

Inferring protein-protein interactions from amino acid sequences: Application to two-component systems

Most of this can be found in:

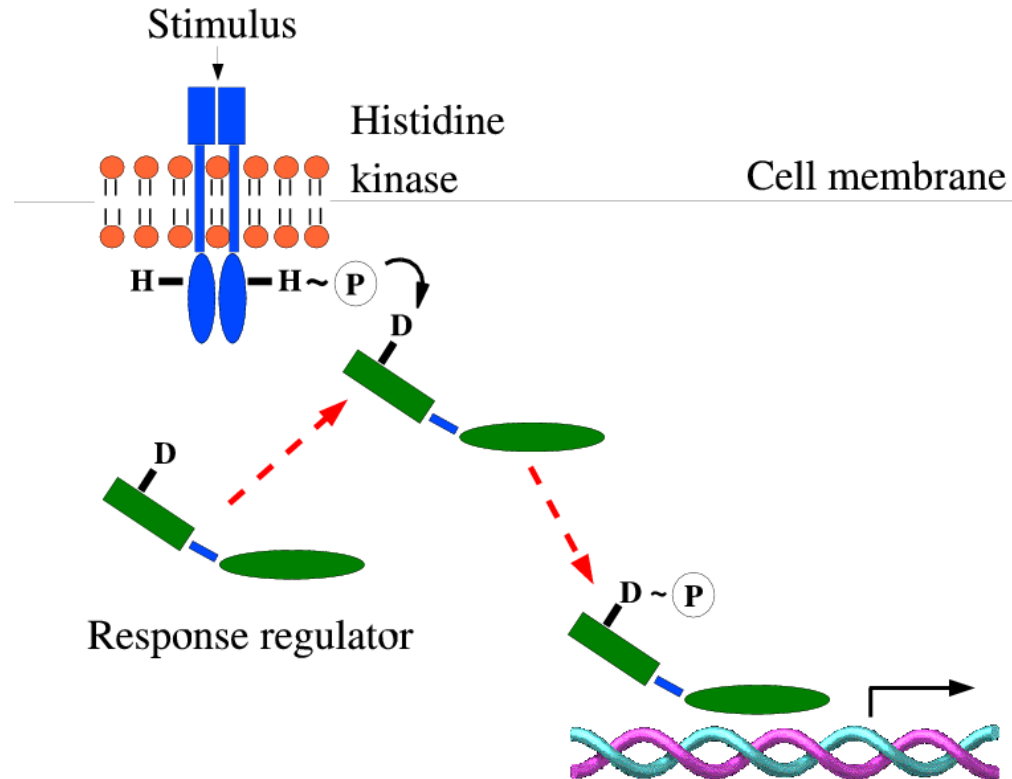
A Bayesian algorithm for reconstructing two-component signaling networks
Lukas Burger and Erik van Nimwegen
Proceedings 6th International Workshop, WABI 2006, Zurich, Switzerland
Lecture Notes in Computer Science , Vol. **4175** , 2006, pp. 44-55

Lukas Burger and Erik van Nimwegen



*Division of Bioinformatics
Biozentrum, Universität Basel,
Swiss Institute of Bioinformatics*

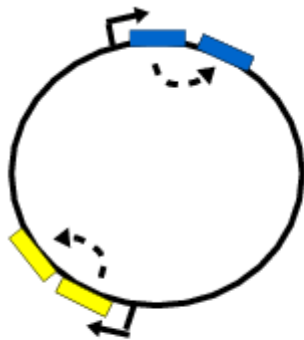
Bacterial signaling: Two-component systems



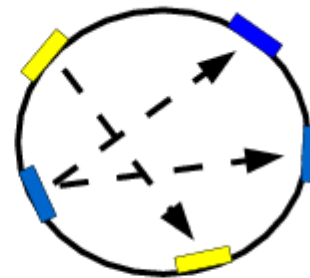
- Responsible for most signal transduction in bacteria.
 - *E. coli*: aerobic/anaerobic switch, chemotaxis
 - *B. subtilis*: sporulation
 - *C. crescentus*: cell-cycle/differentiation.
- Over 8,500 two-component system genes in 399 sequenced bacterial genomes.

Advantages as model system for predicting protein-protein interactions

- Two-component systems can be easily detected using hidden Markov models of the kinase and receiver domains.
- Enough homology to reproduce reliable multiple alignments: specificity of interaction likely lies in details of amino acids at surface.
- Large number of examples available (good statistics).
- *Training sets*: For about 50% we *know* which kinase interacts with with receiver because they lie in a common operon.



cognate pairs



orphan kinases and regulators

Extracting two-component system proteins

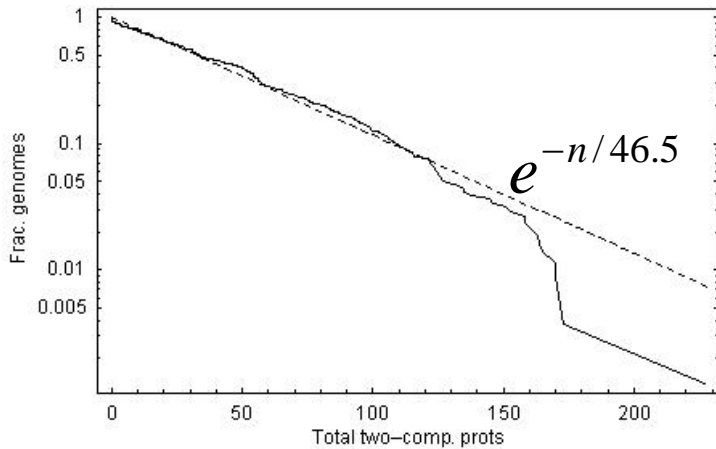
- Receivers: All HMMer hits to Pfam profile Response_reg.
- Kinases: All HMMer hits to the Pfam profiles:
 - HisKA, H2, H3, His_kinase, HWE_HK, HATPase_c, HPT

Name	Architecture	no.cognates	no.orphans
HisKA	HisKA, HATPase_c	3388	2158
H3	H3, HATPase_c	636	183
His_kinase	HATPase_c	245	23
Long hybrid	HisKA, HATPase_c, RR, (RR), Hpt	132	286
Short hybrid	HisKA, HATPase_c, RR, (RR)	126	985
Chemotaxis	Hpt, HATPase_c	89	77
Hpt	Hpt	37	192
HWE	HWE or H2, HATPase_c	34	162

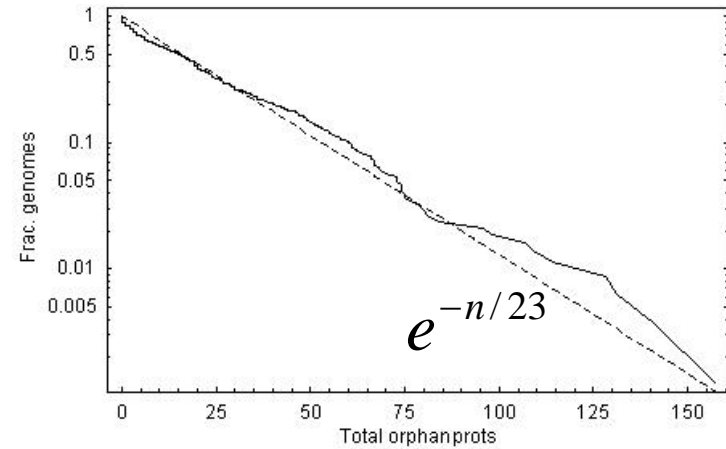
Table 1: Pfam domain combinations of the most abundant kinase architectures and the numbers of their occurrence in both cognates and orphans. Both the short and long hybrid architecture can contain one or two receiver domains.

Global statistics 'signaling networks' in bacteria

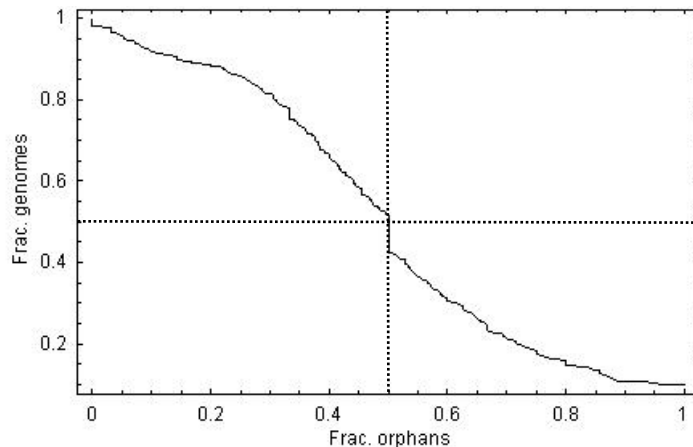
Total number of two-comp. proteins



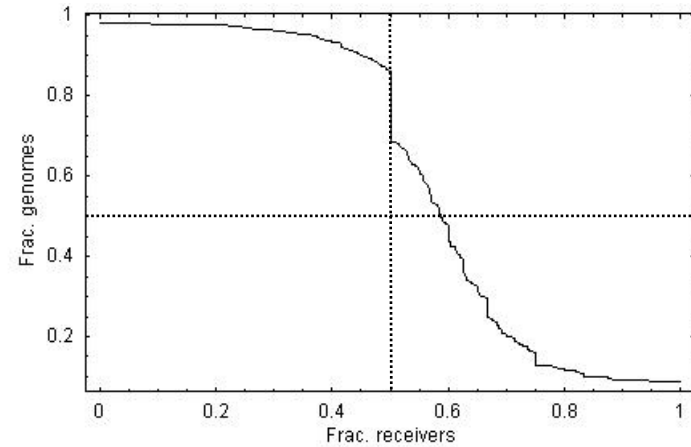
Total number of orphans



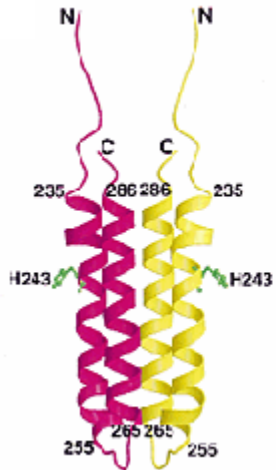
Fraction of all two-comp. that are orphans



Fraction of receivers among orphans



Multiple Alignments



```

..ELMRAVSHEL RTPVAR..
..QLIRGVAHEIKNPLGG..
..VFIANA AHELRTPLTA..
..RFTADASHELRSPLSA..
..QFVADAAHEL RTPITA..
..ALLSSVSHDLRSPLAA..
..IMLAGISHDLRTPLTR..
..DFVDNISHELRTPLTV..
..SFAADVSH ELKNPLTS..
..GLAAAAHELGTPLAT..

```

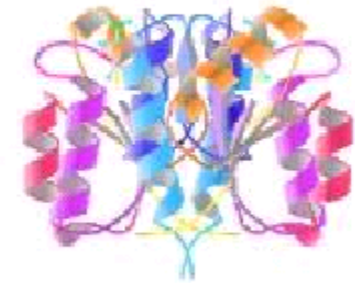
kinase domain alignment

```

..LVVLDVMLPGADGLTV..
..VILTDIRMPGIDGLTF..
..LVLLDLGLPELEGLDV..
..CILLDRGLPKSSGDQV..
..LVLLDLGLPMDGMQV..
..LLILDGLPDMDGAEV..
..LIVLDFMLPVEDGLSI..
..LILLDWMLPGVSGVDL..
..LAIFDIKMPRMDGMEL..
..YAVVDLRLADGNGLV..

```

receiver domain alignment



- Aligned all receivers together. Separate alignment for each kinase domain architecture.
- Produced multiple alignments by aligning to the Pfam profiles.
- Produced independent alignments with ProbCons.
- Keep only positions with greater than 80% agreement by the two methods and less than 50% gaps.
- Gives alignments with *absolute* reference positions for all kinase and receiver sequences.

Final number of positions used:

Receiver domain: 115.

HisKA: 64, H3: 66, Hiskin: 80, Hpt: 84, SH: 66, LH: 84, Chemotaxis: 101, HWE: 83.

Classifying receivers

- For each class c , each position i and each amino acid α estimate the *weight matrix* of probabilities to obtain letter α at position i of a receiver in class c from the cognate kinase/receiver pairs.

Weight matrix:

$$w_{\alpha i}^c = \frac{n_{\alpha i}^c + \lambda}{n^c + 21\lambda}$$

Probability receiver given class:

$$P(\vec{R} | c) = \prod_i w_{R_i i}^c$$

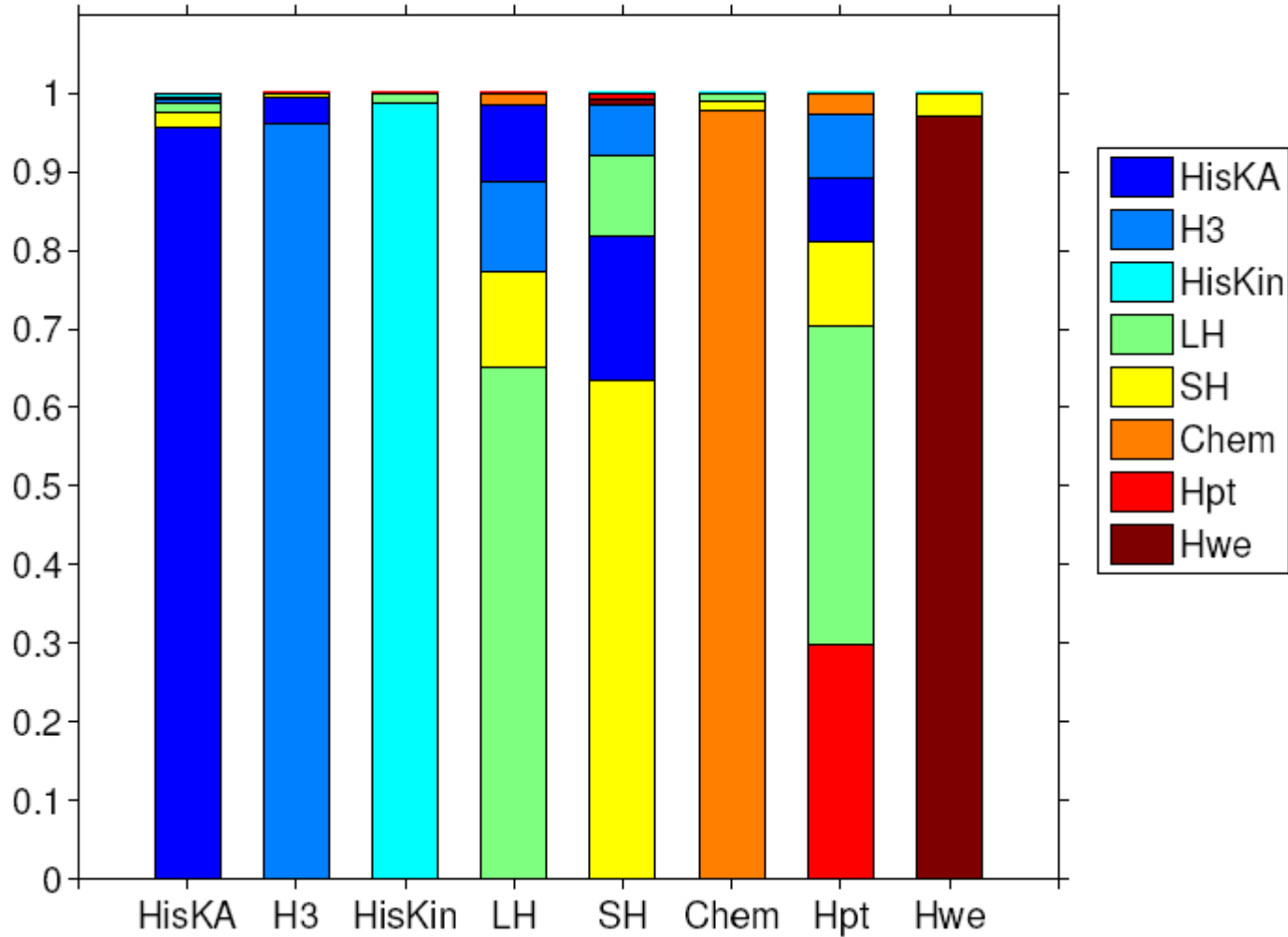
Posterior class given receiver:

$$P(c | \vec{R}) = \frac{P(\vec{R} | c)P(c)}{\sum_{\tilde{c}} P(\vec{R} | \tilde{c})P(\tilde{c})}$$

Some technical details:

- To correct for phylogenetic sampling biases we single-link cognates into clusters of similarity $\geq 90\%$ and give a weight of 1 to each cluster.
- λ is set to $\frac{1}{2}$ (Prior uniform in Fisher-information).
- Gaps are treated as 21st amino acid.
- To get cognates we consider genes separated by ≤ 50 bp on same strand in same operon and take only cognates if there is only 1 kinase and 1 receiver in the operon.
- We use a uniform prior $P(c)$.
- We take out the receiver and all members of its cluster from the counts in the WM when scoring a given receiver.

Results Classifying receivers



Overall 94% of cognate receivers are classified correctly

Interacting amino acids

$\vec{k} = (k_1, k_2, \dots, k_n)$ Amino acid sequence of a kinase.

$\vec{r} = (r_1, r_2, \dots, r_m)$ Amino acid sequence of a receiver.

$(\vec{k}, \vec{r}) = \vec{s}$ Joint sequence of interacting kinase/receiver pair.

$P(\vec{k}, \vec{r}) \equiv P(\vec{s})$ Joint probability of observing the joint sequence for an *interacting* kinase/receiver pair.

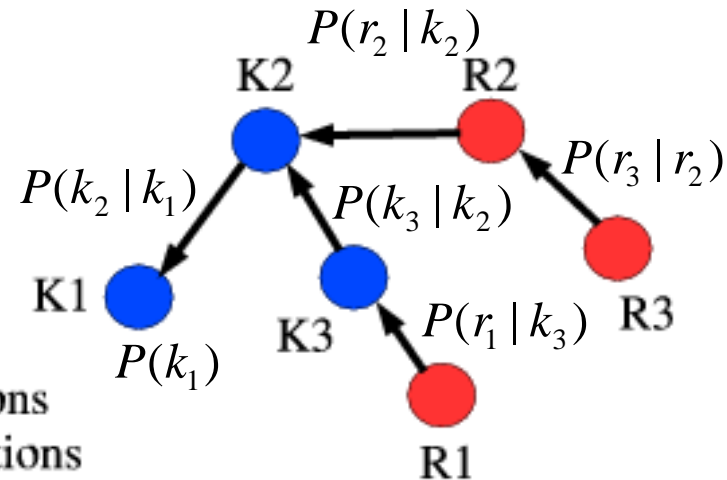
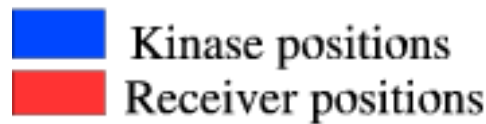
We will factor joint distribution into conditional probabilities of *pairs* of positions.

$$P(\vec{s}) = P(s_r) \prod_{i \neq r} P(s_i | s_{\pi(i)})$$

s_r = Amino acid at root.

$\pi(i)$ = position that amino acid at position i depends on.

Example:



Probability of the data given a dependence tree topology

$$P(s_i | s_j) = \rho_{s_i s_j}^{ij} \quad \text{Parametrization of the conditional probabilities.}$$

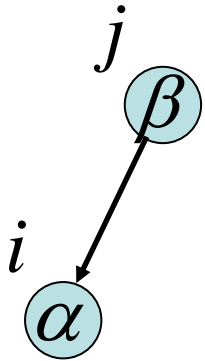
$n_{\alpha\beta}^{ij}$ Number of times amino acid combination α, β occurs at positions i, j in the data.

$$P(D | \rho, \pi) = \left[\prod_{\alpha} (\rho_{\alpha}^r)^{n_{\alpha}^r} \right] \prod_i \left[\prod_{\alpha\beta} (\rho_{\alpha\beta}^{i\pi(i)})^{n_{\alpha\beta}^{i\pi(i)}} \right] \quad \text{Likelihood}$$

$$P(\rho | \pi) \propto \left[\prod_{\alpha} (\rho_{\alpha}^r)^{2^{1\lambda}-1} \right] \prod_i \left[\prod_{\alpha\beta} (\rho_{\alpha\beta}^{i\pi(i)})^{\lambda-1} \right] \quad \text{Prior}$$

$$P(D | \pi) = \int P(D | \rho, \pi) P(\rho | \pi) d\rho \quad \text{Likelihood of the dependence tree}$$

Probability given dependence tree topology



$$P(D_i | \pi, D_j) = \prod_{\beta} \left[\Gamma(21\lambda) \int \prod_{\alpha} \frac{(\rho_{\alpha\beta}^{ij})^{n_{\alpha\beta}^{ij}}}{\Gamma(\lambda)} d\rho_{\alpha\beta}^{ij} \right] =$$

$$\prod_{\beta} \left[\frac{\Gamma(21\lambda)}{\Gamma(n_{\beta}^j + 21\lambda)} \prod_{\alpha} \frac{\Gamma(n_{\alpha\beta}^{ij} + \lambda)}{\Gamma(\lambda)} \right] \equiv M_i R_{ij}$$

Marginal and edge probabilities:

$$M_i = \prod_{\alpha} \frac{\Gamma(n_{\alpha}^i + 21\lambda)}{\Gamma(21\lambda)}, \quad R_{ij} = \prod_{\alpha\beta} \frac{\Gamma(n_{\alpha\beta}^{ij} + \lambda) \Gamma(21\lambda) \Gamma(21\lambda)}{\Gamma(\lambda) \Gamma(n_{\alpha}^i + 21\lambda) \Gamma(n_{\beta}^j + 21\lambda)}$$

Final expression:

$$P(D | \pi) = \frac{\Gamma(21^2 \lambda)}{\Gamma(n + 21^2 \lambda)} \left[\prod_i M_i \right] \left[\prod_{i \neq r} R_{i\pi(i)} \right]$$

Maximal Likelihood dependence tree

$$P(D | \pi) = \frac{\Gamma(21^2 \lambda)}{\Gamma(n + 21^2 \lambda)} \left[\prod_i M_i \right] \left[\prod_{i \neq r} R_{i\pi(i)} \right]$$

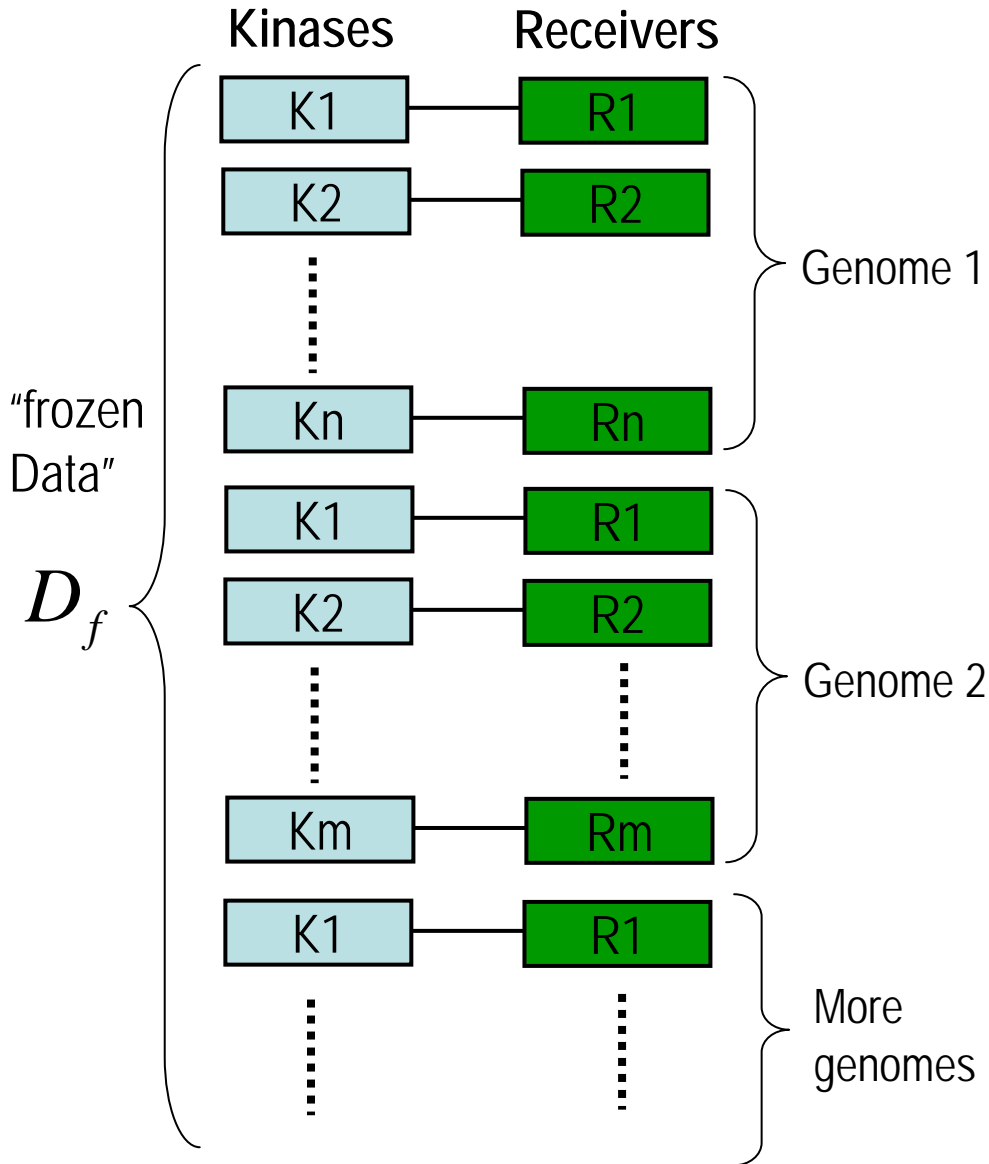
π_* Dependence tree that maximizes $P(D | \pi)$

Chow-Liu algorithm: Start with a complete graph with weights R_{ij} on edges (i,j) . One can find the *maximal spanning tree* of this graph by a simple greedy procedure.

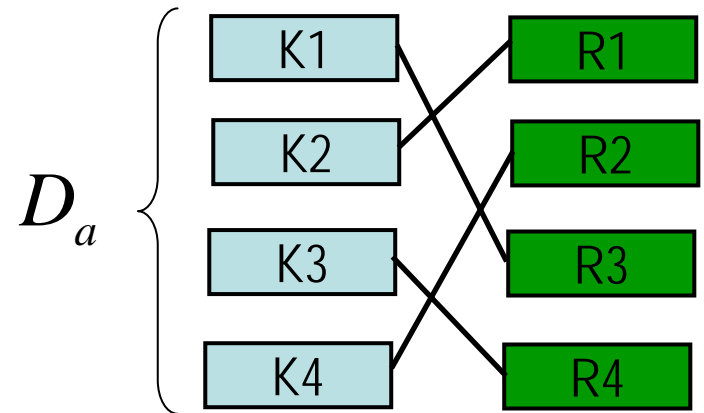
Test: predicting interaction partners for cognate pairs.

- Use dataset D of all cognate pairs to determine best dependence tree.
- For each genome, remove all its cognate pairs from the data-set, and remove also all single-linkage clusters of cognates with $\geq 90\%$ amino acid identity.
- 'Freeze' all other cognate kinase/receiver pairs (training set).
- Use Monte-Carlo to search over all *assignments* of kinase/receiver pairs for the genome under study.
- Probability of assignment is probability of joint data (frozen pairs + assignment).

Predicting Cognate interacting pairs



Genome being sampled:



Likelihood assignment:

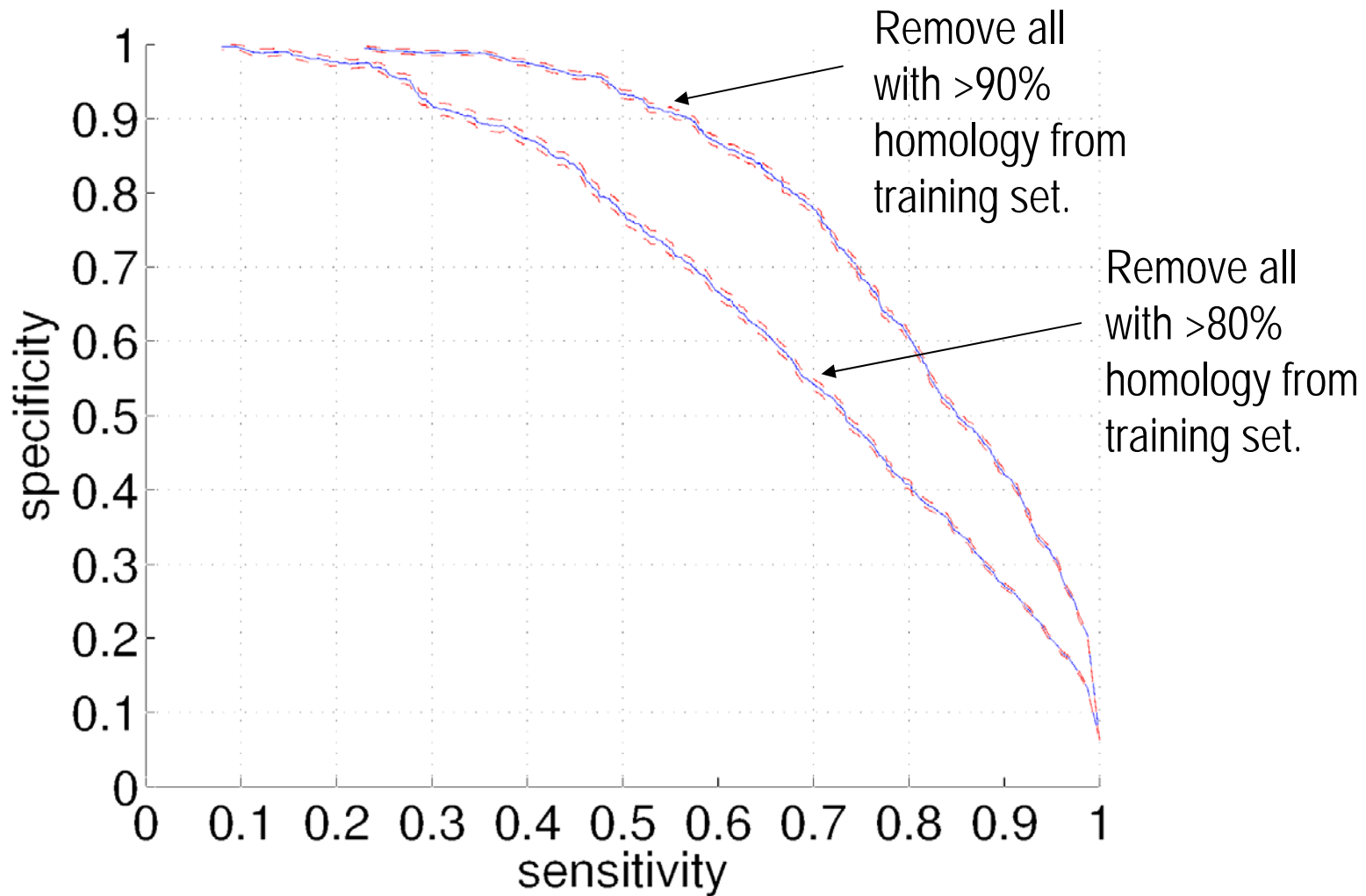
$$P(D_f, D_a | \pi_*, a)$$

Monte-Carlo sampling according to likelihood:

At every time-point pick 2 kinases at random and consider flipping the receivers assigned to them.

Predicting Cognate interacting pairs

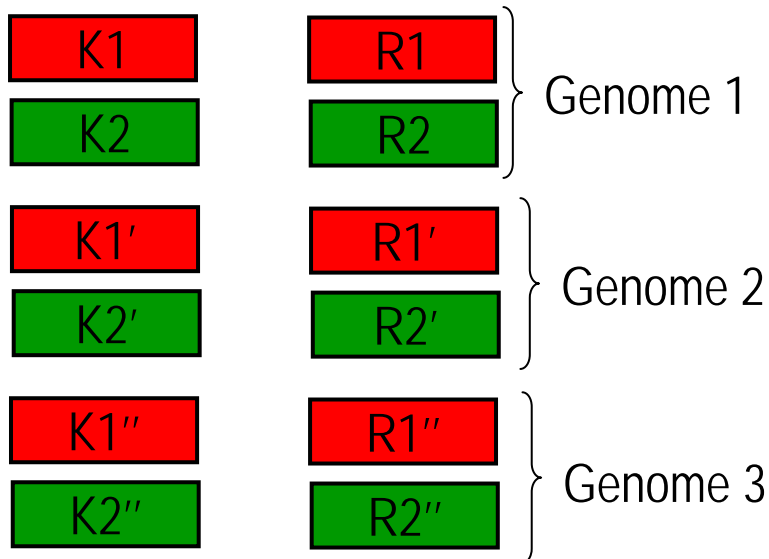
Fraction of predictions that are true cognate pairs.



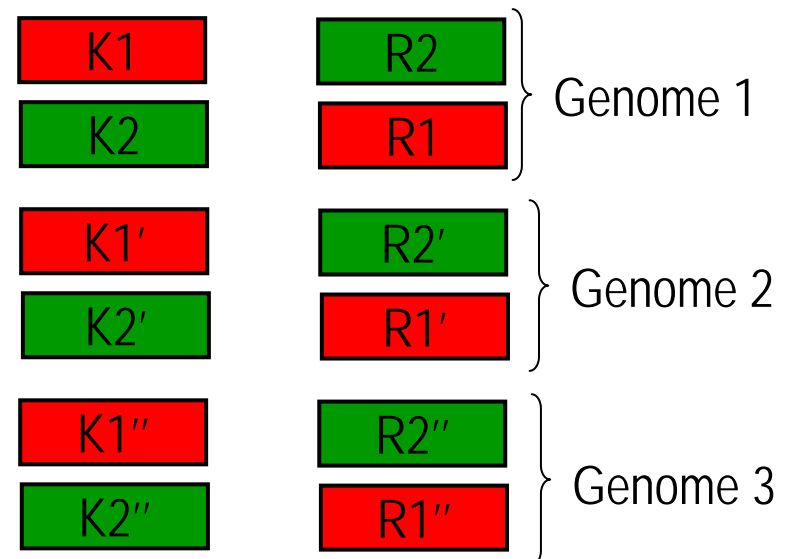
Phylogeny signal vs. physical interactions

- Do positions in kinase receiver just *look* correlated because there are orthologous interacting pairs are evolutionarily related?
- Make a new data-set of `false interacting pairs' that have the exact same evolutionary relationships: Flip assignment of pairs for orthologous groups.

True interacting pairs

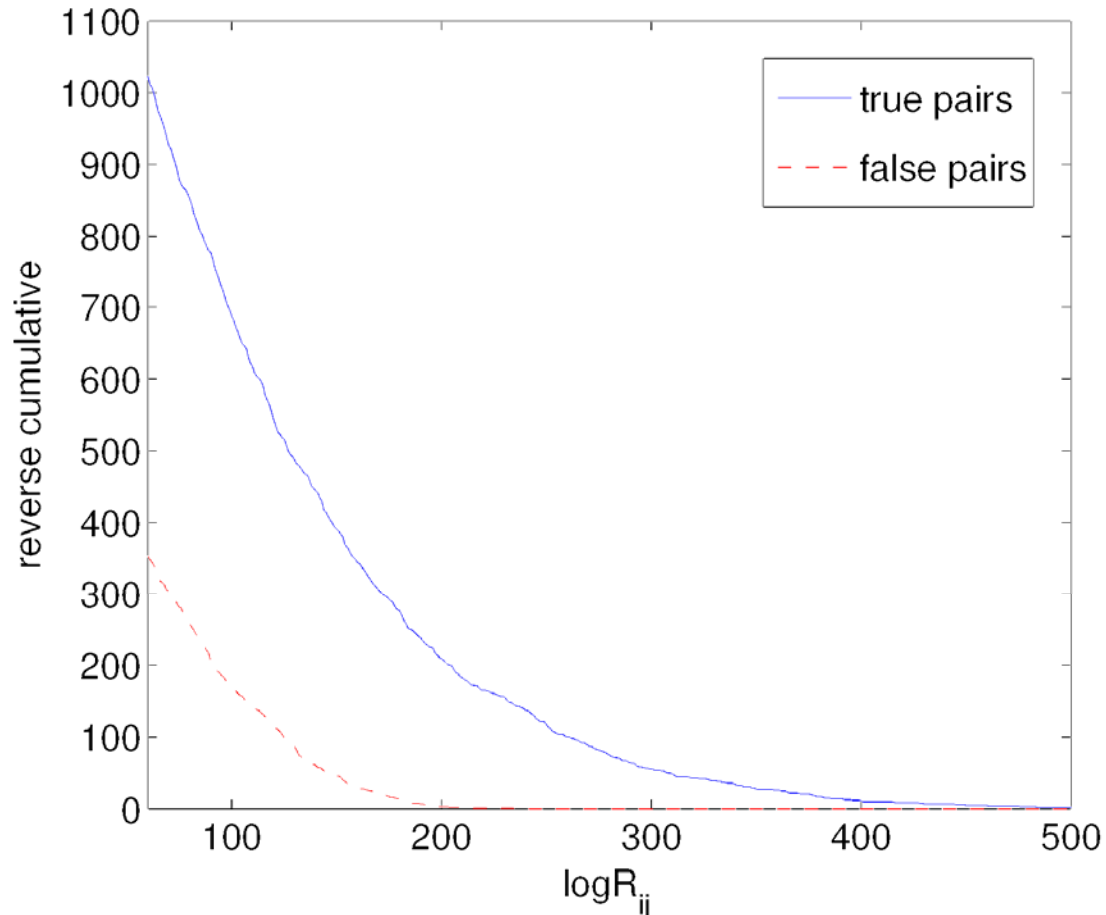


False interacting pairs



Phylogeny signal vs. physical interactions

Number of pairs of positions with 'interaction score' $\log(R_{ij})$ over a given cut-off in true and false pairs.



Predicting orphan interactions

Extensions for doing orphan predictions:

- Use entire set of cognates as a `frozen' training set.
- Run assignments of all orphans from all genomes at the same time.
- Run on all classes of kinases at the same time.
- Since the number of kinases and receivers are not the same, some kinases or receivers will not be hooked up to any partner.
- Each receiver can belong to each of the kinase classes (sampled over).
- Unhooked kinases/receivers are scored according to the simple WM model: one WM for each class of receiver and one for each class of kinase.
- Move-set includes changes in receiver class, and flips of partner (where one of the two members may not have a partner).


Results predicting orphan interactions


Focused on *Caulobacter crescentus* for which most results are available.

C. crescentus has 11 orphan HisKA kinases and 19 orphan receivers.

kinase	regulator	posterior	std	exp evidence
CC0248	CC0247	1.0000	0.0000	putative cognate pair
CC0289	CC0294	0.9931	0.0078	cognate pair, <i>in vitro</i> phosphorylation [13]
CC2765	CC2766	0.9902	0.0154	cognate pair, <i>in vitro</i> phosphorylation [13]
CC2932	CC2931	0.9759	0.0213	putative cognate pair
CC2755	CC2757	0.8681	0.1613	putative cognate pair
CenK	CenR	0.7980	0.1601	<i>in vitro</i> phosphorylation [13]
ChpT	CpdR	0.7766	0.1453	<i>in vitro</i> phosphorylation [5]
pleC	DivK	0.6639	0.2088	<i>in vitro</i> phosphorylation [13]
CckN	CtrA	0.6379	0.1606	not known
DivL	CC0588	0.6187	0.2272	not known
DivJ	PleD	0.4890	0.2190	<i>in vitro</i> phosphorylation [13]
CckN	PleD	0.2834	0.1143	not known
DivJ	CtrA	0.2147	0.1269	<i>in vitro</i> phosphorylation [18]
PleC	PleD	0.1650	0.1332	<i>in vitro</i> phosphorylation [13]
CenK	CC0588	0.1260	0.0913	false positive, <i>in vitro</i> phosphorylation [13]
DivJ	DivK	0.1045	0.0401	<i>in vitro</i> phosphorylation [13]
DivL	DivK	0.0964	0.0693	yeast two-hybrid screen [10]
CC2755	CC0588	0.0948	0.1424	not known
DivJ	CenR	0.0881	0.0615	false positive, <i>in vitro</i> phosphorylation [13]
ChpT	DivK	0.0877	0.1256	false positive, <i>in vitro</i> phosphorylation [5]
DivJ	CC0588	0.0638	0.0214	false positive, <i>in vitro</i> phosphorylation [5]
DivL	CtrA	0.0637	0.0357	<i>in vitro</i> phosphorylation [18]
CenK	CC0432	0.0608	0.1216	not known
DivL	PleD	0.0574	0.0600	not known
DivL	CC0432	0.0479	0.0958	not known
ChpT	CtrA	0.0415	0.0324	<i>in vitro</i> phosphorylation [5]
CckN	DivK	0.0407	0.0345	yeast two-hybrid screen [10]

 Experimentally confirmed

 Interaction observed in yeast two-hybrid screen.

 No data available either for or against.

 Interaction not observed experimentally when tested.

Probability to get this match by chance:

$$p = 1.4 * 10^{-13}$$

Top predictions for hisKA kinases with known interactions. (of 319 possible interactions). List was cut to have at least 1 prediction for all kinases for which an interaction is known.

Results predicting orphan interactions

One hisKA orphan interaction known in *Helicobacter pylori*.

kinase	regulator	posterior	std	exp evidence
HP0244	HP0703	1	0	<i>in vitro</i> phosphorylation [4]
HP0244	HP1067	0	0	
HP0244	HP1043	0	0	
HP0244	HP1021	0	0	
HP0244	HP0616	0	0	
HP0244	HP0393	0	0	
HP0244	HP0019	0	0	

Known interaction is predicted with posterior 1.

There are 7 orphan receivers in *Helicobacter pylori*.

Beyond maximum-likelihood spanning tree: summing over spanning trees



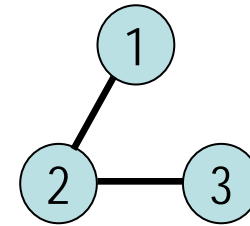
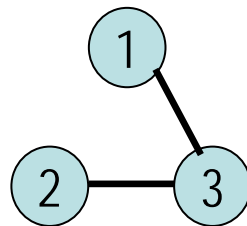
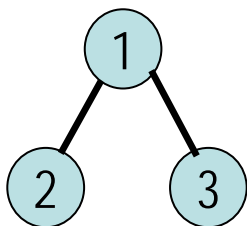
Probability of the data given a spanning dependence tree:

$$P(D | \pi) = \frac{\Gamma(21^2 \lambda)}{\Gamma(n + 21^2 \lambda)} \left[\prod_i M_i \right] \left[\prod_{i \neq r} R_{i\pi(i)} \right]$$

We would like to *sum* over all possible trees:

$$P(D) = \sum_{\pi} \frac{P(D | \pi)}{|\pi|} = \frac{\Gamma(21^2 \lambda)}{|\pi| \Gamma(n + 21^2 \lambda)} \left[\prod_i M_i \right] \sum_{\pi} \left[\prod_{i \neq r} R_{i\pi(i)} \right]$$

Example: for 3 positions we would sum over the three spanning trees:



$$P(D) \propto R_{12}R_{13} + R_{13}R_{23} + R_{12}R_{23}$$

Generalization of the matrix-tree theorem

Define Laplacian matrix: $L_{ij} = \delta_{ij} \sum_k R_{ik} - R_{ij}$

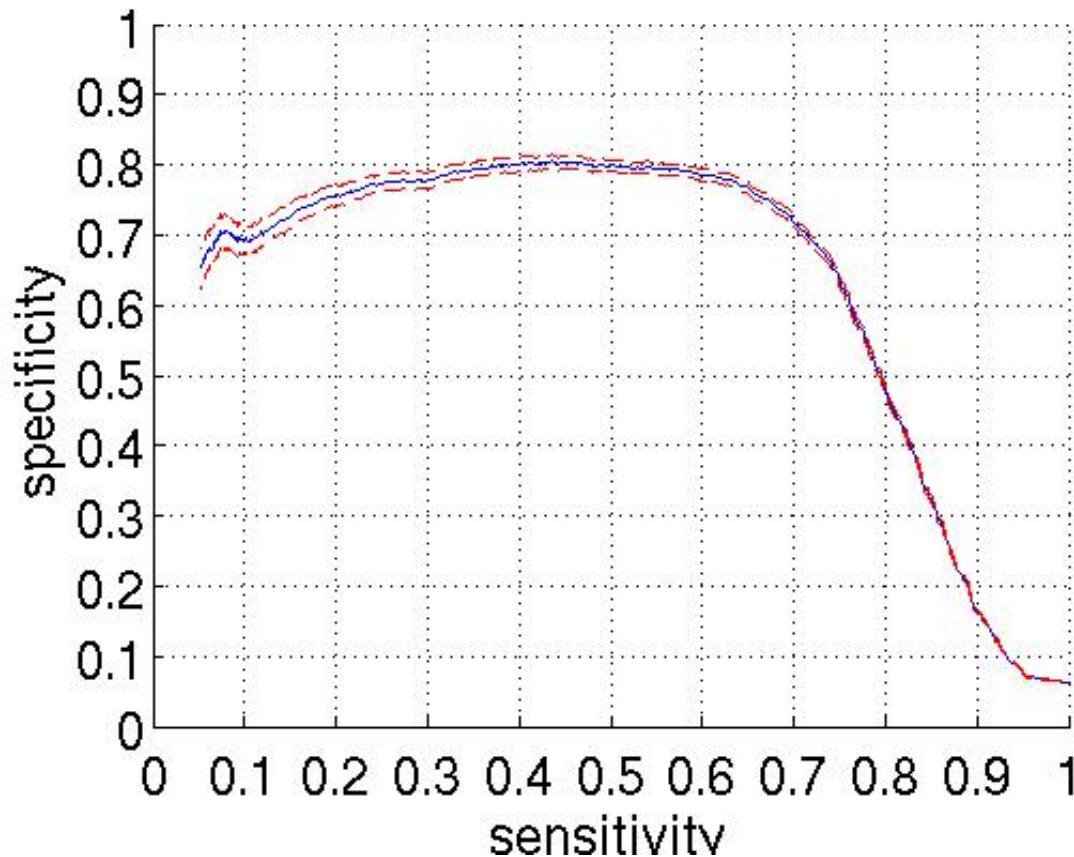
$\tilde{L} =$ Matrix L with a single row and column removed

Theorem:
$$\sum_{\pi} \left[\prod_i R_{i\pi(i)} \right] = \det(\tilde{L})$$

One catch: R has some components that are extremely large and others that are extremely small. We so far have found no numerically stable way of calculating the determinant, only uncontrolled approximations.

Test:

- Take all cognate kinases and receivers and run Monte-Carlo assigning all genomes at the same time.
- Score the combination of assignments from all genomes using the determinant.
- **Note:** No training set!

Results predicting cognate interactions *ab initio*

Caveat: We cannot show that the Monte-Carlo has converged (and believe it has in fact not).