File S1

Supporting Methods

Sequence Reweighting and Pseudocounts

In order to control for sequence bias in our MSA, sets of sequences that exceed a certain identity threshold are down-weighted as a group (Weigt *et al.* 2009; Marks *et al.* 2011; Morcos *et al.* 2011; Hopf *et al.* 2012). For every sequence *m* in an MSA, the number of "identical" sequences k_m is defined as

$$k_m \equiv \sum_{n=1}^{M} \theta\left(\sum_{i=1}^{L} \delta(A_i^m, B_i^n) - xL\right) \quad \textbf{[S1]}$$

where ϑ is a step function equal to one if its argument is greater than or equal to zero and zero if the summation is negative, δ is the Kronecker symbol used for counting, which is equal to one if A_i^m equals B_i^n and to zero otherwise, and x is the identity threshold, defined here as 0.7. When counting pair and single amino acid frequencies, the contribution of sequence *m* is down-weighted by $1/k_m$. The effective number of sequences in an alignment is therefore not M but M_{eff}, where

$$M_{eff} = \sum_{m=1}^{M} \frac{1}{k_m}. \quad \textbf{[S2]}$$

Pair and single amino acid frequencies are then calculated according to the relationships

$$f_i(A) \equiv \frac{1}{\lambda + M_{eff}}\left(\frac{\lambda}{q} + \sum_{m=1}^{M} \frac{1}{k_m} \delta(A_i^m, A)\right) \quad \textbf{[S3A]}$$

$$f_{ij}(A,B) \equiv \frac{1}{\lambda + M_{eff}}\left(\frac{\lambda}{q^2} + \sum_{m=1}^{M} \frac{1}{k_m} \delta(A_i^m, A)\delta(B_j^m, B)\right) \quad \textbf{[S3B]}$$

where λ is a pseudocount term used to ameliorate statistical noise due to underrepresented amino acids and pairs. Here we set λ equal to M_{eff}. Note that the empirical correlation matrix is not invertible before pseudocounts are incorporated.

DCA

According to DCA, the coupling between columns *i* and *j* in an MSA is given by the direct information, DI_{ij}, score according to the relationship

$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}(A,B) \ln\left(\frac{P_{ij}(A,B)}{f_i(A) f_j(B)}\right) \quad \textbf{[S4]}$$

where $P_{ij}(A,B)$ represents the inferred probability of finding amino acid pair (A,B) at positions i and j in the absence of interactions with other residues, $f_i(A)$ and $f_j(B)$ represent the single amino acid frequencies of A and B at positions i and j, and the summation is evaluated over all 441 pairs (A,B) possible for a q = 21 state system, where the states represent the twenty amino acids and a gap. $P_{ij}(A,B)$ is itself a function of the inferred coupling energy $e_{ij}(A,B)$ and the inferred single residue energies $\tilde{h}_i(A)$ and $\tilde{h}_j(B)$ of amino acids A and B at positions i and j according to

$$P_{ij}(A,B) = \frac{1}{Z_{ij}} \left\{ e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B) \right\} \quad \textbf{[S5]}$$

where Z_{ij} is the partition function. The coupling energies $e_{ij}(A,B)$ are determined as described below by inverting an empirical correlation matrix, **C**.

The empirical correlation matrix **C** is determined from the MSA according to the relationships

$$C_{ij}(A,B)_{i \neq j} = f_{ij}(A,B) - f_i(A) f_j(B) \quad \textbf{[S6]}$$

$$C_{ij}(A,B)_{i=j, A=B} = f_i(A)\left(1 - f_i(A)\right) \quad \textbf{[S7]}$$

where $f_i(A)$ is the frequency of amino acid A in MSA column i, $f_j(B)$ is the frequency of amino acid B in MSA column j, and $f_{ij}(A,B)$ is the frequency of amino acid pair (A,B) in columns i and j. Calculation of correlations $C_{ij}(A,B)$ where $i = j$ but $A \neq B$ is carried out according to Equation S6. Note that pair frequencies $f_{ij}(A,B)$ are set to zero for these entries (despite having a finite value based on pseudocounts, as described below to reflect the fact that no protein sequence contains two different amino acids at a single site. The empirical correlation matrix has the dimensions 20L by 20L despite the fact that we employ a q = 21 state model. This is because one amino acid, in our case the gap, is left out of the analysis in order to serve as a reference energy.

The global nature of the DCA algorithm derives from inversion of the empirical correlation matrix (or the composite matrix **C*** described below), which results in the coupling energy matrix, **e**:

$$\mathbf{e} = -\mathbf{C}^{-1}. \quad \textbf{[S8]}$$

The fields $\tilde{h}_i(A)$ and $\tilde{h}_j(B)$ from Equation S5 are calculated numerically along with the partition function Z_{ij} so that the pair probabilities recapitulate the single amino acid frequencies, $f_i(A)$ and $f_j(B)$, observed in the MSA:

$$\sum_{B=1}^{q} P_{ij}(A,B) \cong f_i(A) \quad \textbf{[S9A]}$$

$$\sum_{A=1}^{q} P_{ij}(A,B) \cong f_j(B). \quad \textbf{[S9B]}$$

Once field and coupling energies have been determined, direct information DI_{ij} scores can be evaluated using Equations S4 and S5. The result is a list of DI_{ij} scores representing the direct information between every pair of positions.

Supporting Literature Cited

Hopf T. A., Colwell L. J., Sheridan R., Rost B., Sander C., Marks D. S., 2012 Three-dimensional structures of membrane proteins from genomic sequencing. Cell **149**: 1607–21.

Marks D. S., Colwell L. J., Sheridan R., Hopf T. A., Pagnani A., Zecchina R., Sander C., 2011 Protein 3D structure computed from evolutionary sequence variation. PloS One **6**: e28766.

Morcos F., Pagnani A., Lunt B., Bertolino A., Marks D. S., Sander C., Zecchina R., Onuchic J. N., Hwa T., Weigt M., 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci **108**: E1293–301.

Weigt M., White R. A., Szurmant H., Hoch J. A., Hwa T., 2009 Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci **106**: 67–72.

Table S1 Strains and plasmids

Strain/plasmid	Genotype and relevant features	Reference
***E. coli* K-12 strains**		
MC4100	F- *araD139 (argF-lac)U169 rpsL150 relA1 flb5301 deoC1 ptsF25 thi*	Boyd et al 2000
JCM158	MC4100 $ara^{r/-}$	Malinverni et al 2006
JCM320	JCM158 Δ*bamA* Δ(λ*att-lom*)::*bla* P_{BAD} *bamA araC*	Wu et al 2005
DPR437	JCM320 pDPR1	Ricci et al 2012
DPR660	JCM320 pBamA^{R661G}	This study
DPR1345	JCM320 pBamA^{D740G}	This study
DPR1346	JCM320 pBamA$^{D740G+R661G}$	This study
DPR1374	JCM320 pBamA$^{D740G+F395V}$	This study
DPR1309	JCM320 pBamA$^{D740G+T423I}$	This study
DPR1310	JCM320 pBamA$^{D740G+E607A}$	This study
DPR1311	JCM320 pBamA$^{D740G+G631V}$	This study
DPR1500	JCM320 pBamAD$^{740G+G631W}$	This study
DPR1313	JCM320 pBamA$^{D740G+F717L}$	This study
DPR1317	JCM320 pBamA$^{R661G+F395V}$	This study
DPR1318	JCM320 pBamA$^{R661G+T423I}$	This study
DPR1319	JCM320 pBamA$^{R661G+E607A}$	This study
DPR1320	JCM320 pBamA$^{R661G+G631V}$	This study
DPR1501	JCM320 pBamA$^{R661G+G631W}$	This study
DPR1321	JCM320 pBamA$^{R661G+F717L}$	This study
Plasmids		
pZS21	Expression vector; λ P_L-driven expression, Kan^r	Lutz & Bujard, 1997
pBamA (pDPR1)	pZS21::*bamA*^{WT}	Kim et al 2007
pBamA^{R661G}	pZS21::*bamAR661G*	This study

pBamA^{D740G}	pZS21::*bamAD740G*	This study
pBamA$^{D740G+R661G}$	pZS21::*bamAD740G+R661G*	This study
pBamA$^{D740G+F395V}$	pZS21::*bamAD740G+F395V*	This study
pBamA$^{D740G+T423I}$	pZS21::*bamAD740G+T423I*	This study
pBamA$^{D740G+E607A}$	pZS21::*bamAD740G+E607A*	This study
pBamA$^{D740G+G631W}$	pZS21::*bamAD740G+G631W*	This study
pBamA$^{D740G+G631V}$	pZS21::*bamAD740G+G631V*	This study
pBamA$^{D740G+F717L}$	pZS21::*bamAD740G+F717L*	This study
pBamA$^{R661G+F395V}$	pZS21::*bamAR661G+F395V*	This study
pBamA$^{R661G+T423I}$	pZS21::*bamAR661G+T423I*	This study
pBamA$^{R661G+E607A}$	pZS21::*bamAR661G+E607A*	This study
pBamA$^{R661G+G631W}$	pZS21::*bamAR661G+G631W*	This study
pBamA$^{R661G+G631V}$	pZS21::*bamAR661G+G631V*	This study
pBamA$^{R661G+F717L}$	pZS21::*bamAR661G+F717L*	This study

Table S2 Primers

BamA mutation	Primer pairs
F395V	5' GAATCGTCTGGGCTTCGTTGAAACTGTCGATAC 3'
	5' GTATCGACAGTTTCAACGAAGCCCAGACGATTC 3’
T423I	5' GTAAAAGAGCGCAACATCGGTAGCTTCAACTTTG 3'
	5' CAAAGTTGAAGCTACCGATGTTGCGCTCTTTTAC 3'
E607A	5' CTGGATCGGATAACGCATACTACAAAGTGAC 3'
	5' GTCACTTTGTAGTATGCGTTATCCGATCCAG 3'
G631V	5' CAAATGGGTTGTTCTGGTGCGTACCCGCTGGG 3'
	5' CCCAGCGGGTACGCACCAGAACAACCCATTTG 3'
G631W	5' CAAATGGGTTGTTCTGTGGCGTACCCGCTGGG 3'
	5' CCCAGCGGGTACGCCACAGAACAACCCATTTG 3'
R661G	5' TTCCAGCACCGTGGGCGGCTTCCAGTCCAATA 3'
	5' TATTGGACTGGAAGCCGCCCACGGTGCTGGAA 3'
F718L	5' CAGCCTCGAGTTAATCACCCCGACG 3'
	5' CGTCGGGGTGATTAACTCGAGGCTG 3'
D740G	5' CTTCCTTCTTCTGGGGTATGGGTACCGTTTG 3'
	5' CCAAACGGTACCCATACCCCAGAAGAAGGAAGTAC 3'

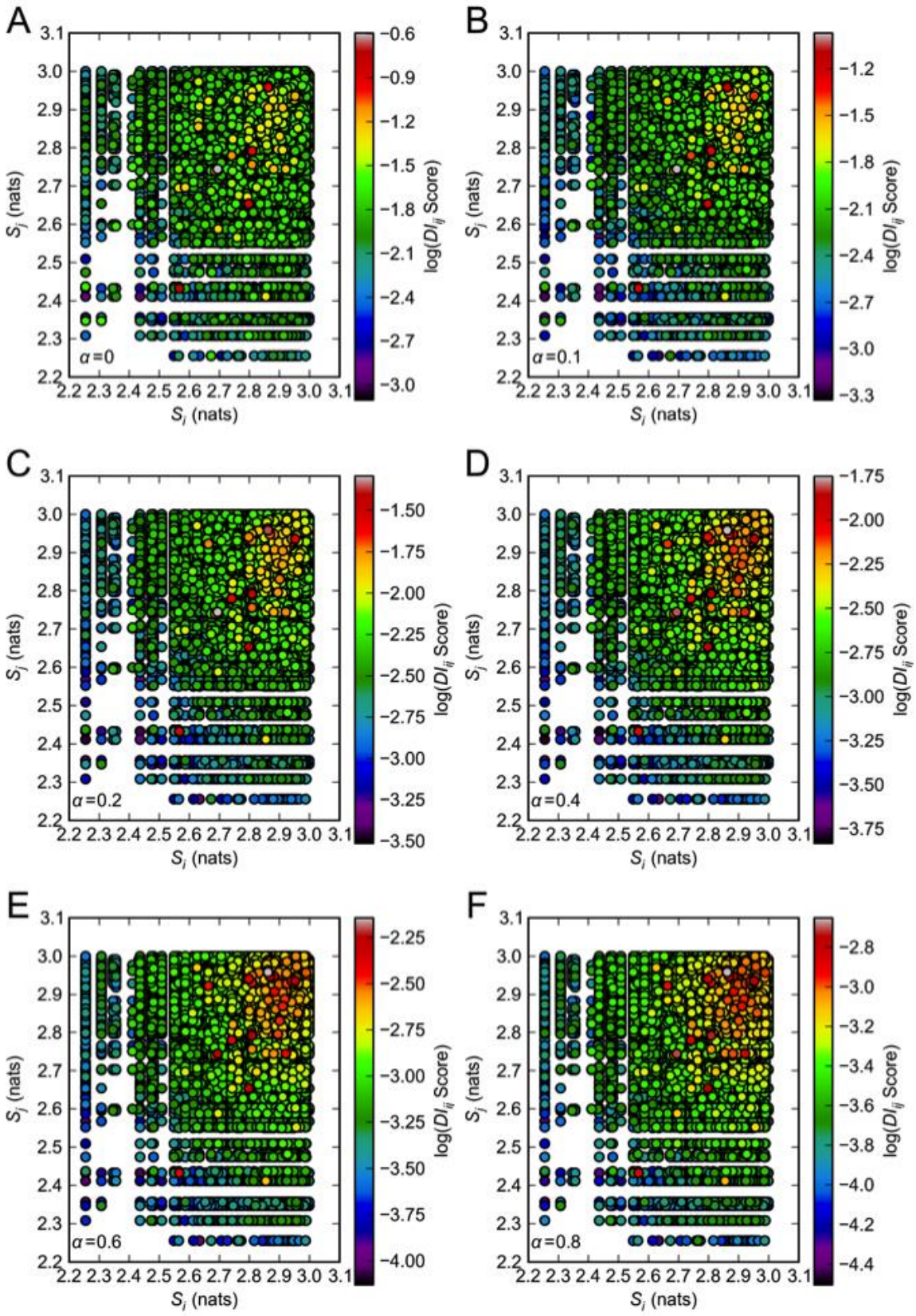

Figure S1 Effect of sequence informational entropy S_i, S_j on pair DI_{ij} score. Log(DI_{ij} Score) is plotted against sequence informational entropies S_i and S_j for all FhaC pairs shown in Figure 1C.

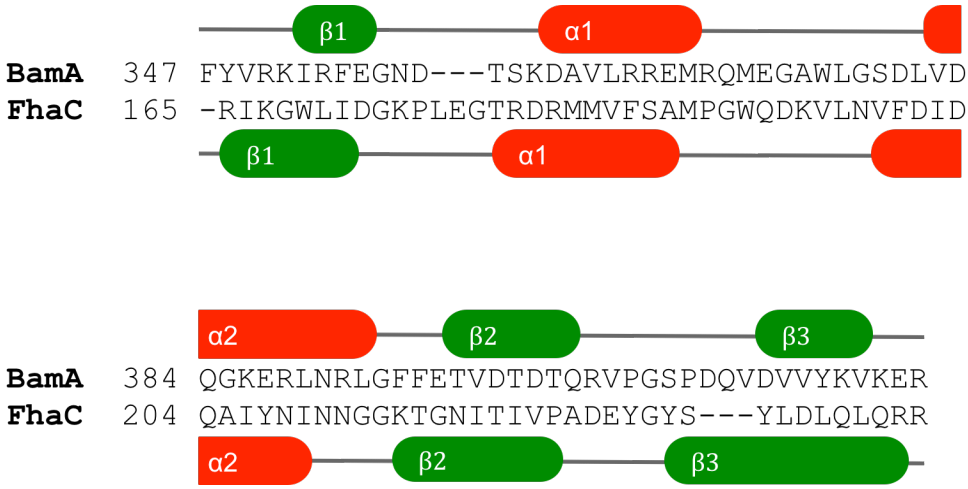

Figure S2 Alignment of BamA POTRA 5 and FhaC POTRA 2 domains. FhaC sequence comprises residues 165 to 238 of *Bordetella pertussis* FhaC. BamA sequence comprises residues 347 to 421 of *Escherichia coli* BamA. Sequences were aligned using COBALT. Secondary structure was determined for FhaC and BamA from crystal structures 2QDZ and 3OG5, respectively.

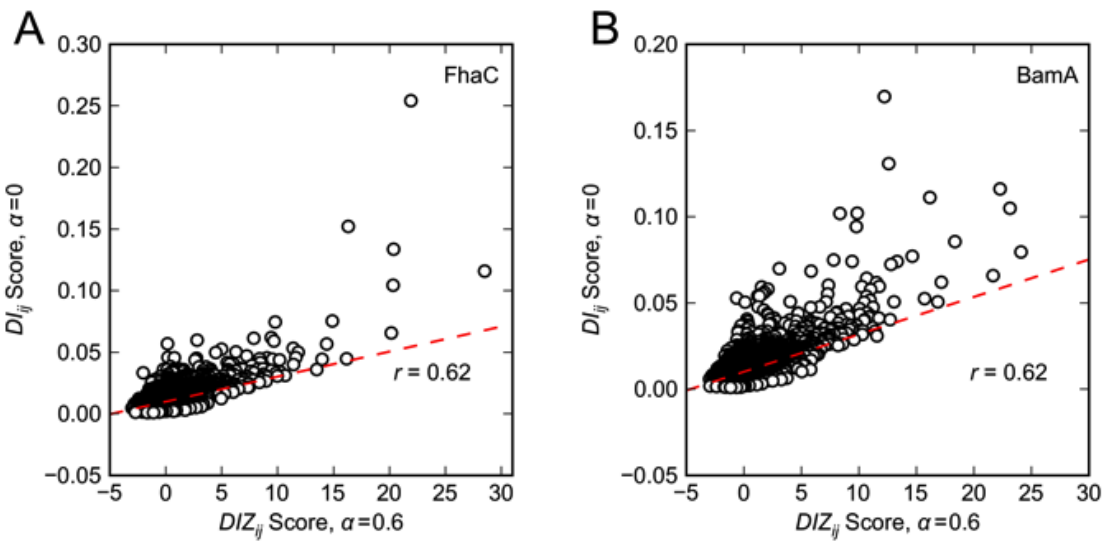

Figure S3 Correlation of $\mathrm{DCA}_{DI_{ij}}^{\alpha=0}$ and $\mathrm{DCA}_{DIZ_{ij}}^{\alpha=0.6}$ scores. DCA was performed as in Figures 2E,F. Least squares regression line (red) is shown along with correlation coefficient *r*.

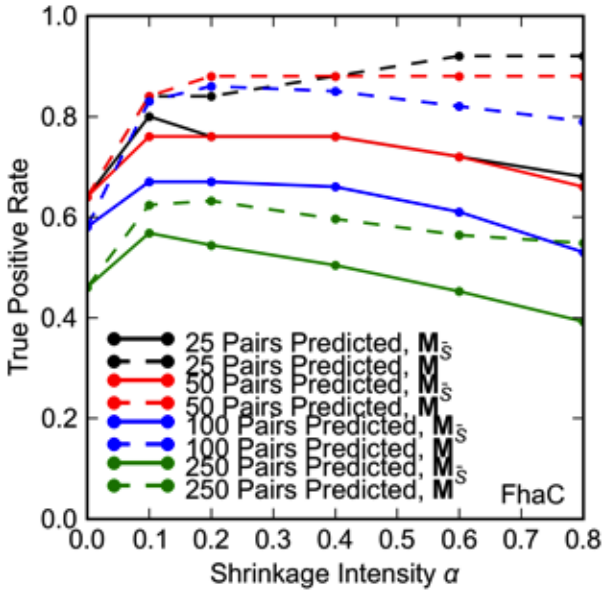

Figure S4 Effect of shrinkage with model matrix $\mathbf{M}_{\tilde{S}}$ on DCA true positive rates. DCA was applied to FhaC as in Figures 1A,B with the same true positive definition. True positive rates are shown for various values of shrinkage intensity α between 0 and 1.