Supplementary Material

Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method

Lukas Burger
Erik van Nimwegen
Biozentrum, the University of Basel, and Swiss Institute of Bioinformatics
Klingelbergstrasse 50/70, 4056-CH, Basel, Switzerland,
email: lukas.burger@unibas.ch and erik.vannimwegen@unibas.ch

November 23, 2007

1 Classifying receiver domains

Similar to previous work, i.e. (Grebe and Stock, 1999; Koretke *et al*, 2000), we found that cognate response regulators that interact with different types of kinases show distinct amino acid compositions in their receiver domains and that these differences can be used to predict, for each receiver domain, what kind of kinase it will interact with.

We divided the multiple alignment of all cognate receiver domains into 8 sub-alignments corresponding to sets of regulators that interact with kinases of each particular kinase class. For each of the 8 alignments we then constructed a position specific weight matrix

$$w^c_{i\alpha} = \frac{n^c_{i\alpha} + \lambda}{\sum_\alpha (n^c_{i\alpha} + \lambda)}. \tag{1}$$

Here $n^c_{i\alpha}$ is the total number receivers of class c that have an amino acid α in column i of the alignment (gaps are treated as a 21st amino acid) and λ is the pseudo-count resulting from the Dirichlet prior (we used the Jeffreys' prior $\lambda = 1/2$). $w^c_{i\alpha}$ is thus the estimated probability of seeing amino acid α in position i of a receiver of class c.

Given a receiver with sequence S we can now determine the posterior probability $P(c|S)$ that it belongs to class c. We have

$$P(c|S) = \frac{P(S|c)P(c)}{\sum_{c'} P(S|c')P(c')} \quad \text{with} \quad P(S|c) = \prod_i w^c_{S_i i}, \tag{2}$$

where S_i is the amino acid in the ith position of receiver sequence S and the product runs over all positions in the receiver. We assumed a uniform prior $P(c) = 1/8$.

We tested to what extent this simple model is capable of correctly classifying receiver sequences. For each cognate receiver we calculated the posterior probability $P(c|S)$ of the class c given the receiver sequence S, using the WMs $w^c_{i\alpha}$ constructed from all receiver sequences. We then assigned the receiver to the class c that maximizes $P(c|S)$. The results in Fig. 1 show that for the three most abundant types of kinases (HisKA, H3, and HisKin), and for the Hwe kinases as well, the classifier predicts almost perfectly which kinase type the respective receivers interact with. For the other classes the classification is still correct in the majority of the cases.

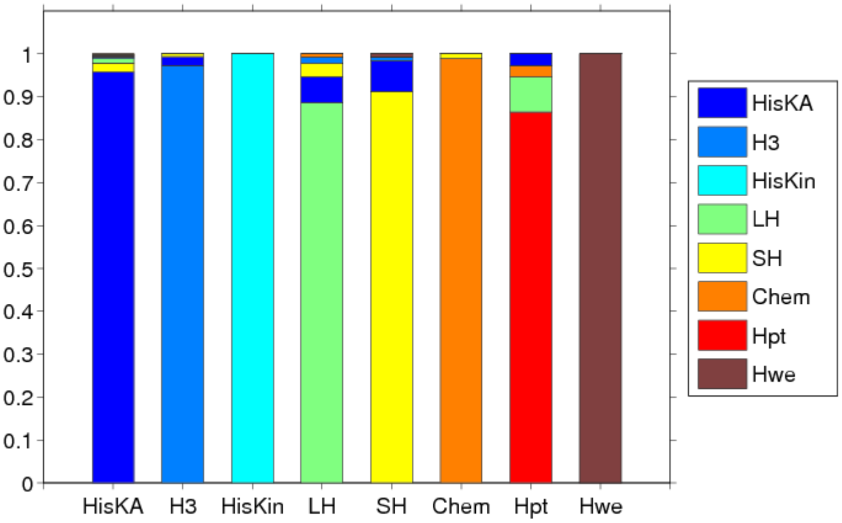


Figure 1: Predicted classification of receivers. Each bar represents the set of all receivers that are member of a cognate pair with kinases of a particular type (indicated below the bar). In each bar the colors indicate what fraction of the cognate receivers of this type is classified with each type of kinase. The legend on the right shows the correspondence between color and kinase type. SH and LH stand for short and long hybrid, respectively, and Chem stands for chemotaxis.

The types of mis-classifications match what is to be expected based on the domain architectures. Both long and short hybrids contain an HisKA domain and their receivers are sometimes mistaken for a receiver that interacts with a single HisKA domain kinase. Both long hybrid kinases and Hpt kinases contain an Hpt domain and the most common misclassification is between receivers that interact with a kinase with a single Hpt domain and receivers that interact with long hybrids. Because of this, and because the number of cognate pairs of the Hpt class is very small, we have treated the Hpt and long hybrid classes as one class in our analysis (leaving 7 classes in total). Although they also contain an

Hpt domain, cognate receivers of chemotaxis kinases are very rarely mistaken with receivers of Hpt and long hybrid kinases, probably due to the fact that cognate regulators of chemotaxis kinases are mainly CheB and CheY regulators which have very specific functions in chemotaxis and correspondingly specific amino acid composition. Overall, the WM model predicts the correct type of kinase for 96% of the cognate receiver domains.

2 Bayesian network model details

We first derive why consistency requires that the pseudo-count λ of the Dirichlet prior for the marginal probabilities w_α is related to the pseudo-count λ' of the joint probabilities $w_{\alpha\beta}$ through

$$\lambda = 21\lambda' \tag{3}$$

In the Methods section of the main paper we calculated expressions for $P(D_i)$ and $P(D_{ij})$ in terms of λ, λ', the joint counts $n^{ij}_{\alpha\beta}$ and the marginal counts n^i_α and n^j_β. The conditional probability is then given by $P(D_i|D_j) = P(D_{ij})/P(D_j)$. However, we could have also calculated the conditional probability by introducing the *conditional* probabilities $w_{\alpha|\beta}$ which give the probability that α occurs at position i given that β occurred at position j. In terms of this parametrization we obtain

$$P(D_i|w, D_j) = \prod_{\alpha,\beta} (w_{\alpha|\beta})^{n^{ij}_{\alpha\beta}}. \tag{4}$$

Using again a Dirichlet prior with pseudo-count λ', the integral over possible conditional probabilities $w_{\alpha|\beta}$ then gives

$$P(D_i|D_j) = \prod_\beta \left[\int P(D_i|w, D_j)P(w)dw_{\alpha|\beta}\right] = \prod_\beta \left[\frac{\Gamma(21\lambda')}{\Gamma(n^j_\beta + 21\lambda')} \prod_\alpha \frac{\Gamma(n^{ij}_{\alpha\beta} + \lambda')}{\Gamma(\lambda')}\right]. \tag{5}$$

It is easy to see that this will only match the conditional probability we calculated through $P(D_i|D_j) = \frac{P(D_{ij})}{P(D_j)}$ if $\lambda = 21\lambda'$. In addition, in the Methods section of the main paper we also noted that equation (7) is independent of the choice of the root. However, this is also only true when $\lambda = 21\lambda'$.

2.1 Probabilities of unassigned kinases and receivers

The calculation of the joint probability $P(J, K, R)$, with J the alignments of assigned pairs, K the alignment of unassigned kinases, and R the alignment of unassigned receiver domains, is identical for each particular class of kinases. We thus focus on a single class. As described in the main paper we make the assumption that K depends only on the kinase sequences in J and R only on the receiver sequences in J. That is, for the probability of the kinases that are not assigned, only the amino acids in the *kinases* of the assigned pairs matter, not the amino acids of the *receivers* in the assigned pairs (and vice versa for the receivers). Formally, we thus assume that we can factorize $P(K, R, J)$ as follows

$$P(K, R, J) = P(K|J^k)P(R|J^r)P(J) \tag{6}$$

with J^k the sequences of the assigned kinases and J^r the sequences of the assigned receivers.

Since the calculation of $P(K|J^k)$ and $P(R|J^r)$ is identical we focus on the calculation of the kinase probabilities $P(K|J^k)$. We first calculate this conditional probability for a specific dependence tree T, i.e. we calculate $P(K|J^k, T)$. Note that, in contrast to the dependence tree for the joint alignment J, this tree includes only positions within the kinase. We now use the general identity

$$P(K|J^k, T) = \frac{P(K, J^k|T)}{P(J^k|T)} \tag{7}$$

and use equations (2), (4), (5), and (7) from the main paper to calculate the factors in numerator and denominator. In particular, let K^{ij} denote the set of counts in the ith and jth columns of the unassigned kinases K, with $K^{ij}_{\alpha\beta}$ the number of times the combination $(\alpha\beta)$ occurs at positions (ij). Similarly let k^{ij} denote the counts in columns i and j of the kinases in J^k with $k^{ij}_{\alpha\beta}$ the number of times combination $(\alpha\beta)$ occurs in columns (ij). We also have the marginal counts K^i_α and k^i_α in columns K^i and k^i. Using equation (7) from the main paper

$$P(D|T) = \left[\prod_i P(D^i)\right] \left[\prod_{i \neq r} R_{i\pi(i)}\right], \tag{8}$$

we have

$$P(K|J^k, T) = \left[\prod_i \frac{P(k^i + K^i)}{P(k^i)}\right] \prod_{i \neq r} \frac{R_{i\pi(i)}(k^{ij} + K^{ij})}{R_{i\pi(i)}(k^{ij})}, \tag{9}$$

where the function $P(n^i)$ of the set of marginal counts n^i is given by expression (2) in the main paper

$$P(n^i) = \frac{\Gamma(21\lambda)}{\Gamma(n + 21\lambda)} \prod_\alpha \frac{\Gamma(n^i_\alpha + \lambda)}{\Gamma(\lambda)}, \tag{10}$$

and the function $R_{ij}(n^{ij})$ of the set of counts n^{ij} in a pair of columns (ij) is given by combining equation (4) from the main paper:

$$P(n^{ij}) = \frac{\Gamma(21^2\lambda')}{\Gamma(n + 21^2\lambda')} \prod_{\alpha\beta} \frac{\Gamma(n^{ij}_{\alpha\beta} + \lambda')}{\Gamma(\lambda')}, \tag{11}$$

with equation (5) from the main paper: $R_{ij}(n^{ij}) = P(n^{ij})/[P(n^i)P(n^j)]$. In summary, the conditional probability $P(K|J^k, T)$ given a dependence tree T

can be determined by using the exact same expressions as used for calculating $P(J|T)$ in the main paper, only now we calculate the ratio of the probabilities of the alignment containing both counts K and J^k and the alignment containing only counts J^k.

Finally, if we define the ratio of R_{ij} values:

$$\tilde{R}_{ij} = \frac{R_{ij}(k^{ij} + K^{ij})}{R_{ij}(k^{ij})}, \tag{12}$$

we could again calculate the sum over spanning trees by defining the Laplacian matrix

$$\tilde{M}_{ij} = \delta_{ij} \left(\sum_k \tilde{R}_{ik} \right) - \tilde{R}_{ij}, \tag{13}$$

from which one row and column have been removed, and using

$$P(K|J^k) = \left[\prod_i \frac{P(k^i + K^i)}{P(k^i)} \right] \frac{1}{|T|} \det(\tilde{M}). \tag{14}$$

However, as detailed below, calculating this determinant accurately is a challenging numerical problem which has currently not be satisfactorily solved, see e.g. (Cerquides and de Màntaras, 2003), and we instead use the approximation of only using the dependence tree T^* with maximal probability, i.e.

$$P(K|J^k) \approx P(K|J^k, T^*). \tag{15}$$

We choose the dependence tree T^* that maximizes the probability of all the cognate kinases and keep this tree fixed throughout the sampling runs. In addition, to reduce numerical error due to small spurious correlations, we score positions that show no evidence of dependence with any other position according to a WM model. In particular, all positions i for which $\log(R_{ij}) < 10$ for all positions j are excluded from the dependence tree T^* and are scored with a WM model.

2.2 Approximation of the determinant

The matrix components R_{ij} and $\tilde{R}_{ij}$ correspond to the ratios of probabilities of all observed data in columns (i, j) under a general dependent model and under the assumption that i and j are independent. These in turn involve the ratios of products of gamma functions whose arguments, i.e. the number of occurrences of certain combinations of letters in certain columns, can become quite large. As a result, some of the matrix components are extremely large numbers, and others are extremely small numbers. In principle this is no numerical problem because we can easily calculate the logarithms of the matrix entries instead of the matrix entries themselves. However, when we calculate the determinant we need to calculate combinations of products, sums, and differences of these matrix entries and this is numerically very challenging.

In order to approximate the determinant we used the same approach as in (Cerquides and de Màntaras, 2003). We rescaled all matrix entries as follows

$$R_{ij} \to 10^{C(\frac{\log(\frac{R_{ij}}{R_{\min}})}{\log(\frac{R_{\max}}{R_{\min}})}-1)} \tag{16}$$

where $R_{\max}$ ($R_{\min}$) is the maximal (minimal) entry of the matrix R_{ij}. This function essentially rescales and shifts all the $\log(R_{ij})$ values such that they now map to the interval $\left[10^{-C}, 1\right]$. These scaled R values can be considered a more conservative estimate of dependence, as they diminish the relative difference in dependence between different pairs of positions (Cerquides and de Màntaras, 2003).

For our predictions of cognate two-component interactions as well as polyketide synthase interactions, we set $C = 5$, calculated $\log(R_{\max})$ as well as $\log(R_{\min})$ at the beginning of the simulation and kept it fixed during the simulation (the highest $\log(R_{ij})$ values correspond to pairs of residues (ij) that lie in the same protein and thus do not depend on the current assignment). In order to keep the absolute log-probability differences of different assignments approximately the same the resulting determinants need to be rescaled by an appropriate factor in order to counteract the reduction of log-probability differences due to the rescaling of the R-matrix entries. We chose this factor by demanding that the model reduces to the maximum-likelihood tree model in the case of one dominating tree. Let $\det(M')$ be the minor of the Laplacian with scaled R values and $\det(M)$ the minor of the Laplacian with the actual R values. We then approximate $\det(M)$ as

$$\det(M) \approx \left[\frac{\det(M')}{10^{-C(n-1)(1+\frac{\log(R_{min})}{\alpha})}}\right]^{\frac{\alpha}{C\log(10)}} \tag{17}$$

where $\alpha = \log(\frac{R_{max}}{R_{min}})$ and n is the dimension of the matrix R_{ij}. Note that this approximation is also very accurate in the case of a set of dominating dependence trees with similar likelihoods.

In an attempt to reduce numerical error due to positions that show no dependence on other positions to start with, we do not score all columns according to the general model, but filter out a subset of positions that show either very low variability, or that show no dependence on any of the other positions. In particular all positions with entropy less than 10% of the maximum possible entropy log(21), and all positions with more than 50% gaps are filtered out. These positions are scored using a simple WM model, i.e. with the probability of the letter independent of other columns. Again, this complication is to reduce numerical errors and would not be necessary if we had a better numerical procedure for calculating the determinant.

2.3 Sampling scheme for the sum-over-trees model

Without any prior knowledge about the connectivity nor about the dependence tree structure, our search space is vast and there is a great danger of getting stuck in local optima during the sampling procedure. In order to deal with this problem we used simulated annealing starting from a relatively high 'temperature'. We sample from the distribution $P(D)^{1/T}$ setting $T = 100$ at the start and decreasing T linearly with time until $T = 1$ is reached. Due to the 'heating', the probability distribution over the space of assignments is effectively flattened and it is easier to move out of local maxima in this initial phase. After $T = 1$ is reached we continue sampling at $T = 1$ and allow the system to reach equilibrium. In a final phase of sampling (still at $T = 1$) we record interaction partners to estimate the posterior distribution of interaction for any kinase/regulator pairs. The simulated annealing resulted in a significant improvement in performance compared to simulations where $T = 1$ is used throughout (data not shown).

3 Reconstruction of cognate pairs

3.1 Results for the small classes

The results of the reconstruction of cognate pairs for the smaller kinase classes are shown in figure 2. The smaller kinase classes, particularly the chemotaxis and HWE classes, have only very few kinases and regulators per genome and therefore random scoring, i.e. where every possible kinase/receiver pair inside the same genome is assigned the same probability of interaction, already produces a reasonable number of correct predictions. Additionally, the sizes of the corresponding alignments are very small and there is only little co-evolutionary information. Nonetheless, it is apparent in figure 2 that the method produces highly accurate predictions on these smaller classes as well.

3.2 Performance of the extended model on all cognate pairs

The prediction of orphan interactions requires two extensions to our model. Response regulators must be allowed to interact with kinases of any class and, due to unequal numbers of kinases and regulators, our way of assigning kinases and regulators demands that in every assignment a number of kinases and regulators do not have any interaction partner (see *Methods*).

A simple way of testing the performance of the former extension is to run our MCMC simulation with all cognate pairs of all 7 classes at the same time. The results are shown in figure 3. Due to the fact that the search space is now much bigger as every kinase can interact with any response regulator of the 7 classes, i.e. every regulatory can switch between the 7 classes of kinases, the quality of our predictions, though still quite accurate, generally decreases. It is important to note that although the chemotaxis and HWE families are very small and thus contain very little co-evolutionary information, the algorithm predicts the interaction partners of kinases of these classes with very high accuracy. This is due to the fact that regulators of the chemotaxis and HWE families form clearly distinct subfamilies (see figure 1) and thus, since they come in very small numbers per genome, a correct classification of their class membership is sufficient for determining their right interaction partners.

4 Network structure predictions

As described in the main text, for the prediction of the two-component signaling network structure, we assign a log-ratio score to any kinase/regulator pair of the HisKA class. In figure 4, we show the PPV/sensitivity curve for this log-ratio score. The used cut-off of 1 corresponds to a sensitivity of 0.56 and a PPV-value of 0.48. Note that although, at this cut-off, every second prediction corresponds to a non-cognate pair, the false positive rate is very low (0.04). Also note that for figure 4 we consider all predicted interactions between proteins belonging to different cognate pairs as false positives, which is very conservative since cross-talk between cognates is likely to exist. If we use the log-ratio score to predict HisKA orphan interactions, we get a p-value of 10^{-7} for the set of known *Caulobacter* interactions (and 10^{-3} when in addition the putative cognate pairs are excluded from our dataset (see below)).

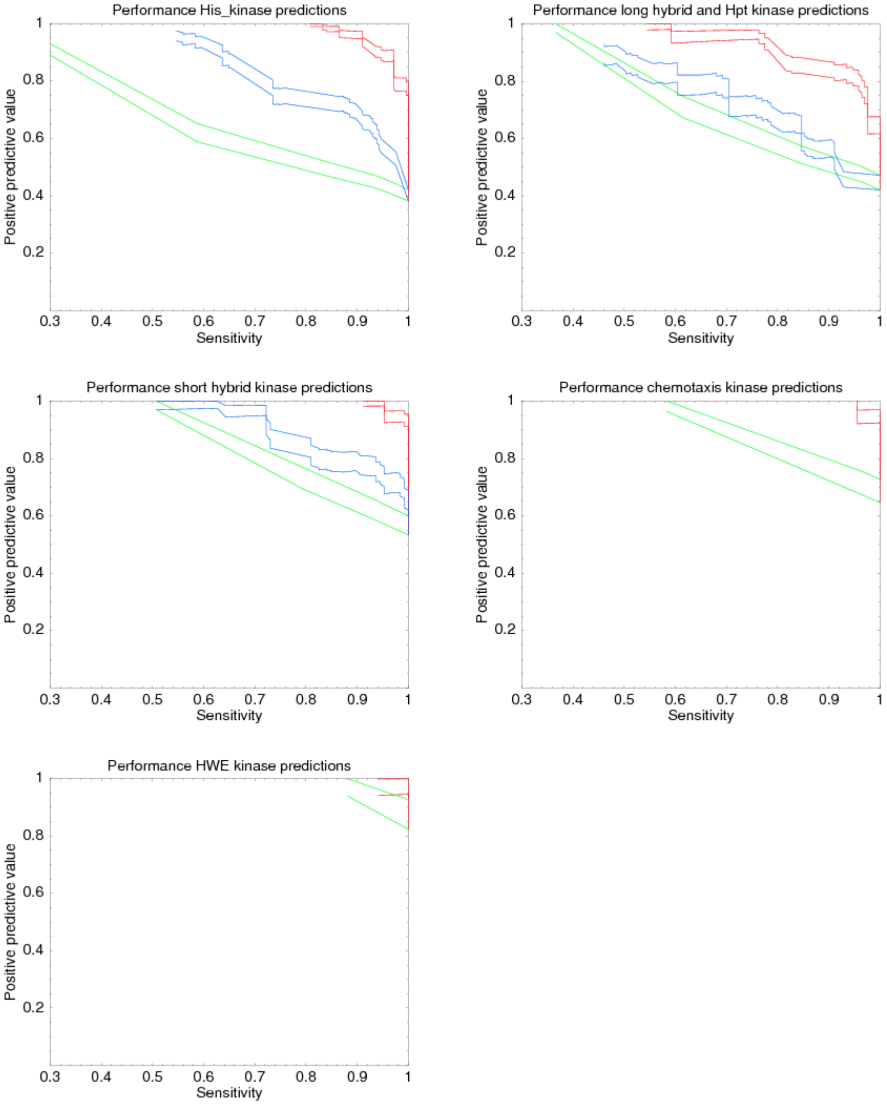
5 P-value calculation

In order to test the significance of our predictions in *Caulobacter Crescentus*, we calculated a p-value as follows. For each of the HisKA kinases with known interactions, we collected the posterior probabilities of interaction for all orphan regulators. We then sorted the entire list of all predictions by posterior and ranked each prediction, starting at rank 0 for the prediction with highest posterior. We then summed the ranks of all known interactions, obtaining the rank-sum r_{tot}, and calculated the probability $P(r \leq r_{\text{tot}})$, of getting a rank-sum r not larger than r_{tot} with random predictions (i.e. a randomly ordered list).

When the total number of predictions, n, is larger than r_{tot}, the probability $P(r \leq r_{\text{tot}})$ can be very well approximated analytically as follows. Let X_i be the rank of the known interaction i ($X_i \in \{0, .., n-1\}$) and l the total number of known interactions.Then, for $m < n$,

$$P(\sum_{i=1}^{l} X_i = m) = \frac{1}{n^l} \binom{m+l-1}{l-1} \tag{18}$$

where $\frac{1}{n^l}$ is the probability for the variables X_i to take on any value between 0 and $n-1$ and $\binom{m+l-1}{l-1}$ is the total number of possible combinations of l numbers that sum up to m. In our problem, two known interactions cannot have the same rank, but this effect should be small as the number of known interactions is small compared to the number of possible interactions. From


Performance His_kinase predictions

Performance long hybrid and Hpt kinase predictions

Performance short hybrid kinase predictions

Performance chemotaxis kinase predictions

Performance HWE kinase predictions

Figure 2: Analysis of the predictions for cognate pairs for the His_kinase (top left), long hybrid/Hpt (top right), short hybrid (middle left), chemotaxis (middle right) and HWE classes (bottom left). In all figures, the red curves show the performance of the model in which $P(D|a,T)$ is averaged over all dependence trees, the blue curve shows the performance of the model $P(D|a,T^{*})$ that uses only the best dependence tree, and the green line shows the performance of random predictions. For the chemotaxis and HWE predictions, the blue curve is not shown as it is identical to the green curve due to the fact that there are no pairs of positions with a $\log(R)$ value higher than our threshold of 10 and the sequences are thus scored with a simple position-specific weight matrix model. All pairs of curves show estimated PPV plus and minus one standard error.

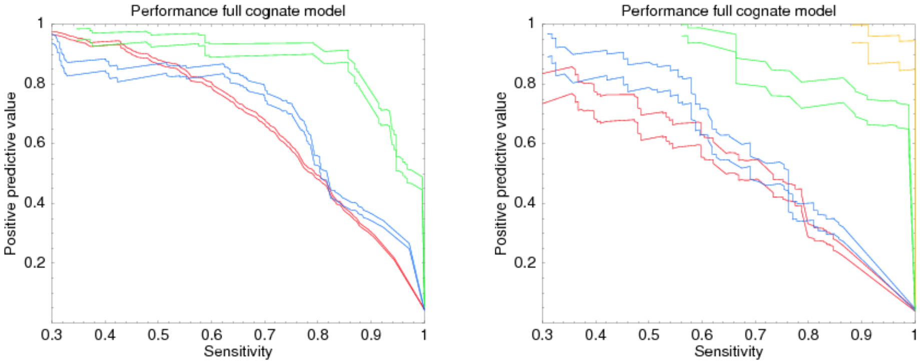


Figure 3: Reconstruction of cognate pairs when response regulators are allowed to interact with kinases of any of the 7 classes.**Left panel:** Quality of predictions for kinases of class HisKA (red line), H3 (blue line) and HisKin (green line). **Right panel:** Quality of predictions for kinases of class long hybrid/Hpt (red line), short hybrid (blue line), chemotaxis (green line) and HWE (orange line). All pairs of curves show estimated PPV plus and minus one standard error.

equation (18), we calculate the p-value,

$$P(r \leq r_{\mathrm{tot}}) = \sum_{f=0}^{r_{\mathrm{tot}}} \frac{1}{n^l} \binom{f+l-1}{l-1} \tag{19}$$

For the orphan predictions in *Caulobacter* we obtain a p-value of $7.5 \cdot 10^{-18}$. Some of the predicted pairs are found to actually lie near each other on the genome (although they were not predicted to be in the same operon, and where thus not classified as orphan pairs). If we exclude these putative cognate pairs the p-value becomes $1.1 \cdot 10^{-9}$.

6 Comparison with orphan interactions

6.1 Orphans in *Caulobacter crescentus*

The orphan kinase ChpT of *Caulobacter crescentus* only has a HisKA domain and does thus not fall into the HisKA class as defined in table 2 in the main text (ChpT does not have an ATP-binding domain). However, to increase the number of experimentally determined interactions that we could use to benchmark our predictions, we added the ChpT kinase as well as its orthologs as defined by COG (Tatusov *et al*, 1997) to our set of orphan HisKA kinases (for our predictions, we only use the HisKA domain (see above), so the absence of the ATP-binding domain does not cause any difficulties).

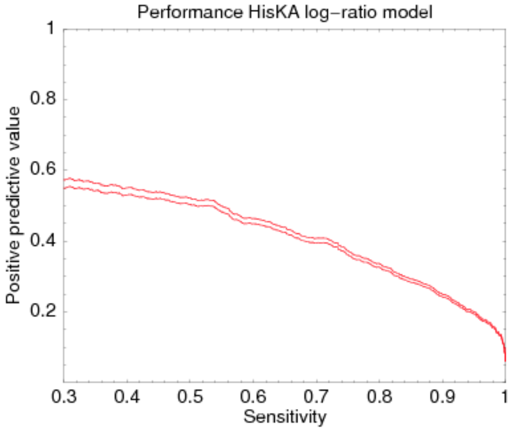


Figure 4: Reconstruction of cognate pairs with the log-ratio model that is used to predict network structure. The curves show estimated PPV plus and minus one standard error.

kinase	regulator	posterior	se	exp evidence
HP0244	HP0703	0.9427	0.0485	(Beier and Frank, 2000)
HP0244	HP1043	0.05336	0.0487	(Beier and Frank, 2000)
HP0244	HP1021	0.0039	0.0022	
HP0244	HP1067	0	0	
HP0244	HP0616	0	0	
HP0244	HP0393	0	0	
HP0244	HP0019	0	0	

Table 1: Predictions for the one orphan HisKA kinase in *Helicobacter pylori* for which an interaction is known. There are 7 orphan regulators in *H. pylori* and we show the posterior probabilities for all of them. Posterior probabilities and their standard errors were calculated over 20 sampling runs.

6.2 Additional orphan interactions

Besides *Caulobacter crescentus* that accounts for the largest part of known orphan interactions, there are three more species with experimentally determined orphan interactions involving HisKA kinases, namely *Helicobacter pylori*, *Bacillus subtilis* and *Ehrlichia chaffeensis*. Our predictions for these species are shown in tables 1, 2 and 3. As in table 1 in the main paper, the list of predictions is shown ordered by posterior, up to and including all known interactions. Correct predictions are shown in green, incorrect predictions (at odds with the experimental results) are shown in red. All other predictions are shown in black. Posterior probability and standard error of the posterior probability over 20 sampling runs are shown for each prediction.

In *H. pylori* the known interaction matches the top prediction of the algo-

kinase	regulator	posterior	se	exp evidence
KinA	Spo0F	0.0361	0.0060	(Piggot and Hilbert, 2004)
KinB	CheV	0.7929	0.0348	
KinB	Spo0A	0.1649	0.0256	
KinB	YneI	0.0412	0.0294	
KinB	Spo0F	0.0006	0.0004	(Piggot and Hilbert, 2004)
KinC	Spo0F	0.6765	0.0731	(Piggot and Hilbert, 2004)
KinD	YneI	0.5215	0.0975	
KinD	Spo0F	0.2840	0.0692	(Piggot and Hilbert, 2004)
KinE	Spo0A	0.4516	0.0768	
KinE	YneI	0.3751	0.0972	
KinE	CheV	0.1649	0.0366	
KinE	Spo0F	0.0028	0.0008	(Piggot and Hilbert, 2004)

Table 2: Predictions for orphan HisKA kinases with known interactions in *B. subtilis*. There are 6 orphan regulators in total in *B. subtilis*. For every known interaction shown there are several kinds of evidence, see (Piggot and Hilbert, 2004). Posterior probabilities and their standard errors were calculated over 20 sampling runs.

kinase	regulator	posterior	se	exp evidence
ECH_0299(NtrY)	ECH_0339(NtrX)	1	0	(Kumagai *et al*, 2006)
ECH_0885(PleC)	ECH_1012(CtrA)	1	0	(Kumagai *et al*, 2006)
ECH_0885(PleC)	ECH_0773(PleD)	0	0.2236	(Kumagai *et al*, 2006)

Table 3: Predictions for the two orphan HisKA kinases with known interactions in *E. chaffeensis*. There are 3 orphan regulators in total in *E. chaffeensis*. Posteriors and their standard errors were calculated over 20 sampling runs.

rithm which is assigned a 94% posterior probability.

In *B. subtilis* it is known that the regulator Spo0F interacts with all Kin kinases, i.e. KinA, KinB, KinC, KinD, and KinE. Indeed we predict that Spo0F interacts with all these kinases with nonzero probability. The interaction probabilities of Spo0F with all other kinases is zero (data not shown). Table 2 shows, however, that the fraction of time Spo0F is associated with each of these kinases varies significantly across the different Kin kinases, with Spo0F associating with KinC more than 65% of the time, with KinD 28% of the time and only roughly 4% with the other Kin kinases. Note also that some of the Kin kinases are predicted to interact with other regulators as well.

For kinase ECH_0299 of *E. chaffeensis* (an ortholog of NtrY), we correctly predict that it interacts only with ECH_0339 (an ortholog of NtrX). Kinase ECH_0885 is the only example where our predictions clearly disagree with the experimental evidence. Whereas the experimental evidence suggests that ECH_0885 interacts only with ECH_0773, we assign 100% posterior probability to ECH_0885 interacting with ECH_1012.

	CK	OK	-
CR	9.184	0.009	3.59
OR	0.015	0.055	1.02
-	1.326	0.346	382.5

Table 4: Ortholog statistics for cognate pairs. For each cognate kinase/receiver pair and each of the 398 other genomes, there can be either: no orthologs for both (-,-), two orthologs that form a cognate pair (CK,CR), no ortholog for the kinase and an ortholog for the receiver which is an orphan receiver (-,OR), etcetera. The table shows the average number of times each of the 9 possible combinations occurs for cognate kinase/receiver pairs.

7 Ortholog Statistics

Our predictions suggest that orphan kinases interact predominantly with orphan regulators, that cognate kinases interact predominantly with cognate regulators, and that there is relatively little interaction between orphan kinases and cognate regulators or between cognate kinases and orphan regulators. Since orphans and cognates almost certainly share a common phylogenetic ancestry, we decided to investigate to what extent cognates and orphans change class on relatively short evolutionary time scales. To this end we determined orthologous genes for each cognate kinase/regulator pair, for each orphan kinase, and for each orphan regulator.

Table 4 shows the ortholog statistics for cognate pairs. For each cognate kinase/regulator pair there are 9 possibilities for its orthologs in each of the 398 other genomes varying from the cognate pair mapping to another cognate pair in the other genome, to absence of orthologs for both genes in the pair. The table shows the average number of occurrences of each of the 9 possibilities.

The table shows that in on average over 380 genomes there are no orthologs for either gene. The next most common occurrence is that the cognate pair maps to a cognate pair (in on average 9.184 genomes). After that it is by far most likely that only one of the two genes has an ortholog. In all cases cognates are significantly more likely to map to cognates than to orphans.

Similarly, for each orphan kinase we counted the number of times that it has no ortholog in each of the 398 other genomes, the number of times the ortholog is itself an orphan, and the number of times the ortholog is part of a cognate pair. Finally, for each orphan receiver we counted the number of times it has no ortholog in each of the other genomes, the number of times its ortholog is an orphan, and the number of times its ortholog is part of a cognate pair. These orphan ortholog statistics are shown in table 5.

The table shows that for both orphan kinases and for orphan receivers there are on average a handful of genomes with orthologs. In both cases, if there is an ortholog, it is much more likely to be an orphan as well.

	Orphan	Cognate	-
Orphan kinase	3.78	0.61	393.6
Orphan receiver	4.595	1.153	392.25

Table 5: Ortholog statistics for orphans. For both orphan kinases and orphan receivers, the table shows how many of 398 other genomes on average have: an ortholog that is also an orphan, an ortholog that is part of a cognate pair, or no ortholog.

8 Prediction of polyketide synthase interactions: classification model

In order to compare the quality of our predictions to the simple classification scheme proposed in (Thattai *et al*, 2007), we calculated posterior probabilities of interaction using only the information about the class membership of the head and tail sequences as follows. We used the annotation of (Thattai *et al*, 2007) to label every head (tail) as H1 (T1), H2 (T2), H3 (T3) or as 'unclustered'. For a given head sequence of class H_i of a genome g, we assign an interaction probability of 0 to all tails of classes T_j with $j \neq i$ and a probability of $1/n_g^i$, where n_g^i is the number of tails of class i of genome g, to all tails of class i of genome g. If the head belongs to the class of unclustered heads, it is assigned a probability of $1/n_g$ to interact with any of the n_g tails of genome g. In other words, for the H1, H2 and H3 classes, each head sequence can only interact with tail sequences of the correct corresponding tail class, but within the corresponding class, every tail is equally likely to be an interaction partner. Heads that are unclustered can interact with any tail of the same genome with equal probability.

References

Beier D, Frank R (2000) Molecular characterization of two-component systems of *Helicobacter pylory. J Bacteriol* **182**: 2068–2076

Cerquides J, de Màntaras RL (2003) Tractable Bayesian learning of tree augmented naive bayes classifiers. *Proceedings of Twentieth International conference on Machine Learning*

Grebe T, Stock J (1999) The histidine protein kinase superfamily. *Adv Microb Physiol* **41**: 139–227

Koretke K, Lupas A, Warren P, Rosenberg M, Brown J (2000) Evolution of two-component signal transduction. *Mol Biol Evol* **17**: 1956–1968

Kumagai Y, Cheng Z, Lin M, Rikihisa Y (2006) Biochemical activities of three pairs of *Ehrlichia chaffeensis* two-component regulatory system proteins involved in inhibition of lysosomal fusion. *Infect Immun* **74**: 5014–5022

Piggot PJ, Hilbert DW (2004) Sporulation of *Bacillus subtilis*. *Curr Opin Microbiol* **7**: 579–586

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* **278**: 631–637

Thattai M, Burak Y, Shraiman BI (2007) The origins of specificity in polyketide synthase protein interactions. *PLoS Comp Biol* **3**: e186