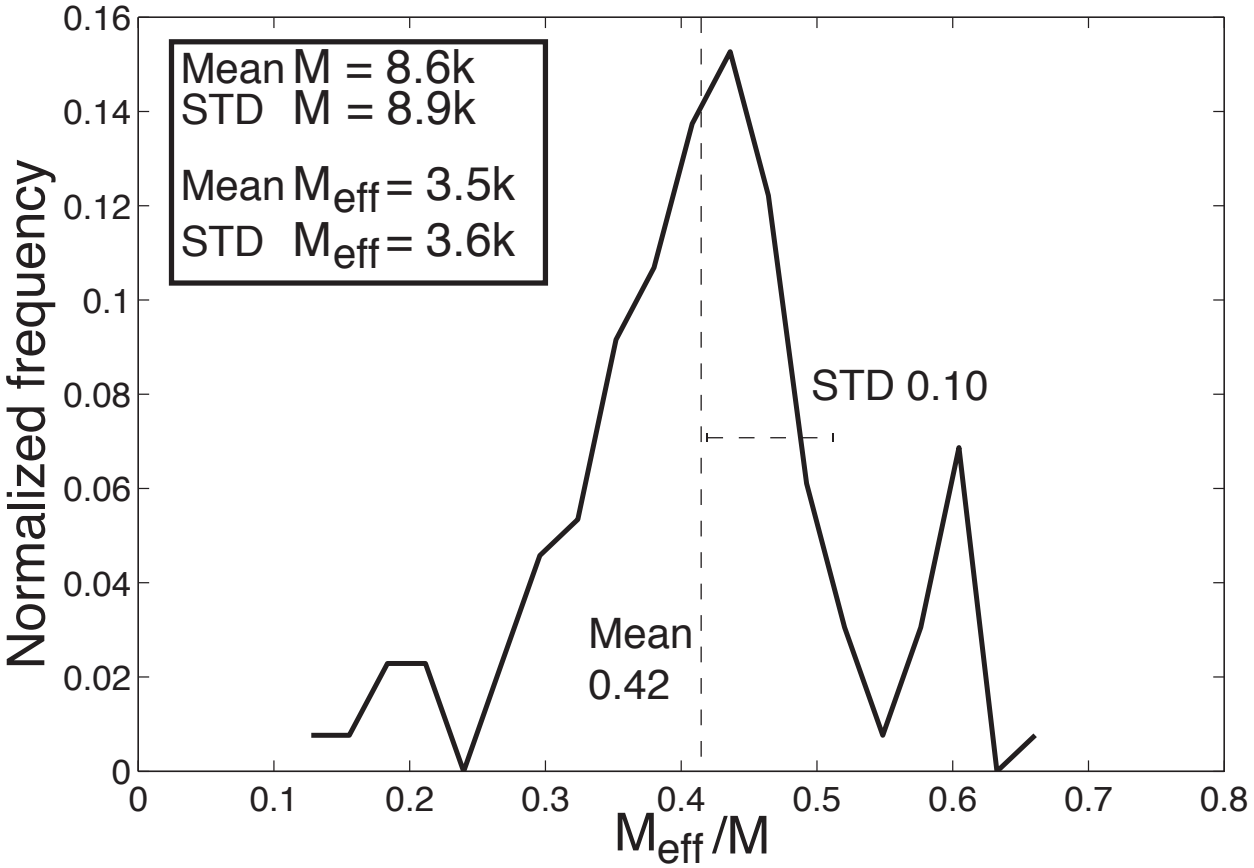

Figure S2. Distribution of the ratio Meff /M for the dataset of 131 domain families used in this study. MSA for all these families have a mean value of 8,600 sequences with a mean of 3,600 effective sequences.

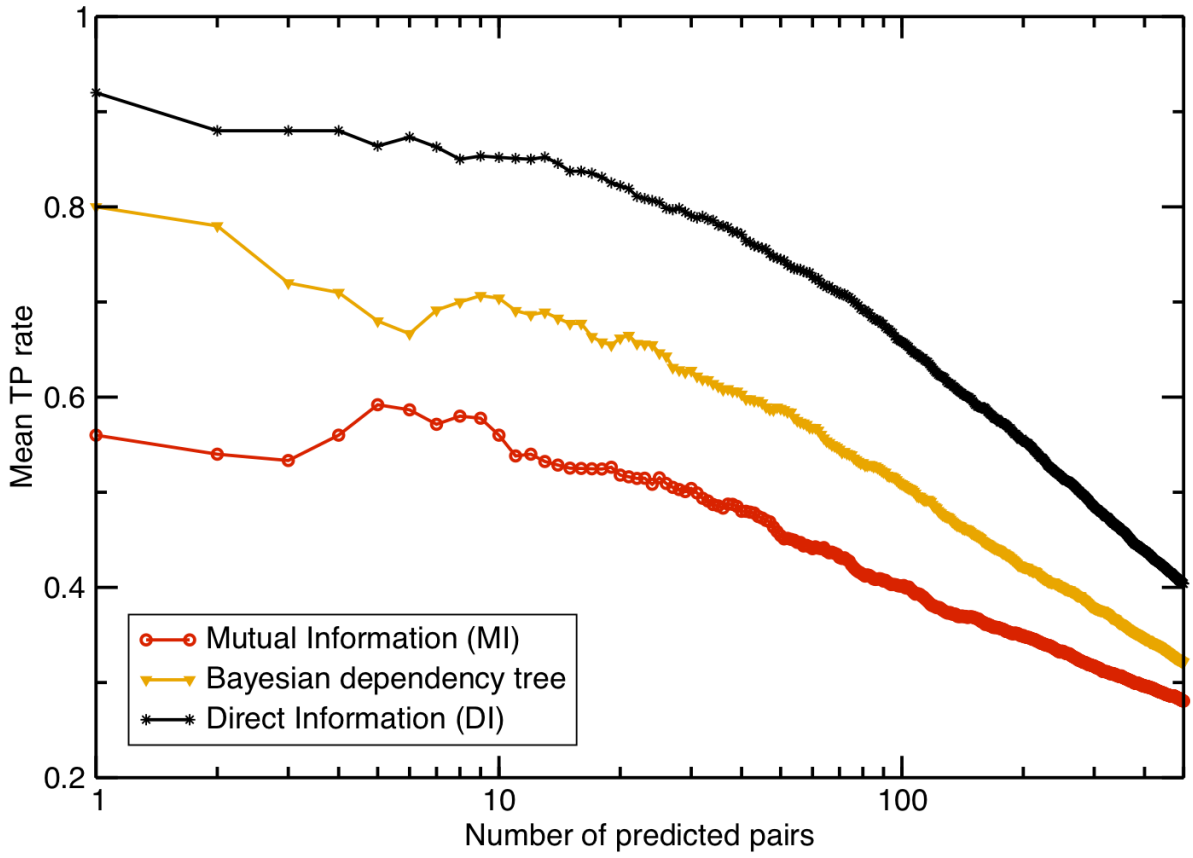

Figure S3. Mean prediction performance for 25 eukaryotic domain families with more than 2000 sequences. The figure shows equivalent results as the ones obtained for bacterial sequences (Fig. 2A and Fig. S5). This suggests that the applicability of DI-based predictions to eukayotic is plausible.

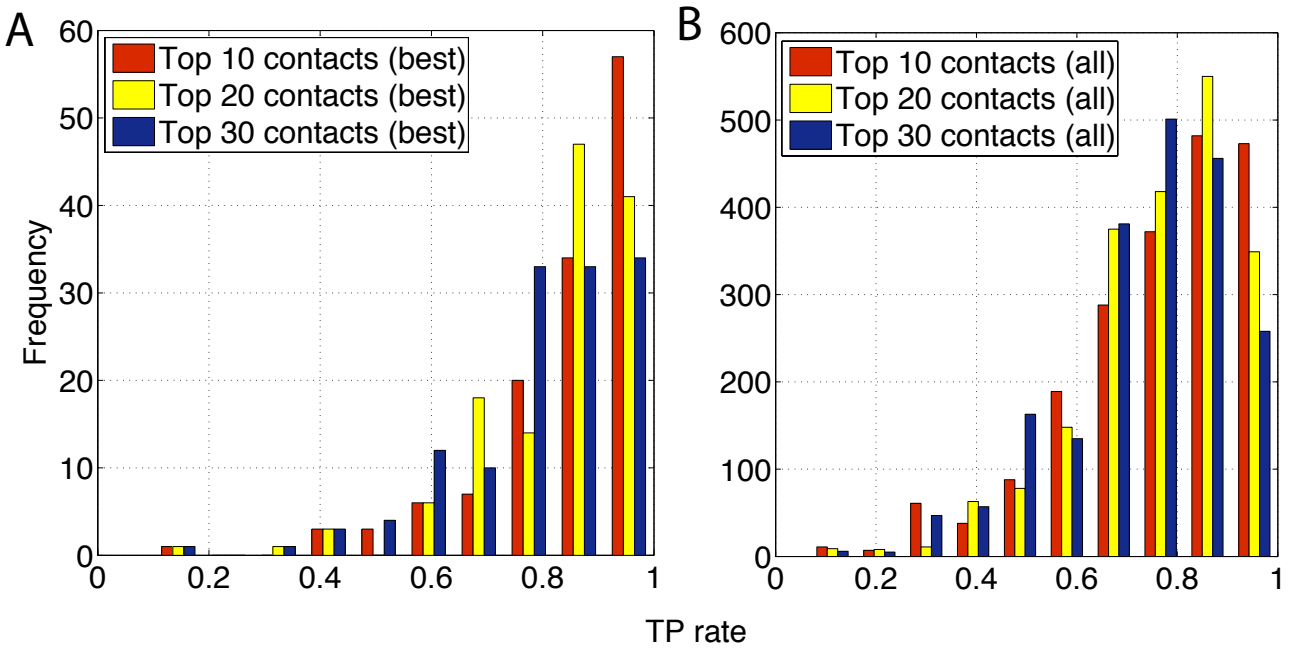

Figure S4. A) Distribution of TP rates for the 131 domains studied and computed with the best predicted structures per domain using mfDCA with sampling correction. Results are shown for the top 10,20 and 30 predicted pairs. B) Distribution of TP rates for the 131 domains studied and all PDB structures using mfDCA and sampling correction. Top 10,20 and 30 pairs seem to have a peak of the TP rate distribution around 0.8-0.9.

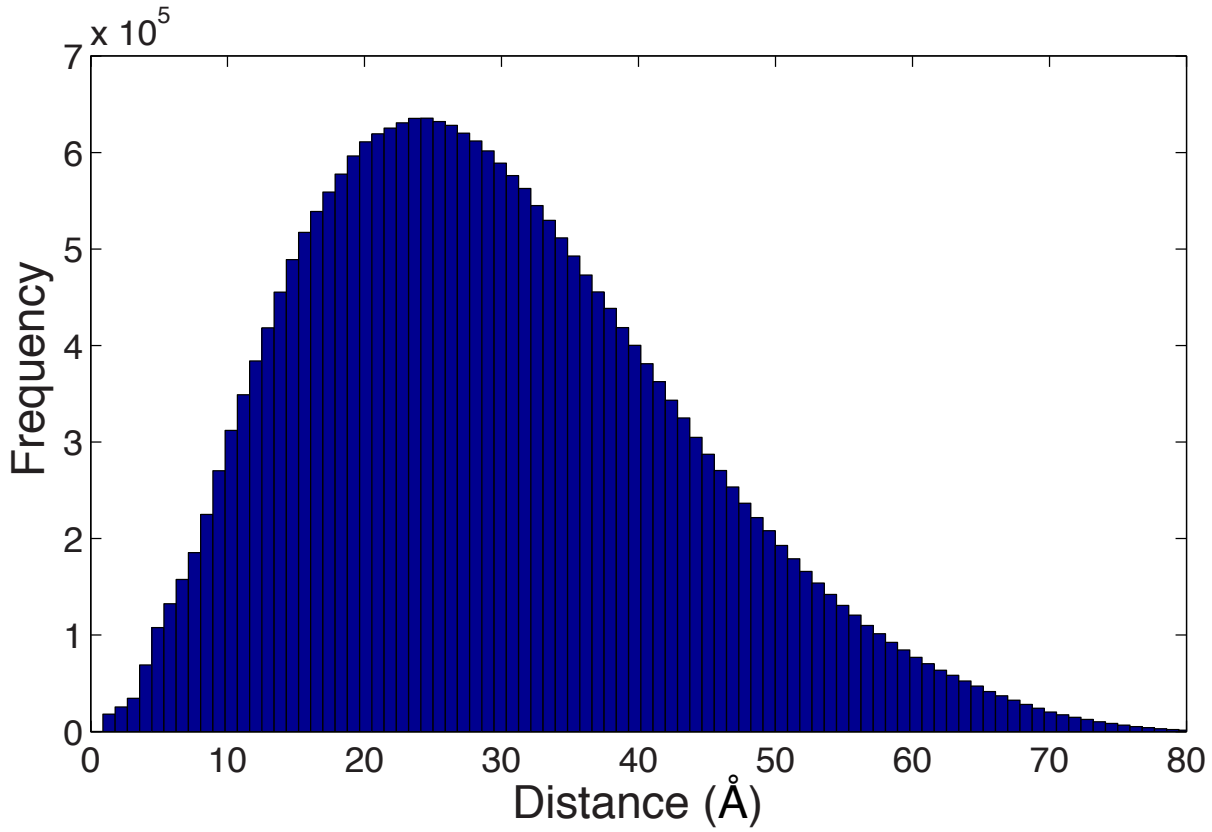

Figure S5. Histogram of all background pairwise atomic distances for 10 random PDB structures in our dataset. The peak of the distribution around 25 Å explains a small bump observed in Figure 2B near the same distance (20-25 Å) in the distribution.

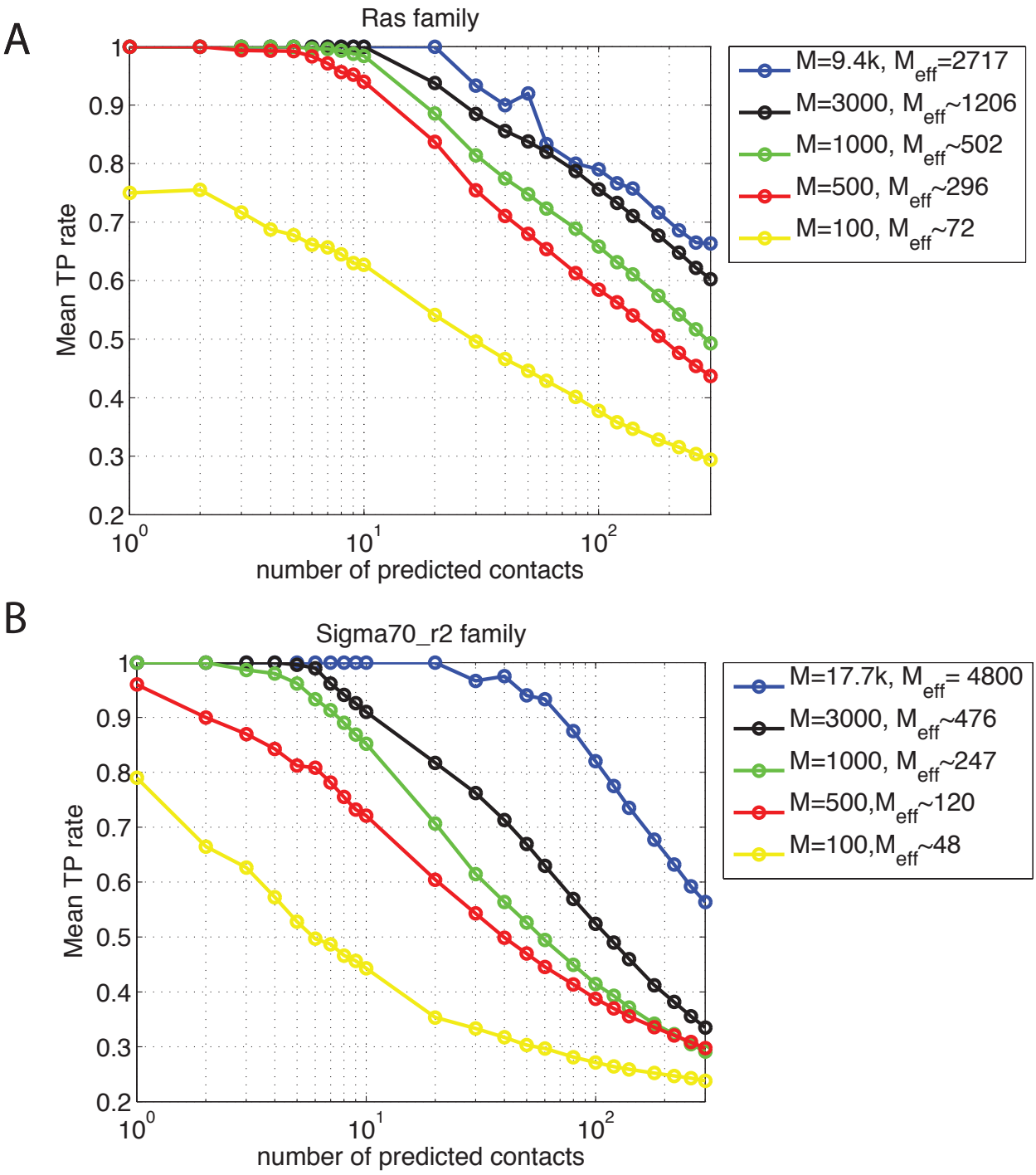

Figure S6. Sensitivity analysis of the performance of mfDCA for random sub-alignments of different lengths. Results are shown for two domain families: (A) the Ras domain family (PF00071) and (B) the DNA-recognition domain (Region 2) of the bacterial Sigma-70 factor (Pfam ID PF04542) were selected to assess prediction performance for sequence alignments of size M=100, 500, 1000 and 3000, corresponding to Meff values ranging from 72 to 1206. Curves are averaged over 100 randomly generated sub-alignments fore each M. A number of Meff ~ 250 appears to be necessary to get sensitive results, while using Meff ~ 1000 reaches results similar to the ones using full alignments.

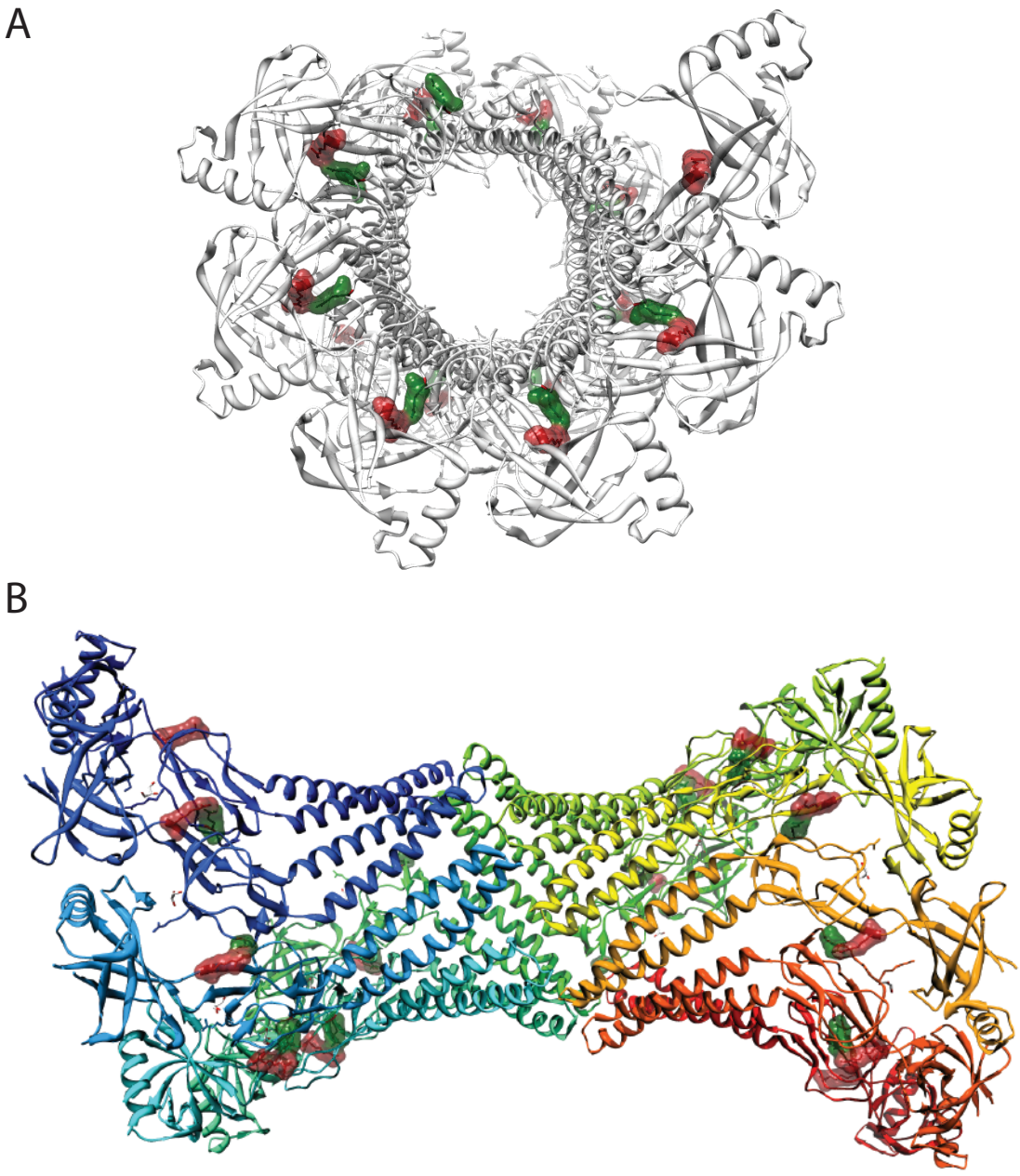

Figure S7. A) Protein MexA (PDB ID 1vf7), showing nine secretion and transporter activity domains HlyD domains (PF00529) forming a funnel like structure used as antibiotic efflux. One of two false positives in the top 20 predictions was a multimerization couplet, shown in green and red. B) Side view of the complex with domains in different colors.

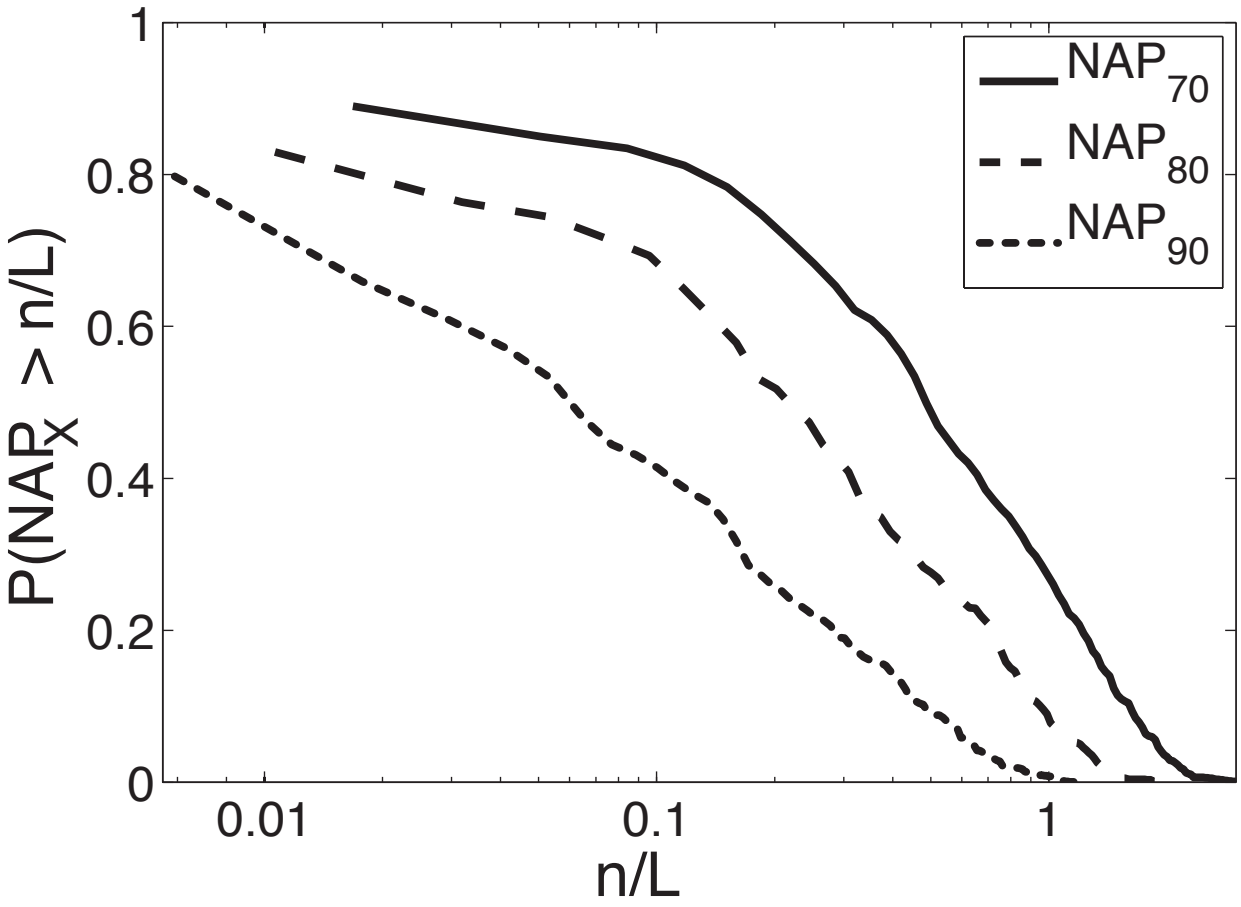

Figure S8. Cumulative distribution of the Number of Acceptable Pairs (NAPx) for a given TP rate x normalized by the length of the domain L. The curves show the probability of NAPx to be larger than a given number n for contacts at given TP rates of 0.9, 0.8 and 0.7. The curves are computed for all 856 PDB structures in the dataset.

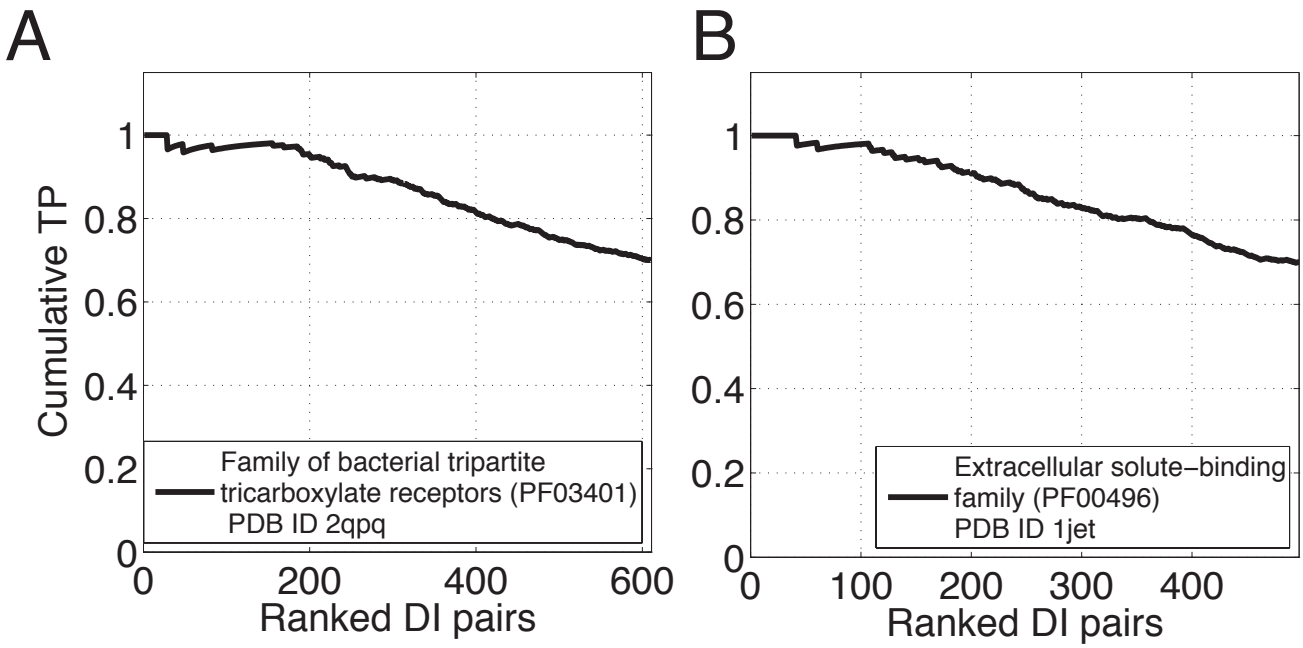

Figure S9. A) Family of bacterial tripartite tricarboxylate receptors (PF03401), NAP70 is 600, i.e.,70% of the top 600 DI pairs correspond to true contacts when mapped to structure PDB ID 2qpq. B) The extracellular solute-binding family (PF00496) mapped to the structure of the periplasmic oligopeptide-binding protein OppA of S. typhimurium (PDB ID 1jet) has a NAP70 of 497. Approximately 350 contacts are true positives.

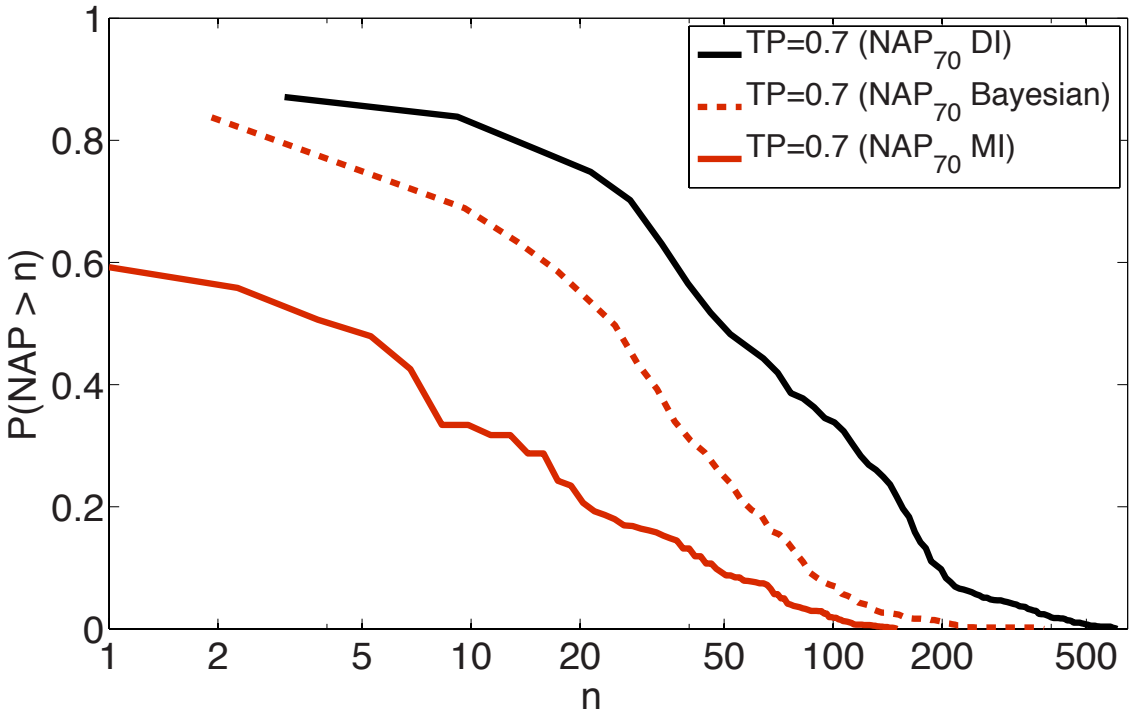

Figure S10. Comparison of the probability function of the Number of Accepted Pairs (NAP70) to be larger than a certain number of pairs for three methods: DI, Bayesian approach and MI. DI shows a clear improvement against MI (red curve) and the Bayesian approach by Burger et al. (dashed red) which becomes more evident as NAP grows larger.

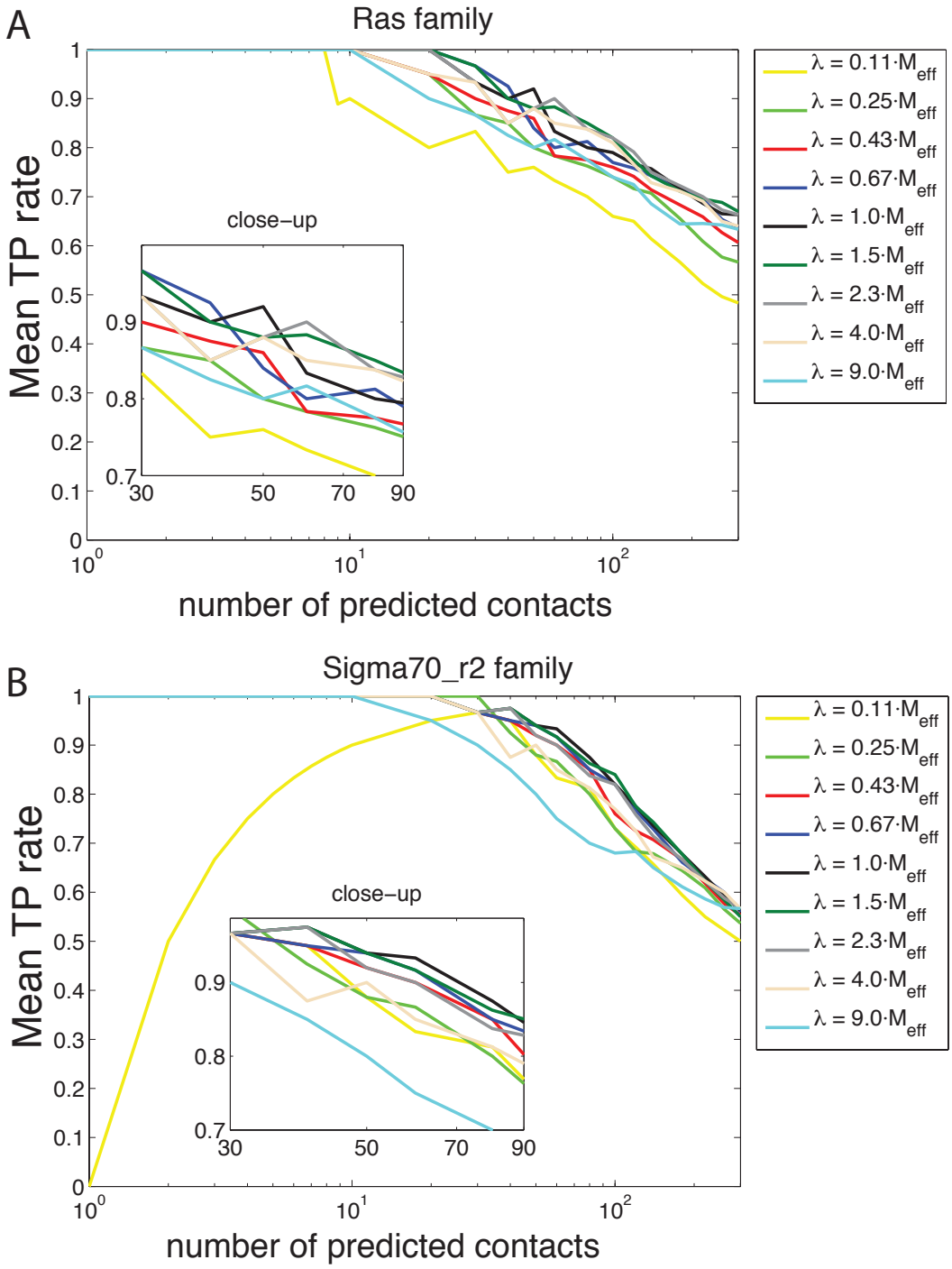

Figure S11. Performance of mfDCA for different values of the pseudocount parameter λ . Mean TP rates are shown for two domain families (A) the Ras domain family (PF00071) and (B) the DNA-recognition domain (Region 2) of the bacterial Sigma-70 factor (Pfam ID PF04542). The pseudo-count values used depend on the number of effective sequences Meff and a weighting parameter, pseudo-count weight w as λ = w Meff. Mean TP rates are computed for different w values between 0.11 and 9. A relatively small variance in performance for values of w > 0.5 is observed with the optimum between 1-1.5. λ = Meff was used as a fixed parameter in all the results shown in this study.

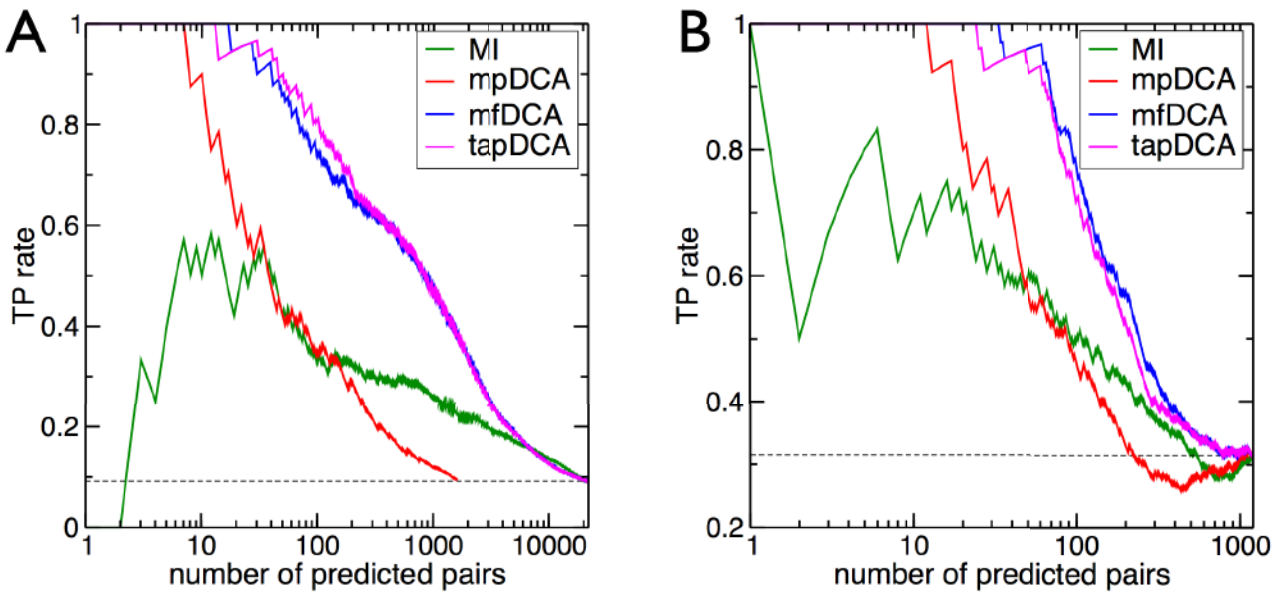

Figure S12. Comparison of different DCA approximations for (A) Trypsin (PF00089, PDB 3TGI) and (B) Trypsin inhibitor (PF00014, PDB 5PTI). Whereas all DCA algorithms outperform the contact prediction by mutual information (green line), we find the new mfDCA (blue line) to be superior to the previous mpDCA (red line). Going beyond mfDCA to the next order of the smallcoupling expansion (tapDCA, pink line), cf. Methods, does not systematically improve over mfDCA, but leads to a substantially slower algorithm. The fact that the red curve in panel A finishes at a smaller number of pairs results from the fact, that mpDCA can be run only on subalignments of up to 70 columns due to the algorithmic complexity of the approach.

Table S1. List of PDB structures analyzed in this study.

PDB IDs									
153l	1gbs	1lqp	1qgs	1vz0	2bkn	2gd9	2oqg	2z1e	3e10
154l	1gdt	1lr0	1qhg	1w55	2bko	2gj3	2oqr	2z1f	3e38
1a04	1gg4	1ls9	1qhh	1w6s	2bkp	2gjg	2oxo	2z1u	3e4r
1a0b	1gqy	1lsp	1qks	1w77	2bm4	2gkg	2oyo	2z2l	3e4v
1a0p	1gu9	1lss	1qpz	1w78	2bm5	2glk	2p19	2z2m	3e7l
1ae9	1gug	1luc	1qsa	1w8i	2bm6	2gm5	2p4g	2z4g	3e8o
1al3	1gun	1lvw	1qte	1wet	2bm7	2gms	2p5v	2z4p	3eag
1atg	1gus	1m65	1qtw	1wmi	2bnm	2gmy	2p7o	2z6r	3ec2
1b7e	1gut	1m68	1qu7	1woq	2brc	2gqp	2paq	2z8x	3ecc
1b9m	1h3l	1m6k	1qwy	1wp1	2byi	2gsk	2pbq	2z98	3ech
1b9n	1h4i	1m70	1qxx	1wpm	2c2a	2gu1	2pfx	2z9b	3ecp
1bia	1h7l	1m7j	1r1m	1wpn	2c81	2guf	2ph1	2zau	3edp
1bib	1h7q	1ma7	1r1t	1wpp	2ce0	2guh	2pjr	2zbc	3eet
1bl0	1h8z	1mb3	1r1u	1ws6	2cg4	2gup	2pkh	2zc3	3efm
1boo	1h98	1mdo	1r23	1x74	2ch7	2gxg	2pmh	2zc4	3eiw
1bsl	1h9g	1mkm	1r62	1x9h	2cvi	2gza	2pn6	2zcm	3eix
1byi	1h9j	1mkz	1r8d	1x9i	2cwq	2h1c	2pq7	2zdp	3eko
1byq	1h9k	1mm8	1r8e	1xa3	2cyy	2h98	2pt7	2zf8	3elk
1c02	1h9m	1mnz	1r9x	1xc3	2d1h	2h99	2puc	2zie	3eus
1c52	1h9s	1moq	1r9y	1xd7	2d1v	2h9b	2pud	2zif	3ex8
1c5k	1hfe	1muh	1r9z	1xi2	2d5m	2haw	2px7	2zig	3eyw
1c75	1hm9	1mur	1ra0	1xja	2d5n	2hek	2q0o	2zki	3ezu
1cb7	1hw1	1mus	1ra5	1xk6	2d5w	2heu	2q0t	2zkz	3f1c
1ccw	1hxd	1muw	1rak	1xk7	2dbb	2hkl	2q1z	2zod	3f1n
1cp2	1i0r	1mv8	1req	1xkw	2dek	2hmt	2q4f	2zov	3f1o
1crx	1i1g	1mw8	1rhc	1xkz	2df8	2hmu	2q8p	2zxj	3f1p
1crz	1i52	1mw9	1rio	1xma	2dg6	2hmv	2qb6	3b4y	3f2b
1ctj	1i58	1n2z	1rk6	1xo0	2di3	2hnh	2qb7	3b6i	3f44
1d4a	1i5n	1n9l	1rp3	1xoc	2dql	2hoe	2qb8	3b8x	3f52
1d5y	1i74	1n9n	1rrm	1xw3	2dvz	2hof	2qcz	3b9o	3f6c
1dad	1i8o	1nfp	1rtt	1y0h	2dxw	2hph	2qdf	3bcv	3f6o
1dae	1i9c	1nki	1rzu	1y1z	2dxx	2hq0	2qdl	3be6	3f6v
1dag	1icr	1nly	1rzv	1y20	2e15	2hqs	2qeu	3bem	3f8b
1dah	1id0	1nnf	1s5m	1y7m	2e1n	2hs5	2qgq	3bg2	3f8c
1dai	1id1	1nox	1s5n	1y7y	2e4n	2hsg	2qgz	3bhq	3f8f
1dak	1ihc	1nqe	1s8n	1y80	2e5f	2hsi	2qi9	3bkh	3fd3
1dd9	1ihr	1nw5	1sfx	1y82	2e7w	2hwv	2qj7	3bkv	3fgv
1dde	1ihu	1nw6	1sg0	1y9u	2e7x	2hxv	2qm1	3bm7	3fis
1di6	1ii0	1nw7	1si0	1yc9	2e7z	2i0m	2qmo	3bpk	3fms
1di7	1ii9	1nw8	1sig	1ydx	2eb7	2i5r	2qpq	3bpq	3fwy
1dlj	1ini	1nwz	1sly	1ye5	2ecu	2ia2	2qsx	3bpv	3fwz
1dts	1inj	1ny5	1sqe	1yf2	2efn	2ia4	2qwx	3bqx	3fxa
1dur	1ir6	1ny6	1sqs	1yg2	2eh3	2ibd	2qx4	3bre	3fzv
1e2x	1iuj	1o1h	1sum	1yio	2ehl	2ict	2qx6	3bs3	3g13
1e3u	1ixc	1o2d	1suu	1yiq	2ehz	2ift	2qx8	3bvp	3g5o
1e4d	1ixg	1o61	1t3t	1ylf	2ek5	2ikk	2r01	3bwg	3g7r
1e4f	1ixh	1o69	1t5b	1yoy	2esh	2ipl	2r0x	3c1q	3gdi
1e4g	1iz1	1o7l	1t72	1ysp	2esn	2ipm	2r1j	3c29	3gfa
1e8c	1j5y	1oad	1ta9	1ysq	2esr	2ipn	2r25	3c3w	3gfv
1ecl	1j6u	1oap	1td5	1yvi	2ewn	2is1	2r4t	3c48	3gfx
1efa	1jbg	1odd	1tf1	1z05	2ewv	2is2	2r6g	3c57	3gfy
1efd	1jbw	1odv	1tqg	1z19	2eyu	2is4	2r6o	3c7j	3gfz
1eg2	1je8	1oj7	1tqq	1z7u	2f00	2is6	2r6v	3c85	3gg0
1ek9	1jet	1olt	1tv8	1zat	2f2e	2is8	2ra5	3c8f	3gg1
1esz	1jeu	1opc	1tvl	1zi0	2f5x	2iu5	2rb9	3c8n	3gg2
1etk	1jev	1opx	1tzb	1zlj	2f6g	2iuy	2rc7	3c9u	3ghj

1eto	1jft	1or7	1tzc	1zvt	2f6p	2iv7	2rc8	3can	3gp4
1etv	1jh9	1ot6	1u07	1zvu	2f7a	2iw1	2rca	3ccg	3gpv
1etw	1jiw	1ot9	1u2w	1zzc	2f7b	2iw4	2rde	3cij	3gr3
1etx	1jlj	1ota	1u8b	2a0b	2f8l	2iwx	2rii	3cix	3guv
1ety	1jnu	1otb	1u8t	2a3n	2f9f	2jba	2ril	3ckj	3h4o
1ezw	1jpu	1oxk	1uaa	2a5h	2fa1	2jcg	2rsl	3ckn	3h5t
1f07	1jq5	1p2f	1uc8	2a5l	2fa5	2jfg	2uag	3ckv	3h87
1f1u	1jyk	1p31	1uc9	2a61	2fb2	2nip	2v25	3clo	3hfi
1f44	1k20	1p3d	1us4	2aa4	2fbh	2npn	2v2k	3cnr	3hh0
1f48	1k2v	1p7d	1us5	2aac	2fcj	2nq2	2v9y	3cnv	3hhh
1f5v	1k38	1p9r	1usc	2ad6	2fdn	2nq9	2vha	3cp5	3hl0
1f9i	1k4f	1p9w	1usf	2ad7	2fe1	2nqh	2vjq	3ctp	3hmz
1fca	1k54	1pb0	1uux	2ad8	2fez	2nt3	2vk2	3cuo	3hn7
1fdn	1k56	1pb7	1uuy	2aef	2ff4	2nt4	2vke	3cwr	3hoi
1fep	1kap	1pb8	1uyl	2aej	2ffu	2o08	2vkr	3cx4	3htv
1fia	1kb0	1pjr	1v4y	2afh	2fhp	2o0y	2vlg	3cyi	3hvw
1fip	1kbu	1pnz	1v51	2am1	2fn9	2o3j	2vma	3cyp	3pyp
1fp6	1kgs	1po0	1v8p	2anu	2fnu	2o4d	2vmb	3cyq	3uag
1fr3	1kmo	1pt7	1v96	2ap1	2fpo	2o7i	2vpz	3d5k	4aah
1fse	1kmp	1pvp	1vct	2ar0	2fsw	2o7p	2vsh	3d6z	4crx
1fxo	1kq3	1q05	1ve2	2ara	2fvy	2o8x	2w27	3d7i	4req
1g1l	1ku3	1q06	1vf7	2arc	2fw0	2o99	2w8b	3dbo	4uag
1g1m	1ku7	1q07	1vgt	2azn	2g2c	2o9a	2w8i	3df7	5req
1g20	1kv9	1q08	1vgw	2b02	2g6v	2obc	2yve	3df8	6req
1g28	1kw3	1q09	1vhd	2b0p	2g7u	2ofy	2yx0	3dma	7req
1g5p	1kw6	1q0a	1vhv	2b13	2gai	2ogi	2yxb	3dr4	8abp
1g60	1l3l	1q35	1vim	2b3z	2gaj	2ojh	2yxo	3drf	
1g6o	1lj9	1q7e	1vj7	2b44	2gci	2okc	2yxz	3drj	
1g72	1lq9	1qg8	1vke	2bas	2gd0	2olb	2yye	3dsg	
1g8k	1lqk	1qgq	1vlj	2bfw	2gd2	2ooc	2yz5	3du1	

Table S2. List of Pfam domain families analyzed in this study.

Pfam Domain Names				
ABM	Fe-ADH	HlyD	PAS	SBP_bac_1
AIRS	FecCD	Hpt	PASTA	SBP_bac_3
AIRS_C	Fer4	HxlR	PAS_3	SBP_bac_5
AP_endonuc_2	Fer4_NifH	IclR	PD40	SIS
ATP-grasp_3	Flavin_Reduct	IspD	PHP	SLBB
Amidohydro_3	Flavodoxin_2	IstB	PIN	SLT
AraC_binding	FtsA	LacI	PQQ	Sigma54_activat
ArsA_ATPase	GGDEF	LysR_substrate	PadR	Sigma70_r2
AsnC_trans_reg	GSPII_E	MCPsignal	ParBc	Sigma70_r4
B12-binding	GSPII_F	MarR	Pentapeptide	Sigma70_r4_2
BPD_transp_1	GerE	MerR-DNA-bind	Peptidase_M23	Surf_Ag_VNR
Bac_luciferase	Glycos_transf_1	MerR	Peripla_BP_1	TOBE
Bug	Glycos_transf_2	Methylase_S	Peripla_BP_2	TOBE_2
CMD	Glyoxalase	MoCF_biosynth	Phage_integr_N	TP_methylase
CbiA	GntR	Molybdopterin	Phage_integrase	TetR_N
CheW	HATPase_c	Molydop_binding	PhoU	TonB
CoA_transf_3	HD	Mur_ligase	PilZ	TonB_dep_Rec
Cons_hypoth95	HTH_1	Mur_ligase_C	Plasmid_stabil	Toprim
Cytochrom_C	HTH_11	Mur_ligase_M	Plug	Trans_reg_C
DHH	HTH_3	N6_Mtase	ROK	Transpeptidase
DHHA1	HTH_5	N6_N4_Mtase	Radical_SAM	Transposase_11
DNA_gyraseA_C	HTH_8	NMT1	Resolvase	TrkA_N
DegT_DnrJ_EryC1	HTH_AraC	NTP_transferase	Response_reg	TrmB
EAL	HTH_IclR	Nitroreductase	RibD_C	UDPG_MGDP_dh_N
FCD	HemolysinCabind	OEP	RimK	UTRA
FMN_red	HisKA	OmpA	Rrf2	UvrD-helicase
				YkuD

Table S3. Pfam domain families and their respective PDB structure with oligomerization TP contacts.

Pfam Domain	PDB structure
AsnC_trans_reg	2z4p
Bac_luciferase	3b4y
CMD	1vke
EAL	2r6o
Flavodoxin_2	1t5b
FMN_red	2a5l, 2q62
Glyoxalase	2p7o
GSPII_E	2gza
HlyD	2f1m,1t5e
Hpt	1i5n
HTH_IclR	2g7u
HxlR	2f2e
IspD	3f1c
MCPsignal	2ch7
MerR-DNA-bind	3gp4
Mur_ligase	2am1
Resolvase	2gm5
Sigma54_activat	1ny6
TOBE	1h9s
TOBE_2	2awn
TP_methylase	1vhv

Table S4. Top-30 prediction of mfDCA for the Serine protease data of (41). The first two columns specify the residue pair, the third column provides the DI value, and the last one the native distance in rat trypsin (PDB ID 3tgi). Residues belonging to the sectors defined in (41) are indicated, using the color scheme of (41).

Res. 1	Res. 2	DI	Dist/Å
136	201	0.52	2.0
32	40	0.47	2.8
191	220	0.37	2.2
189	226	0.34	3.3
57	195	0.34	2.7
42	58	0.28	2.0
44	52	0.25	4.3
30	139	0.25	2.7
72	77	0.24	3.0
72	78	0.23	8.0
59	104	0.23	3.9
51	105	0.22	3.8
190	213	0.20	3.7
34	40	0.19	3.4
116	127	0.18	23.7
26	157	0.18	4.9
45	209	0.18	3.8
117	127	0.17	23.9
46	112	0.16	4.0
71	78	0.15	8.5
71	79	0.15	6.9
117	122	0.15	13.3
161	184	0.15	3.1
138	213	0.14	4.2
116	122	0.14	13.1
53	209	0.14	3.5
189	228	0.13	3.9
100	179	0.13	2.3
102	195	0.13	6.1
27	157	0.13	3.8