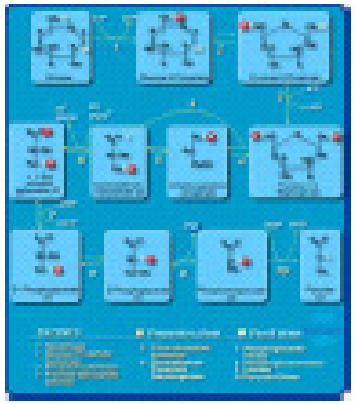
Ab initio prediction of protein interaction

Dirk Husmeier
Biomathematics & Statistics Scotland (BioSS)
JCMB, The King's Buildings, Edinburgh EH9 3JZ
United Kingdom
http://www.bioss.ac.uk/∼dirk

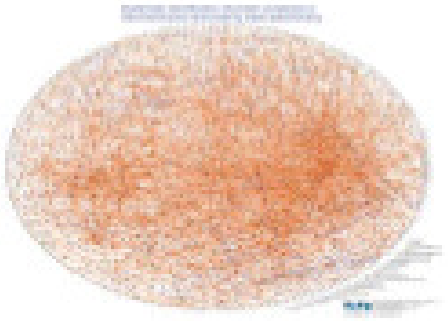
Pathways and systems biology

Glycolysis



Metabolic Pathways

Protein-Protein

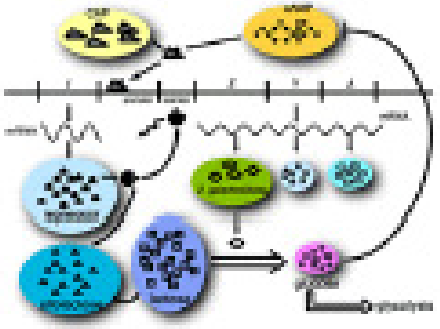


Molecular Interaction Networks

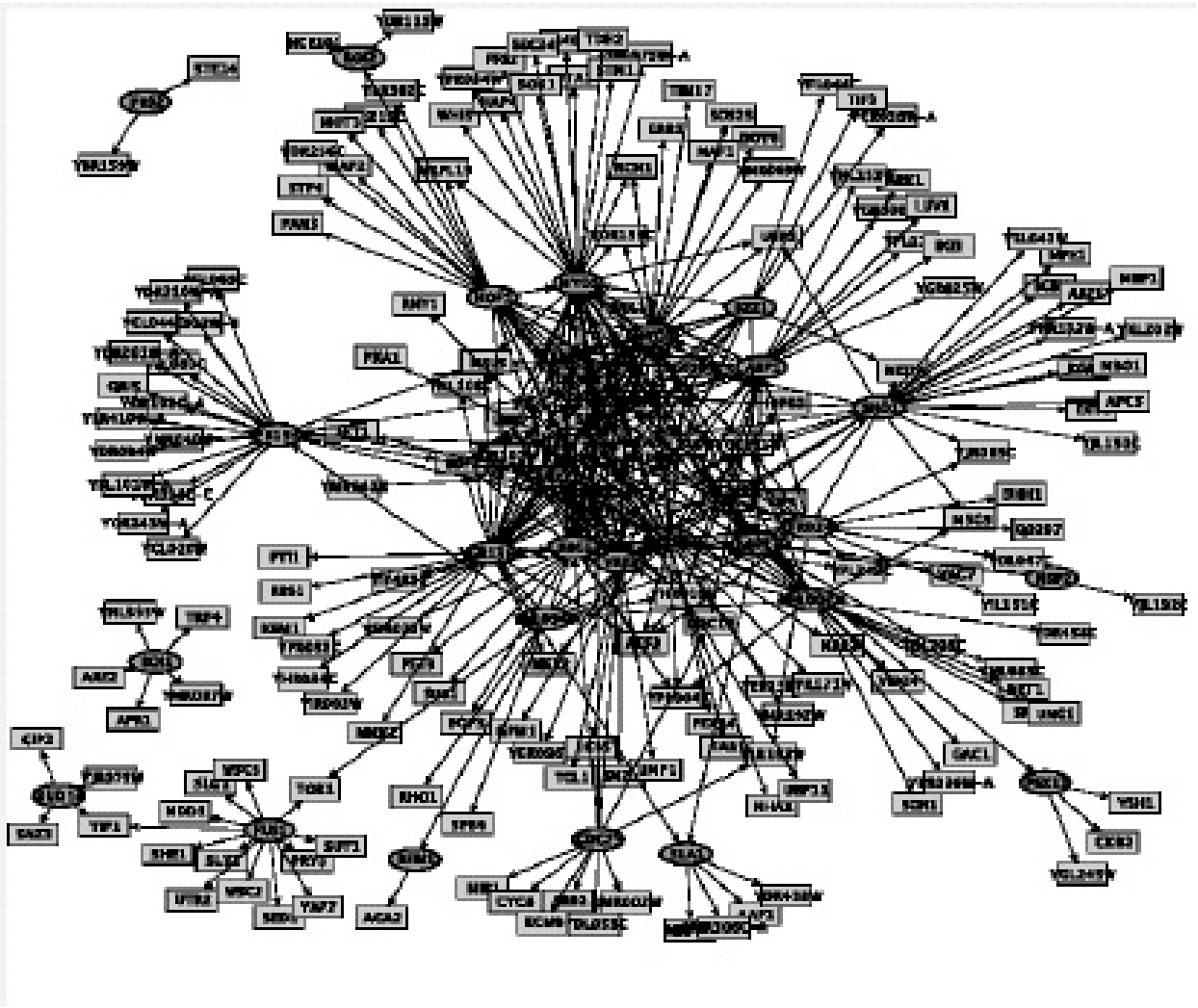
Apoptosis

Signaling Pathways

Lac Operon



Gene Regulation



SH3 domain protein interaction network in *S. cerevisiae*; from Tong et al. (2002)

Experimental high-throughput techniques

Yeast two-hybrid

Phage display

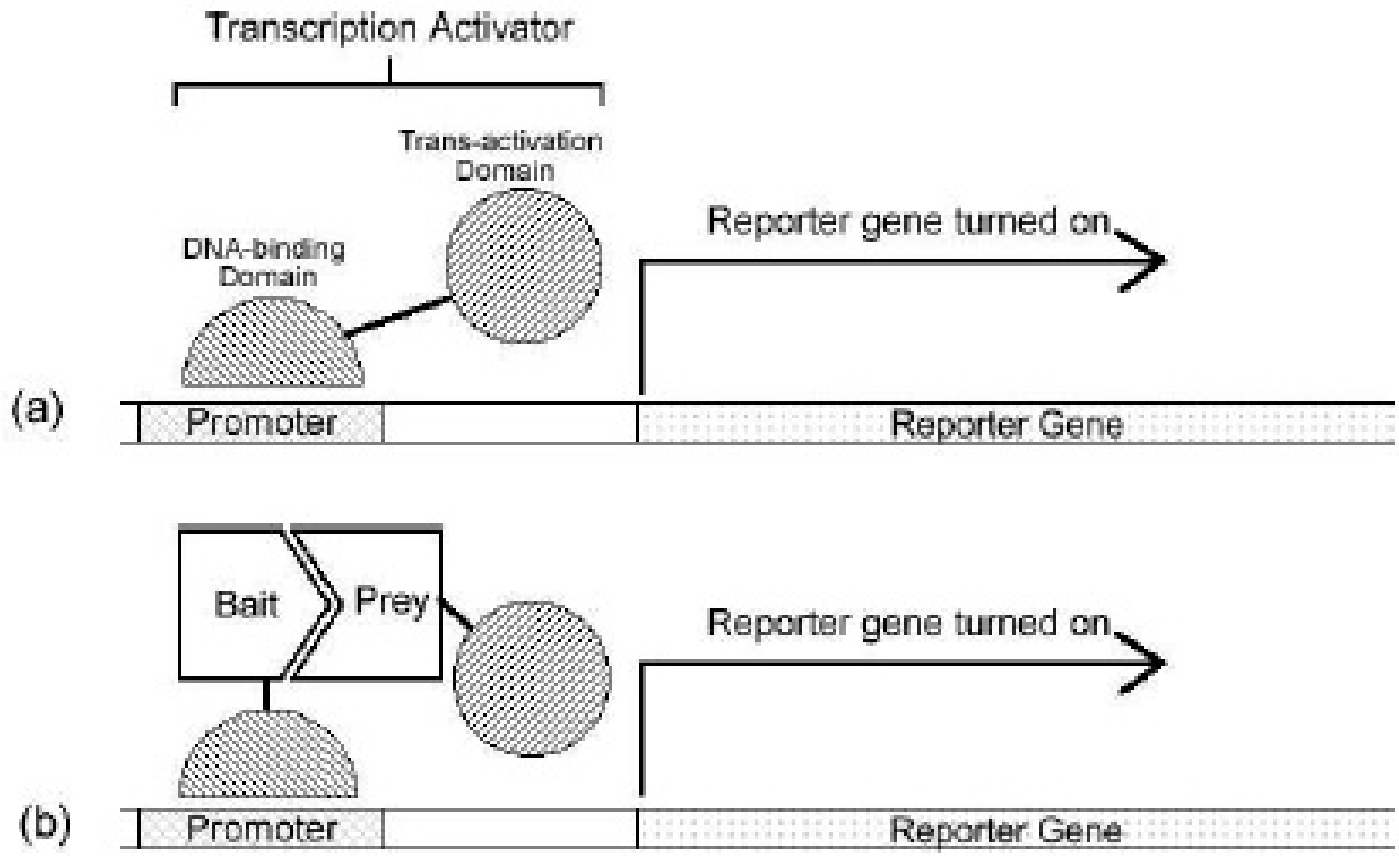


Fig. 1. Interaction detection by yeast two-hybrid assay. (a) Activation of reporter gene by transcriptional activator; (b) Activation of reporter gene by reconstituted transcriptional activator.

From See-Kiong Ng and Soon-Heng Tan, J. Bioinf. Comp. Bio. (2004)

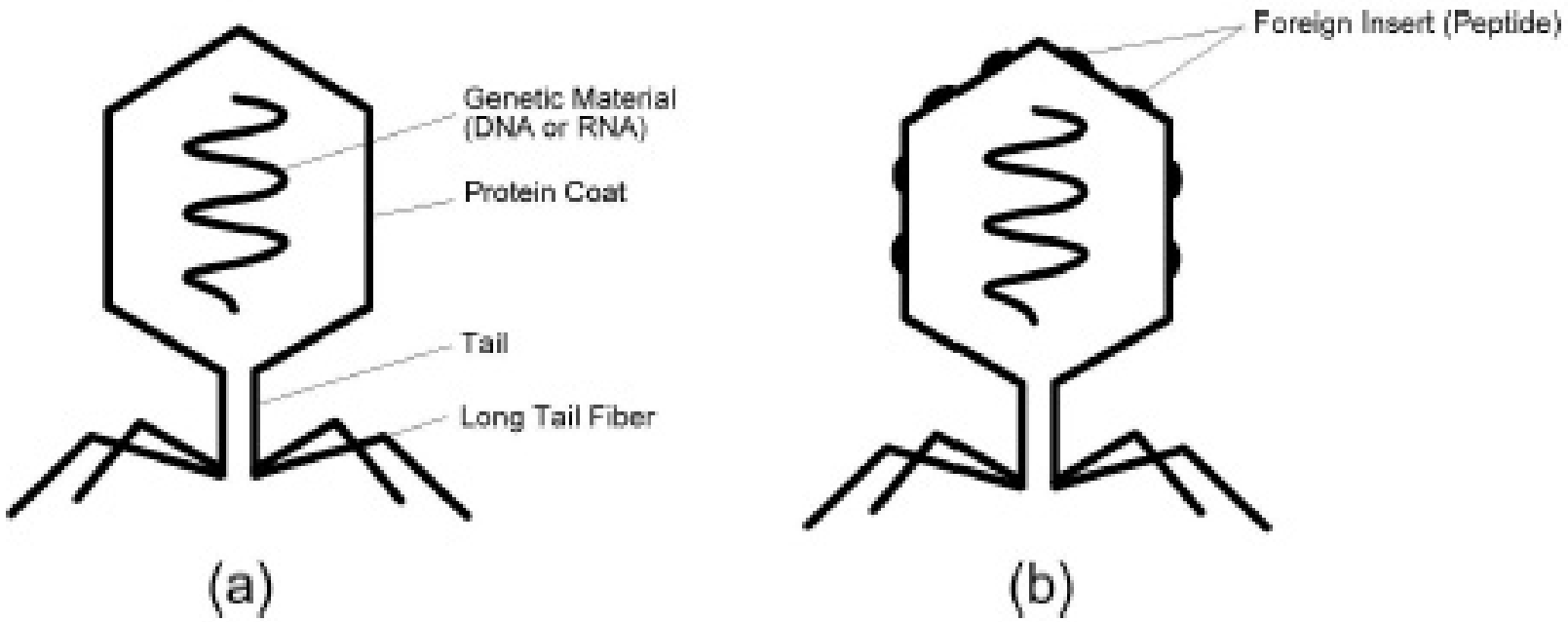


Fig. 2. Schematic diagrams of (a) a phage; and (b) interaction detection by phage display.

From See-Kiong Ng and Soon-Heng Tan, J. Bioinf. Comp. Bio. (2004)

Tong et al. (2002), Science 295, 321-324.

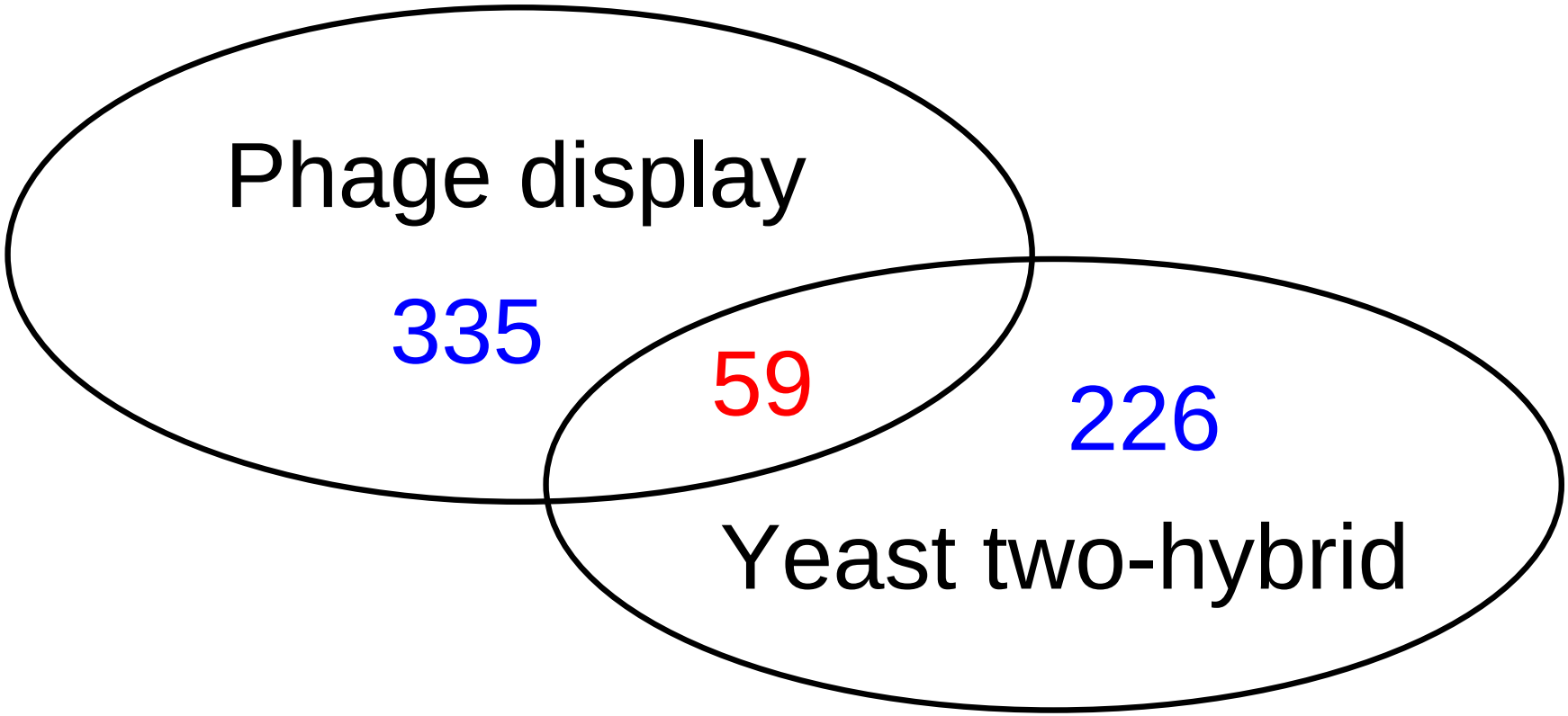
SH3 domain proteins in *Saccharomyces cerevisiae*.

Yeast two-hybrid interaction network

285 interactions between 28 SH3 proteins
and 143 binding peptides

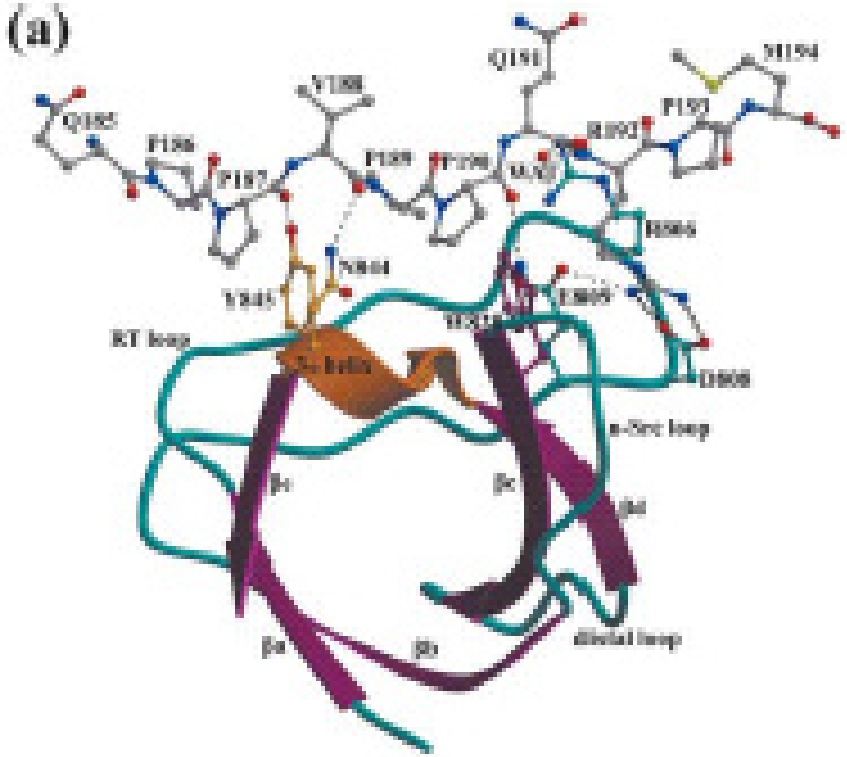
Phage display interaction network

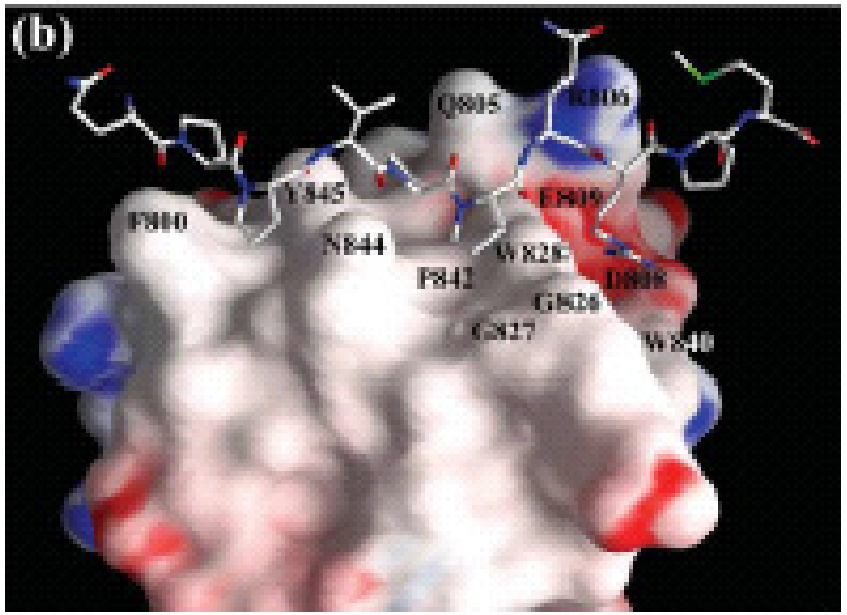
394 interactions between 28 SH3 proteins
and 178 binding peptides


Phage display
335
59
226
Yeast two-hybrid

- High-throughput experiments (yeast two-hybrid, phage display) are expensive and intrinsically noisy.
- It would be desirable to more specifically target or partially bypass them with complementary *in silico* approaches.

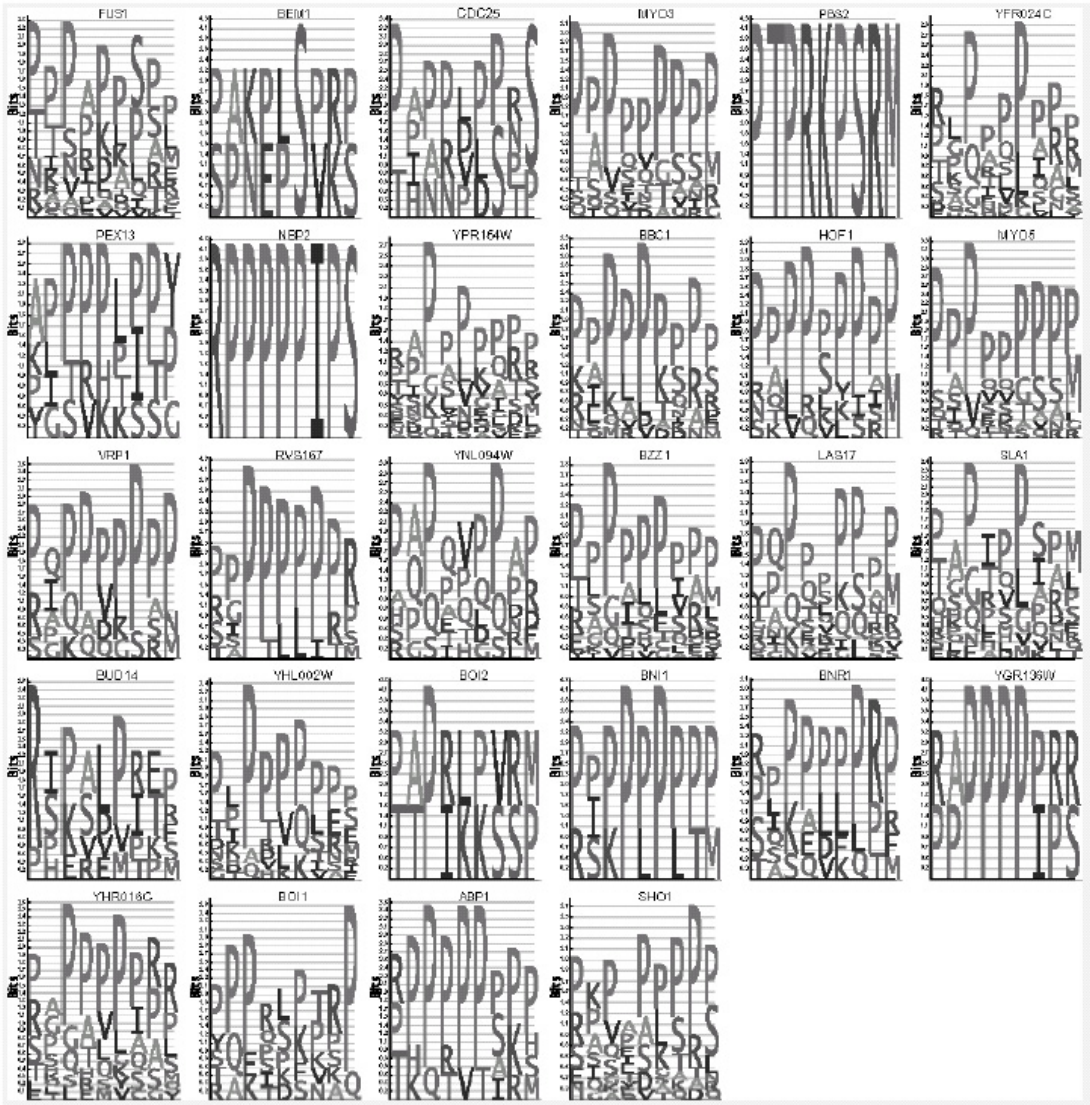
- High-throughput experiments (yeast two-hybrid, phage display) are expensive and intrinsically noisy.
- It would be desirable to more specifically target or partially bypass them with complementary *in silico* approaches.
- Objective: develop a probabilistic model to predict protein-protein interactions from sequence data.
- Method: We want to capture the way protein recognition modules recognise and bind to peptide ligands that contain a specific binding motif.

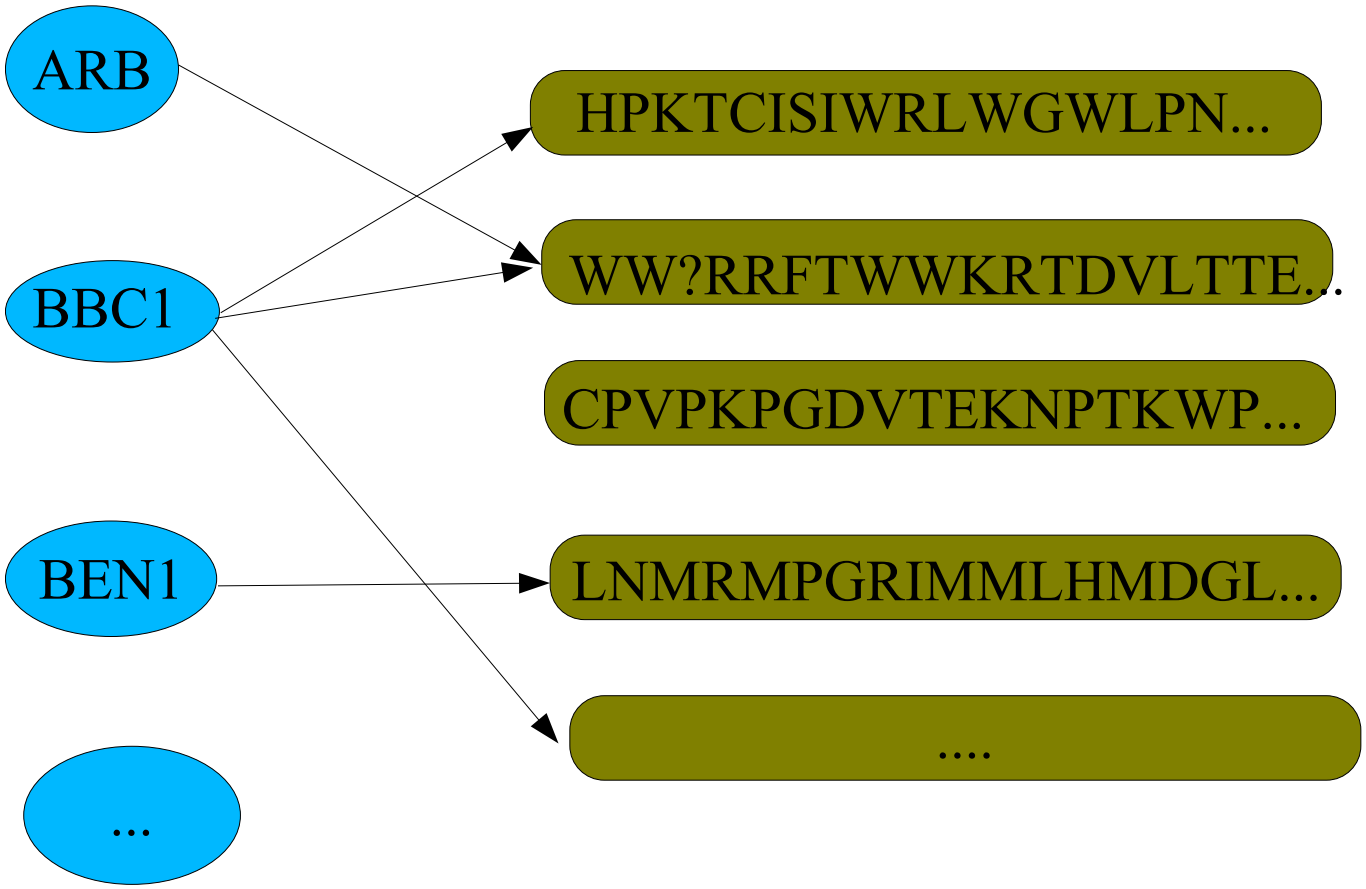




Peptide recognition modules

Example: SH3 domain


FUS1
BEM1
CDC25
MYO3
PBS2
YFR024C
PEX13
NBP2
YPR164W
BBC1
HOF1
MYO5
VRP1
RVS167
YNL094W
BZZ1
LAS17
SLA1
BUD14
YHL002W
BOI2
BNI1
BNR1
YGR136W
YHR018C
BOI1
ABP1
SHO1


ARB
BBC1
BEN1
...
HPKTCISIWRLWGWLPN...
WW?RRFTWWKRTDVLTTE...
CPVPKPGDVTEKNPTKWP...
LNMRMPGRIMMLHMDGL...
....
SH3 domains
Interacting proteins

Motif: TA[C/T] A[C/G]

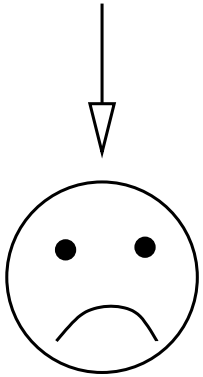
T C G A A T T C T A T A G C C A C

Motif: T A [C/T] A [C/G]

T C G A A T T C T A T A G C C A C

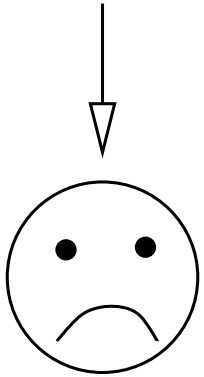
Motif: T A [C/T] A [C/G]

T C G A A T T C T A T A G C C A C



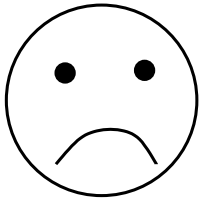
Motif: TA[C/T] A[C/G]

T C G A A T T C T A T A G C C A C



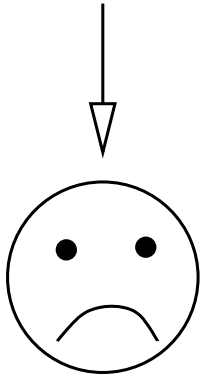
Motif: TA[C/T] A[C/G]

T C G A A T T C T A T A G C C A C



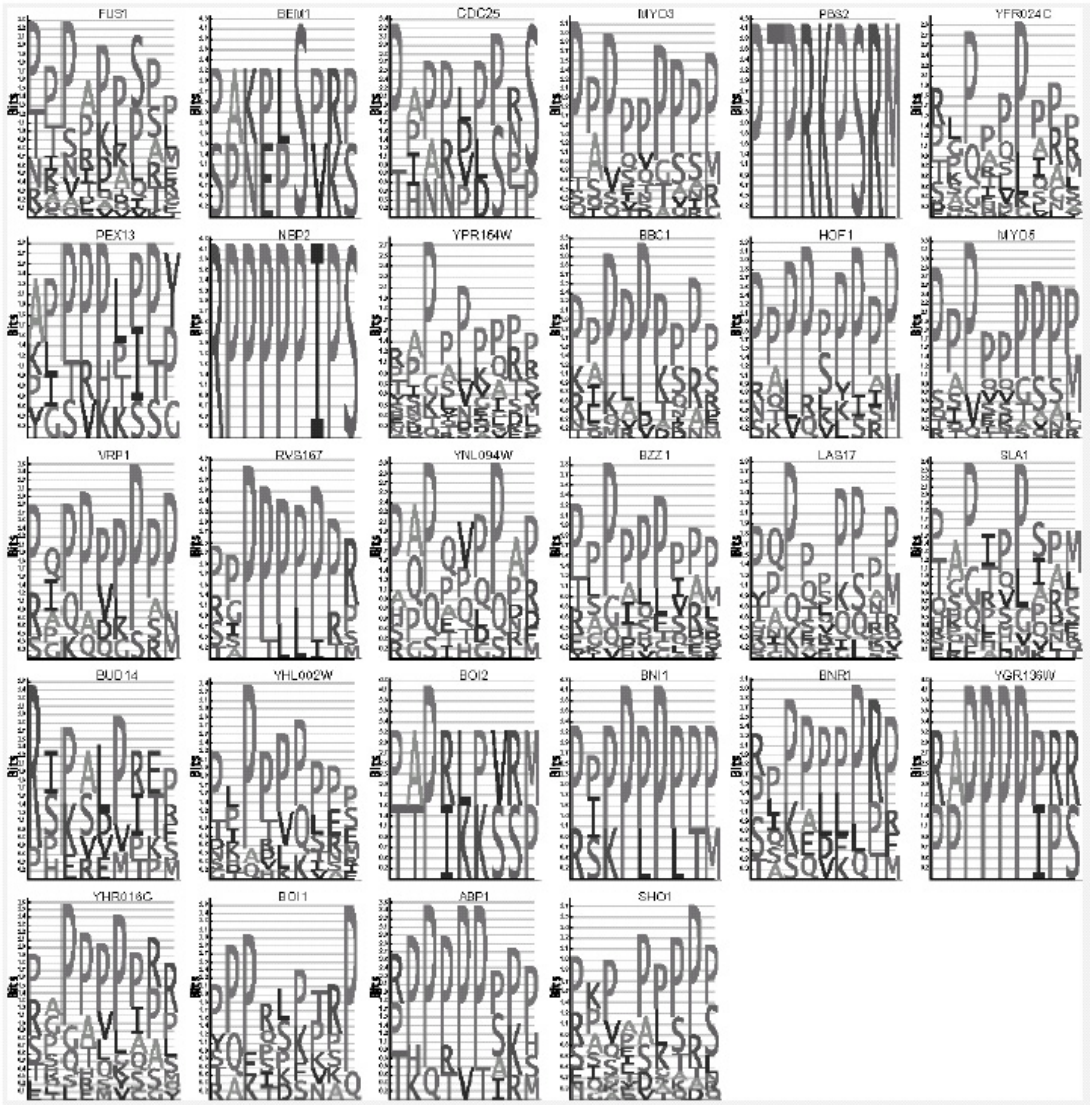
Motif: T A [C/T] A [C/G]

T C G A A T T C T A T A G C C A C



Motif: TA[C/T] A[C/G]

T C G A A T T C T A T A G C C A C


FUS1
BEM1
CDC25
MYO3
PBS2
YFR024C
PEX13
NBP2
YPR164W
BBC1
HOF1
MYO5
VRP1
RVS167
YNL094W
BZZ1
LAS17
SLA1
BUD14
YHL002W
BOI2
BNI1
BNR1
YGR136W
YHR018C
BOI1
ABP1
SHO1

Position Specific Scoring Matrix (PSSM)

Search for a motif of length W in binding sequences.

Position Specific Scoring Matrix (PSSM)

Search for a motif of length W in binding sequences.

$W \times 4$ matrix $\psi_k(l)$:
Probability that the nucleotide in the kth position,
$k \in [1, \ldots, W]$, is an $l \in \{A, C, G, T\}$.

Position Specific Scoring Matrix (PSSM)

Search for a motif of length W in binding sequences.

$W \times 4$ matrix $\psi_k(l)$:
Probability that the nucleotide in the kth position,
$k \in [1, \ldots, W]$, is an $l \in \{A, C, G, T\}$.

Background model

for non-binding sequences

4-dim vector $\theta_0(l)$:
Probability of nucleotide l; this distribution is position-independent.

Sequence $S_1, S_2, \ldots, S_N$

Sequence $S_1, S_2, \ldots, S_N$

Non-binding sequence: R=0

$$P(S_1, S_2, \ldots, S_N | R = 0) \; = \; \prod_{t=1}^{N} \theta_0(S_t)$$

Sequence $S_1, S_2, \ldots, S_N$

Non-binding sequence: R=0

$$P(S_1, S_2, \ldots, S_N | R = 0) = \prod_{t=1}^{N} \theta_0(S_t)$$

Binding sequence: R=1, motif starting at position m+1

$$\begin{aligned} & P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) \\ = & \prod_{t=1}^{m} \theta_0(S_t) \prod_{k=1}^{W} \psi_k(S_{m+k}) \prod_{t=m+W+1}^{N} \theta_0(S_t) \\ = & \prod_{t=1}^{N} \theta_0(S_t) \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned}$$

Binding sequence: R=1, motif starting at position m+1

$$P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) = \prod_{t=1}^{N} \theta_0(S_t) \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$

Binding sequence: R=1, motif starting at position m+1

$$P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) = \prod_{t=1}^{N} \theta_0(S_t) \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$

Binding sequence: R=1, motif starting anywhere

$$\begin{aligned} & P(S_1, S_2, \ldots, S_N | R = 1) \\ &= \sum_{m=0}^{N-W} P(start = m + 1) P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) \\ &= \prod_{t=1}^{N} \theta_0(S_t) \frac{1}{N - W + 1} \sum_{m=0}^{N-W} \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned}$$

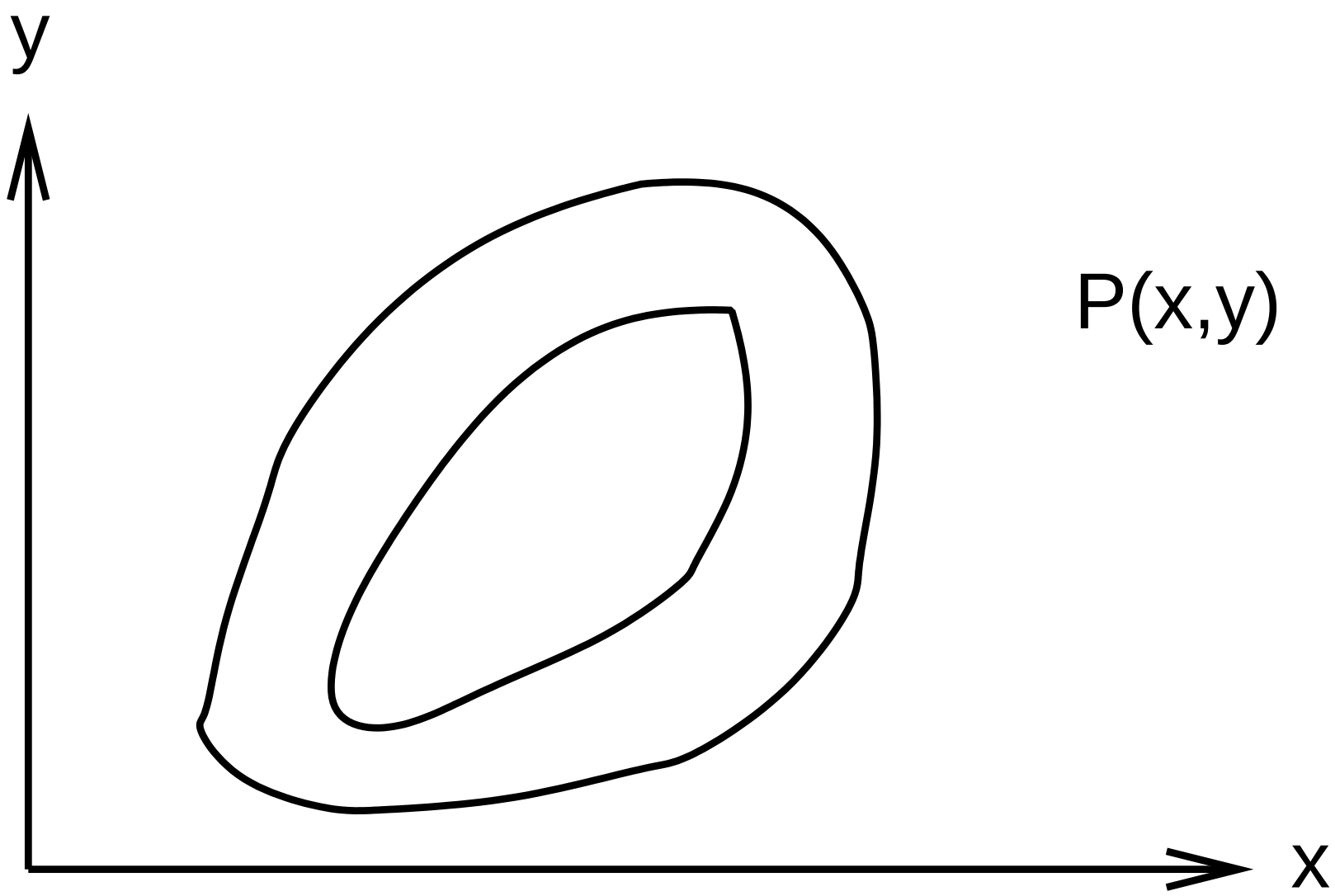
Binding sequence: R=1, motif starting at position m+1

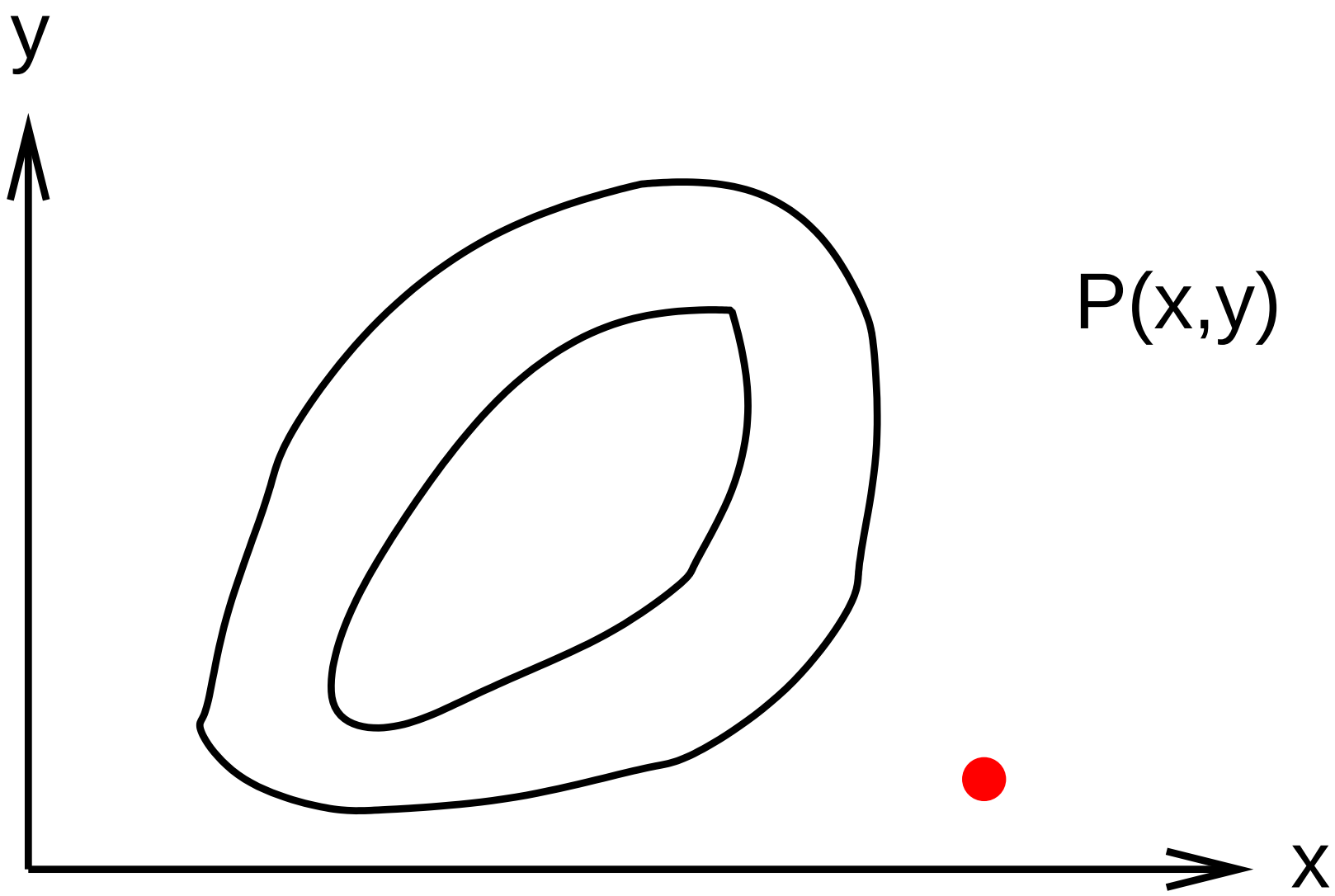
$$P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) = \prod_{t=1}^{N} \theta_0(S_t) \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$

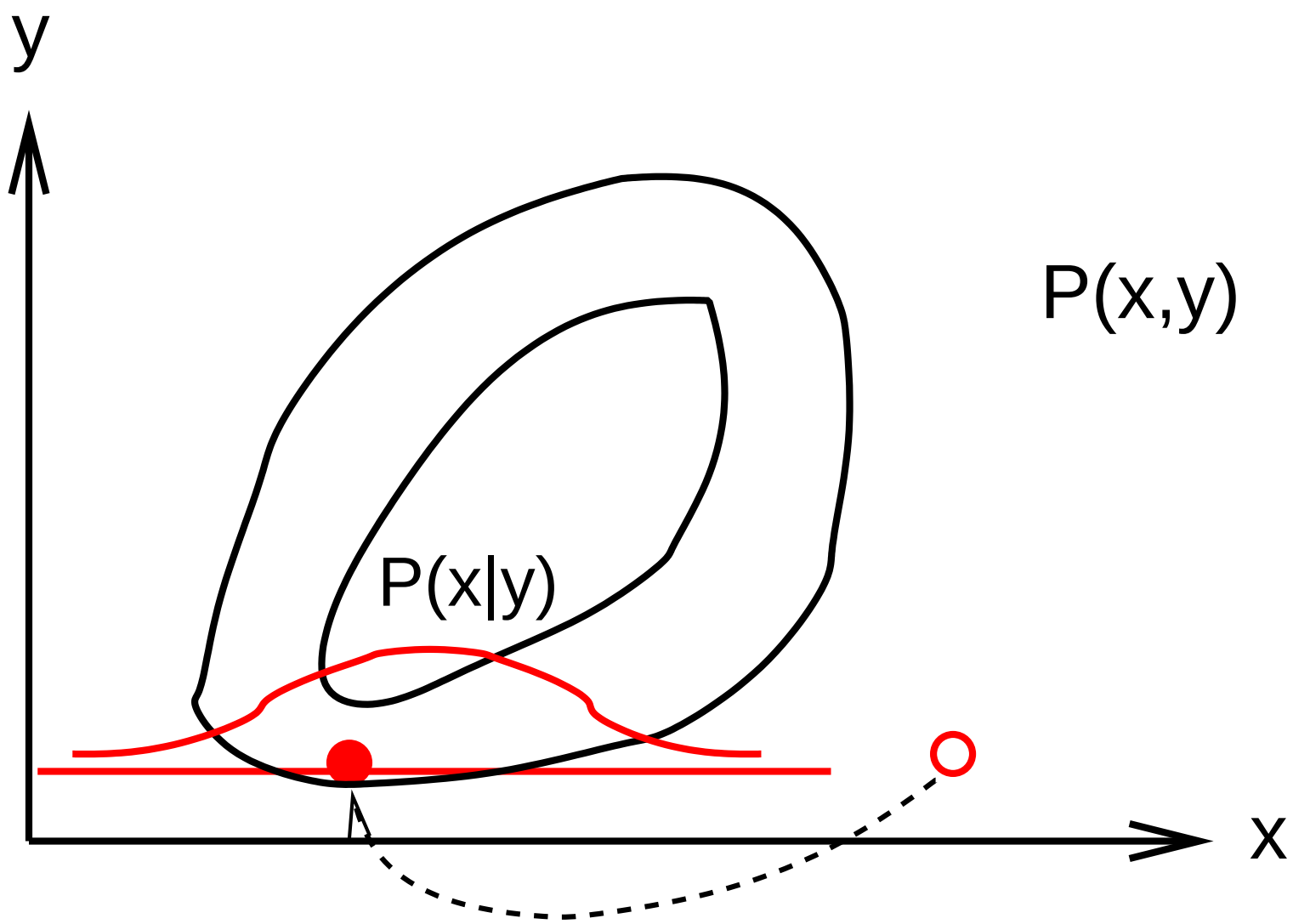
Binding sequence: R=1, motif starting anywhere

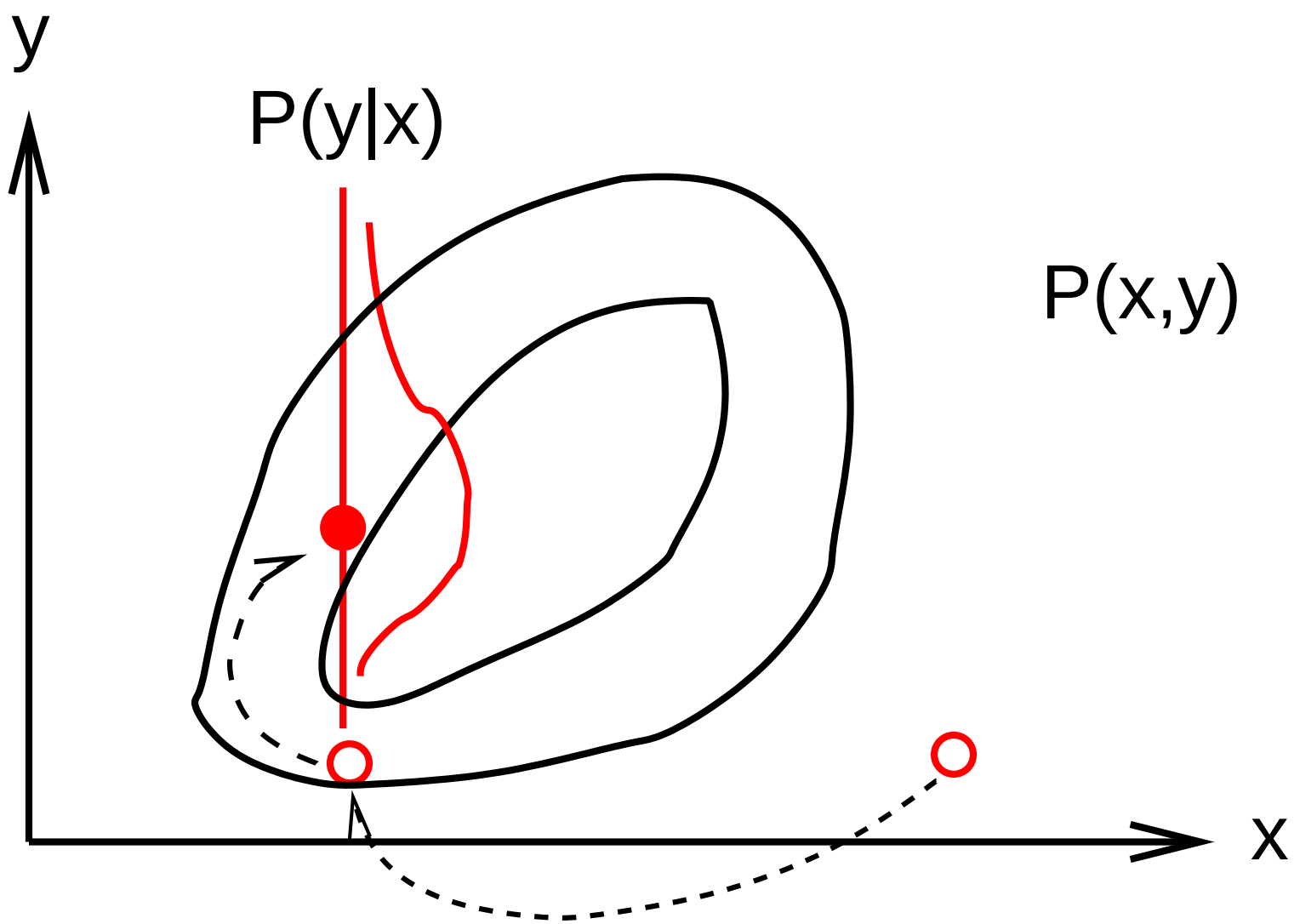
$$\begin{aligned} & P(S_1, S_2, \ldots, S_N | R = 1) \\ &= \sum_{m=0}^{N-W} P(start = m + 1) P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) \\ &= \prod_{t=1}^{N} \theta_0(S_t) \frac{1}{N - W + 1} \sum_{m=0}^{N-W} \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned}$$

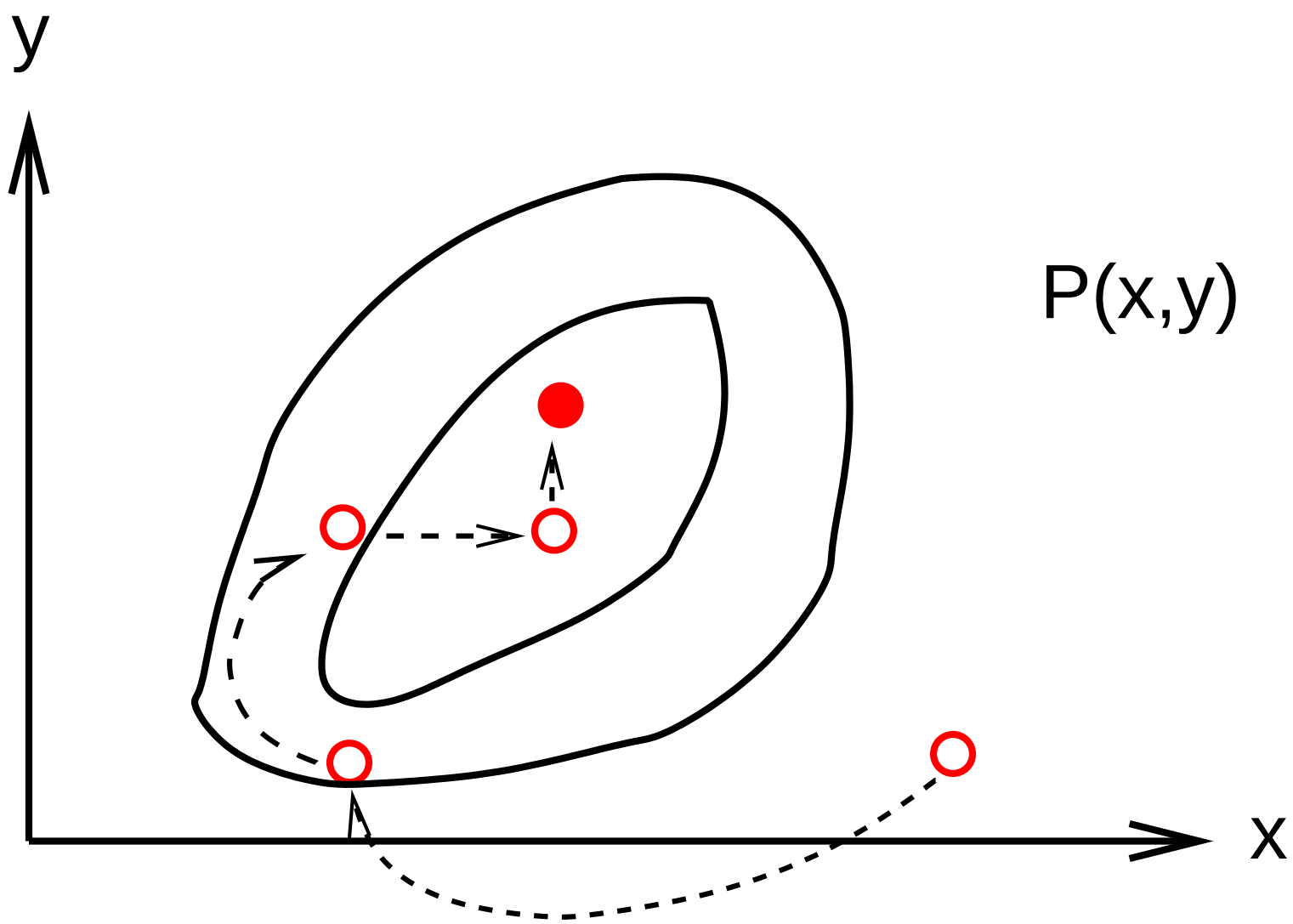
Gibbs sampling


y
P(x,y)
x


y
P(x,y)
x


y
P(x,y)
P(x|y)
x


y
P(y|x)
P(x,y)
x


y
P(x,y)
x

Posterior probability of a binding location, given the parameters

$$P(start = m+1 | S_1, S_2, \ldots, S_N, Parameters) \; =$$

$$\frac{P(S_1, S_2, \ldots, S_N | start = m+1) P(start = m+1)}{P(S_1, S_2, \ldots, S_N)}$$

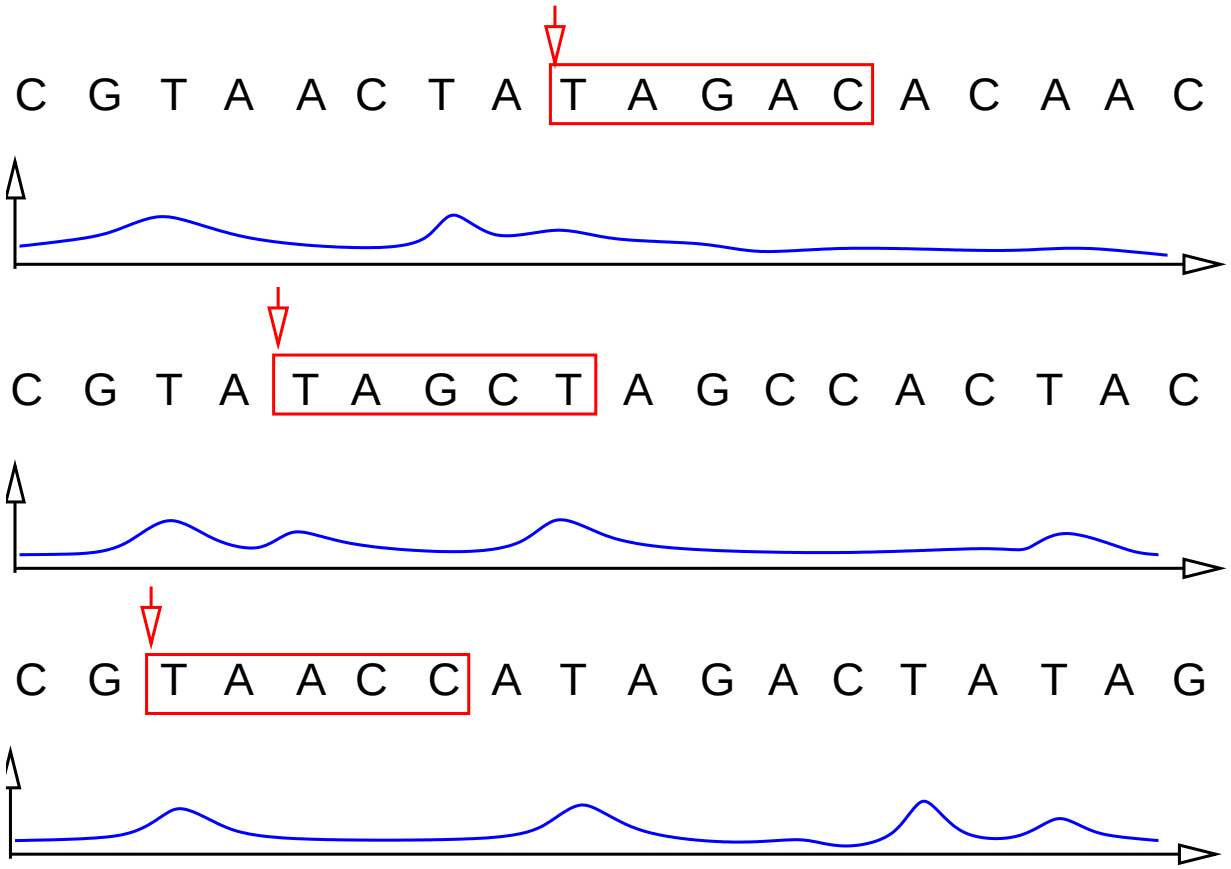
Motif starting at position m+1

$$P(S_1, S_2, \ldots, S_N | start = m+1) \; = \; \prod_{t=1}^{N} \theta_0(S_t) \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$

Motif starting anywhere

$$P(S_1, S_2, \ldots, S_N)$$

$$= \sum_{m=0}^{N-W} P(start = m+1) P(S_1, S_2, \ldots, S_N | start = m+1)$$



Counts =

	1	2	3	4	5
A	0	3	1	1	0
C	0	0	0	2	2
G	0	0	2	0	0
T	3	0	0	0	1

Posterior probability of the parameters, given the binding locations

Sufficient statistics: Count matrix $C_{k,l}$

$C_{k,l}$: Number of times amino acid l appears in position k.

$$P(D, binding\ locations|parameters) = \prod_{k=1}^{W}\prod_{l=1}^{20}\psi_{k,l}^{C_{k,l}}$$

Conjugate prior distribution: Dirichlet

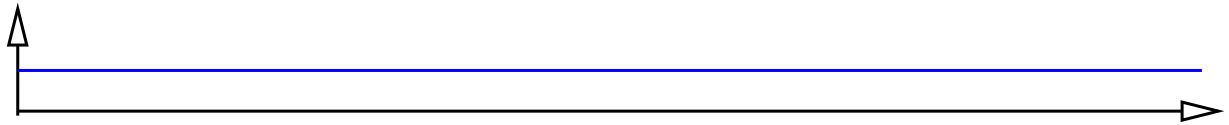
$$P(parameters) \propto \prod_{l=1}^{20}\psi_{k,l}^{\alpha_l-1}$$

Posterior distribution

$$P(parameters|D, binding\ locations) \propto \prod_{k=1}^{W}\prod_{l=1}^{20}\psi_{k,l}^{C_{k,l}+\alpha_l-1}$$

C G T A A C T A T A G A C A C A A C

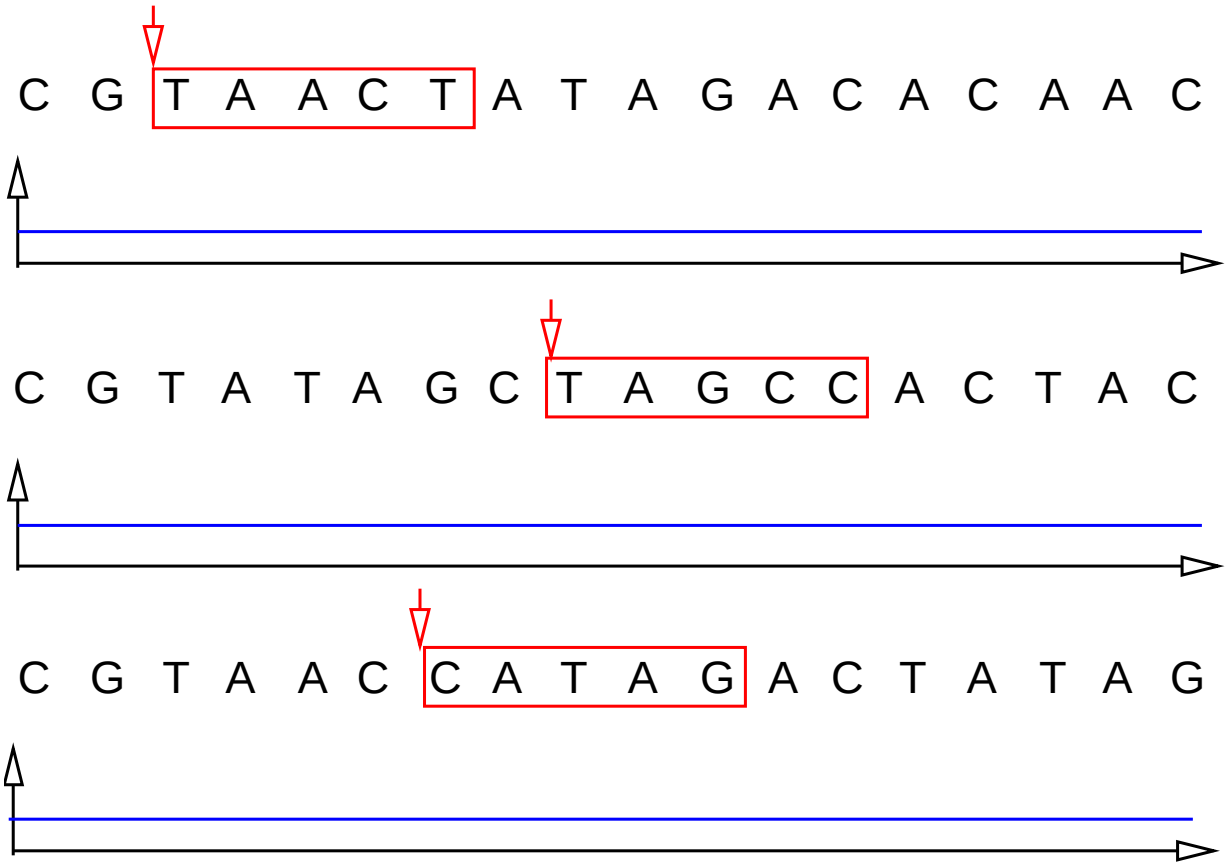
C G T A T A G C T A G C C A C T A C



C G T A A C C A T A G A C T A T A G

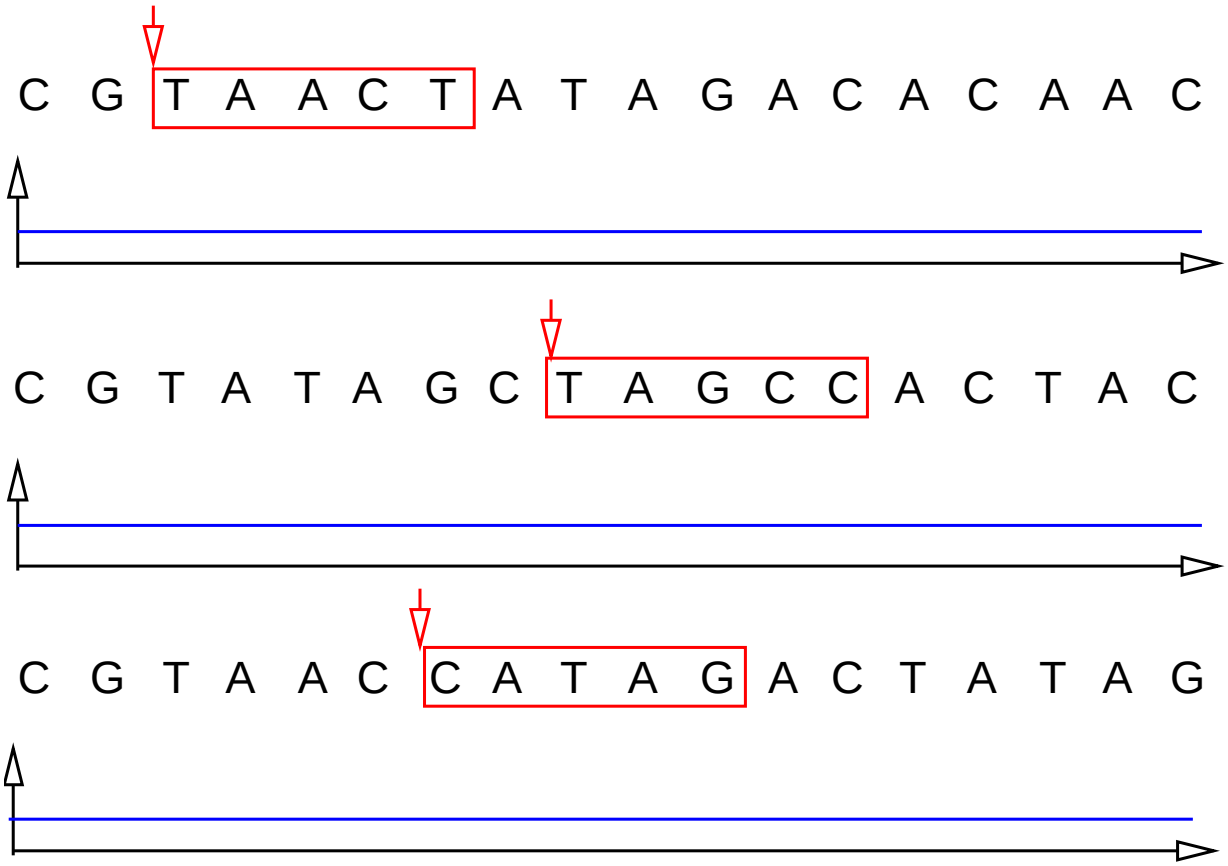
Counts =

A	0	0	0	0	0
C	0	0	0	0	0
G	0	0	0	0	0
T	0	0	0	0	0
	1	2	3	4	5



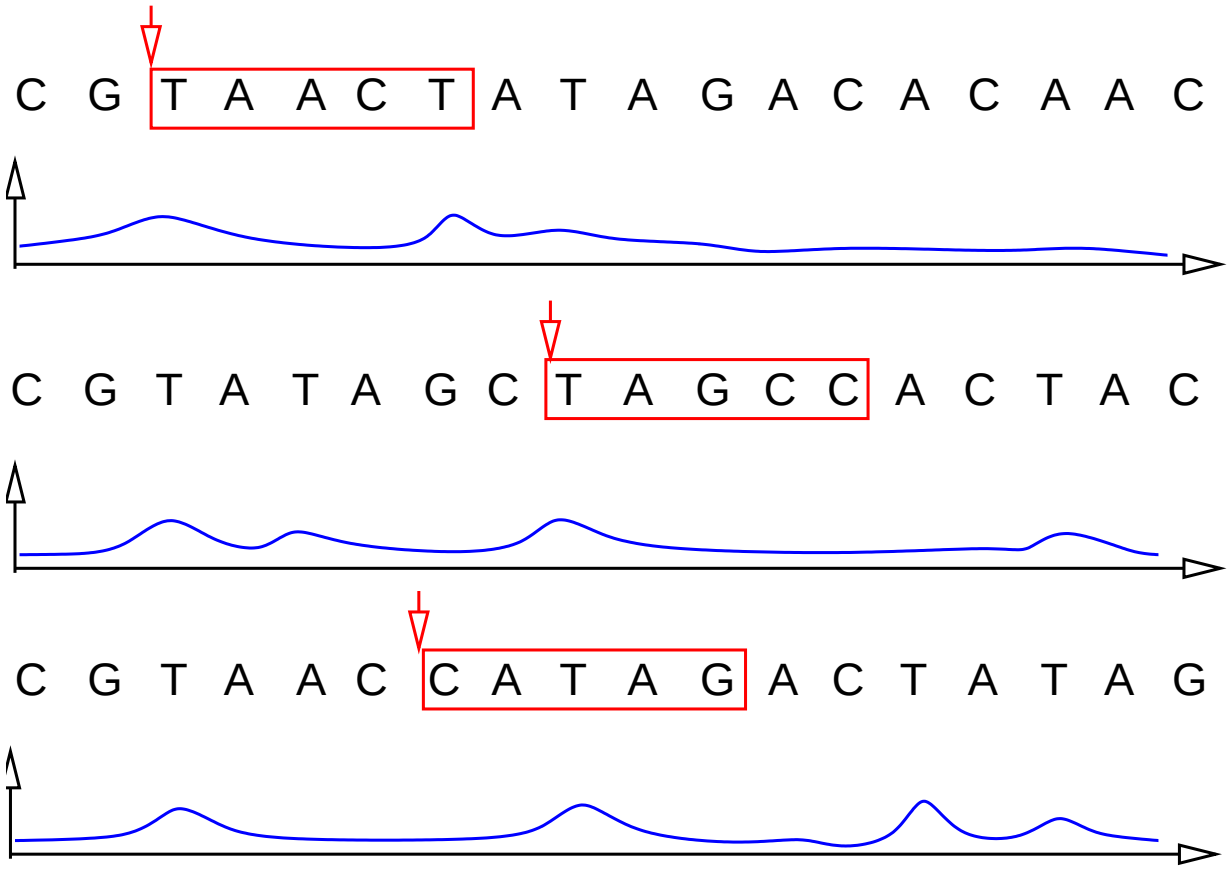
Counts =

	1	2	3	4	5
A	0	0	0	0	0
C	0	0	0	0	0
G	0	0	0	0	0
T	0	0	0	0	0



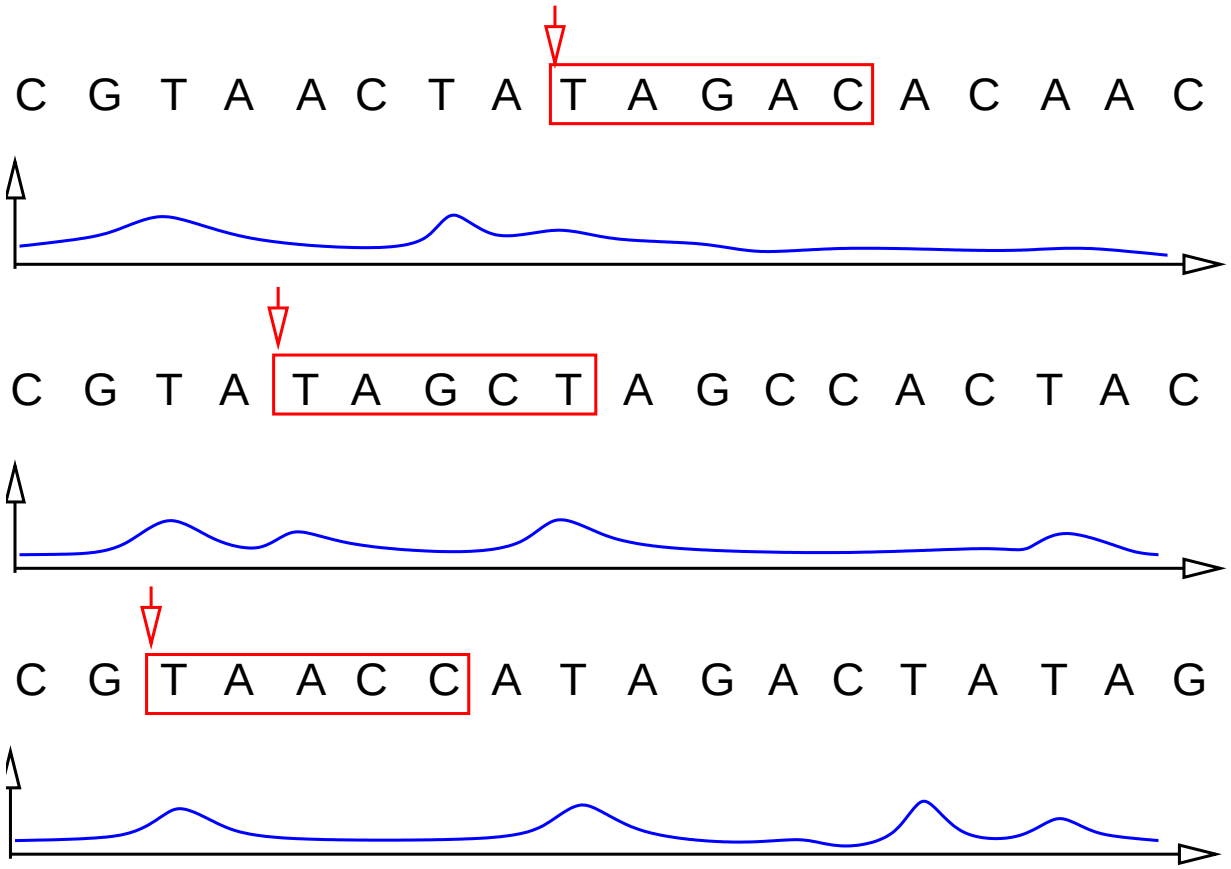
Counts =

A	0	3	1	1	0
C	1	0	0	2	1
G	0	0	1	0	1
T	2	0	1	0	1
	1	2	3	4	5



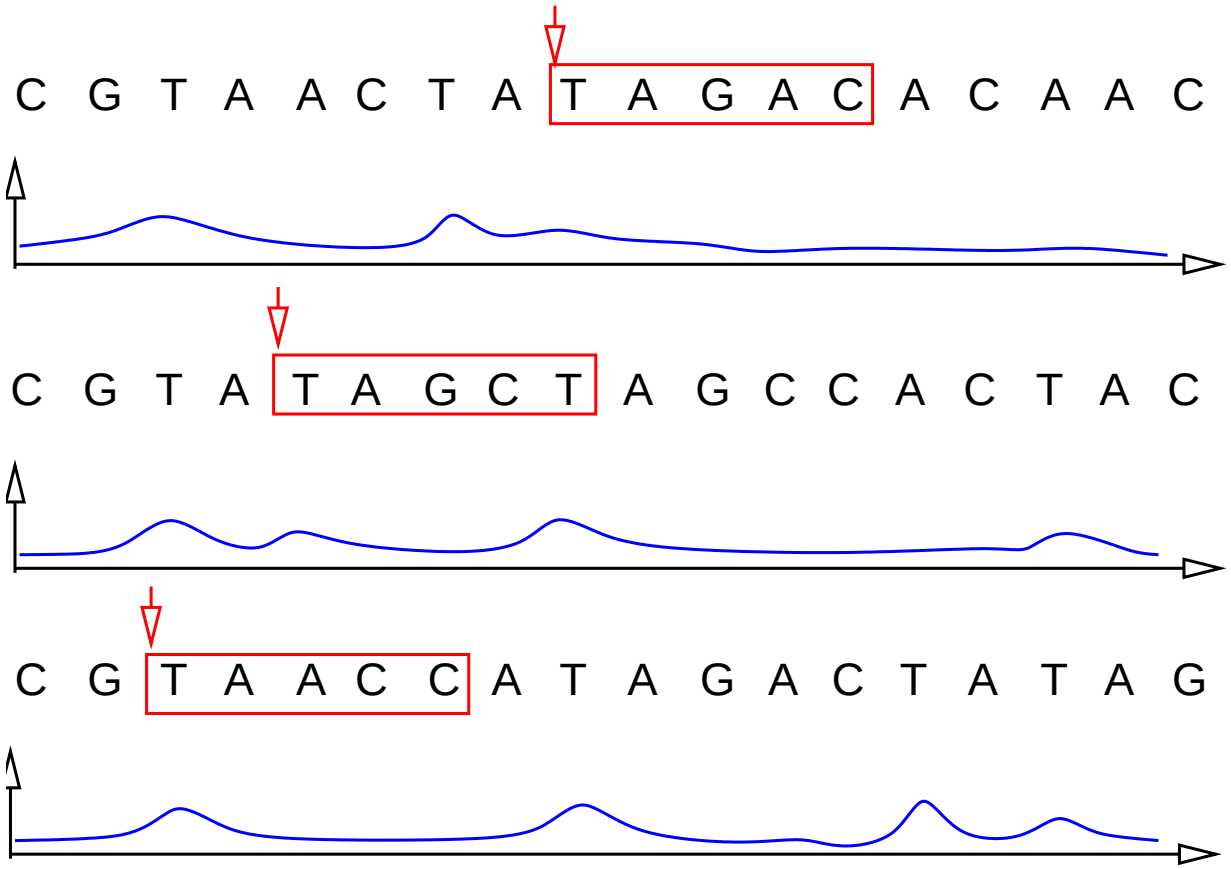
Counts =

	1	2	3	4	5
A	0	3	1	1	0
C	1	0	0	2	1
G	0	0	1	0	1
T	2	0	1	0	1



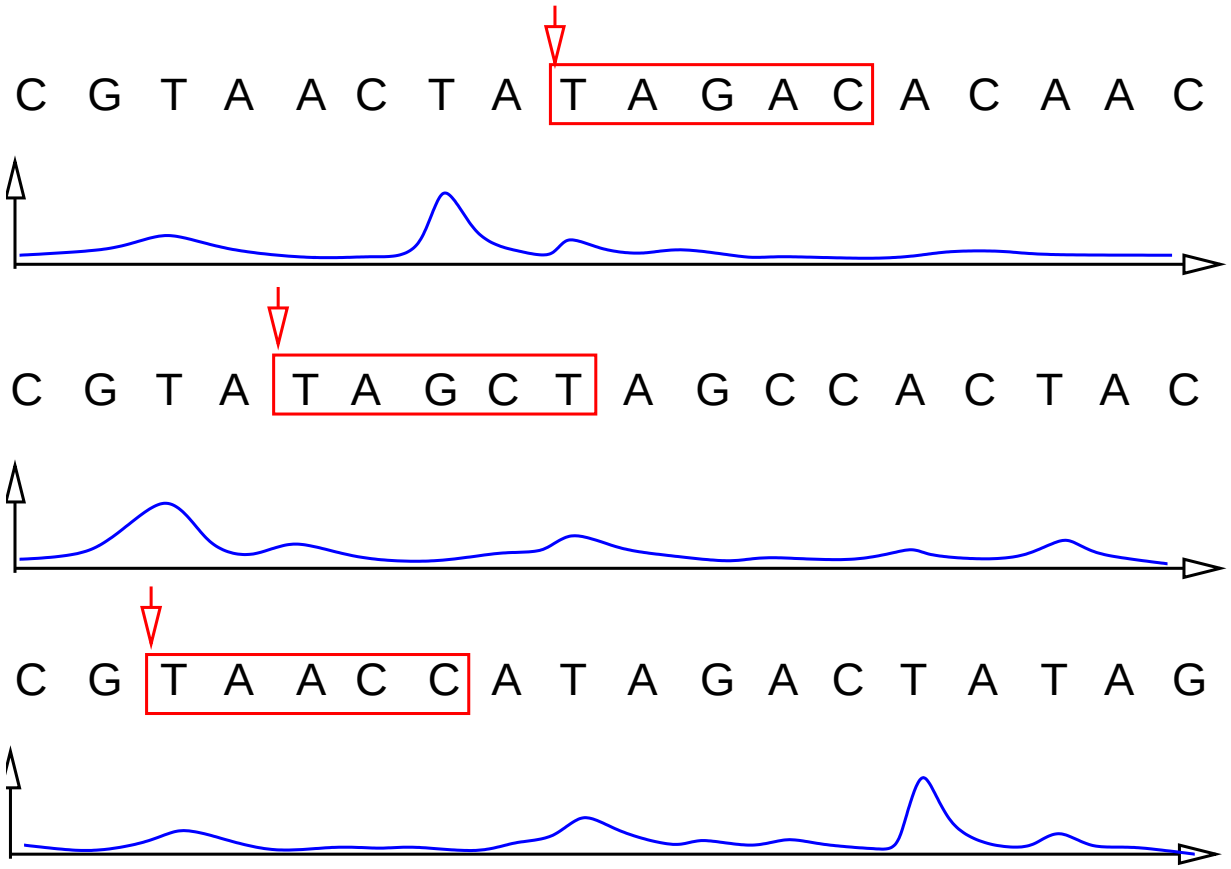
Counts =

	1	2	3	4	5
A	0	3	1	1	0
C	1	0	0	2	1
G	0	0	1	0	1
T	2	0	1	0	1



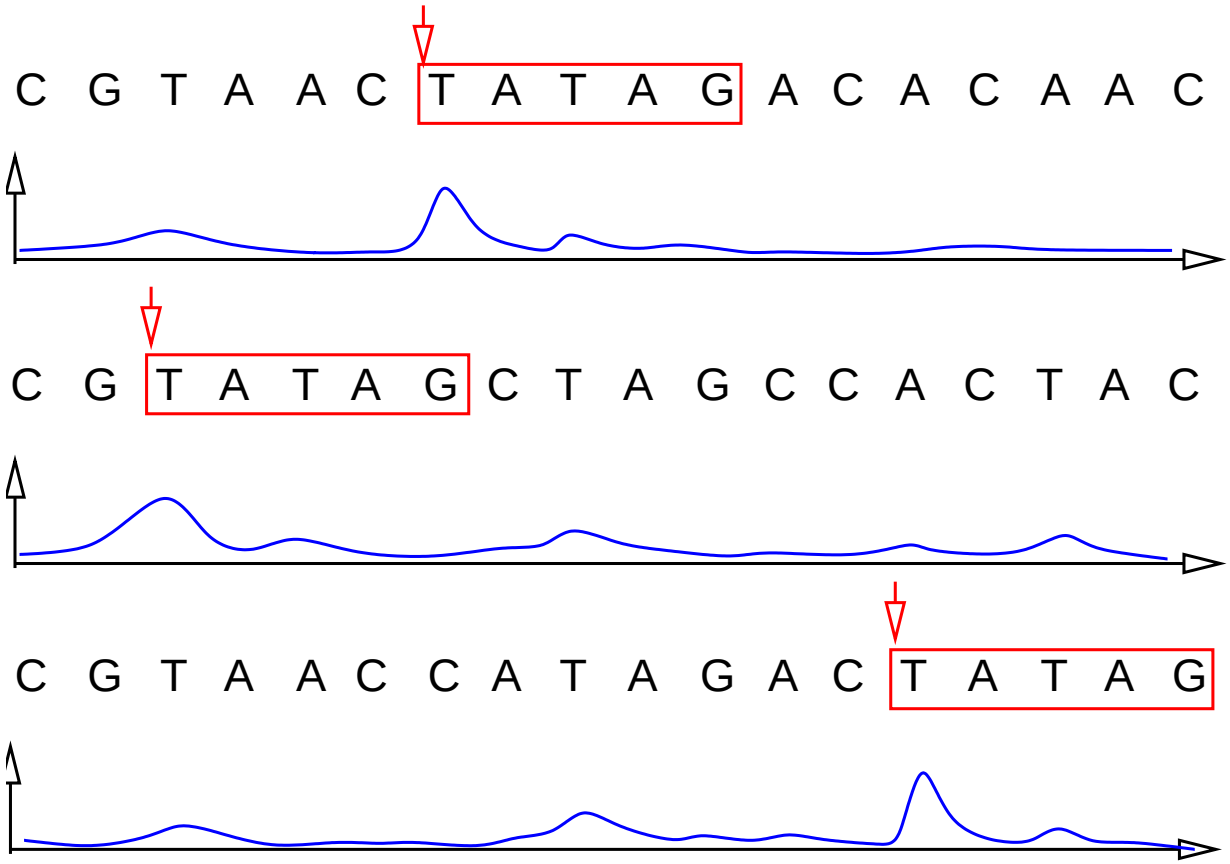
Counts =

	1	2	3	4	5
A	0	3	1	1	0
C	0	0	0	2	2
G	0	0	2	0	0
T	3	0	0	0	1



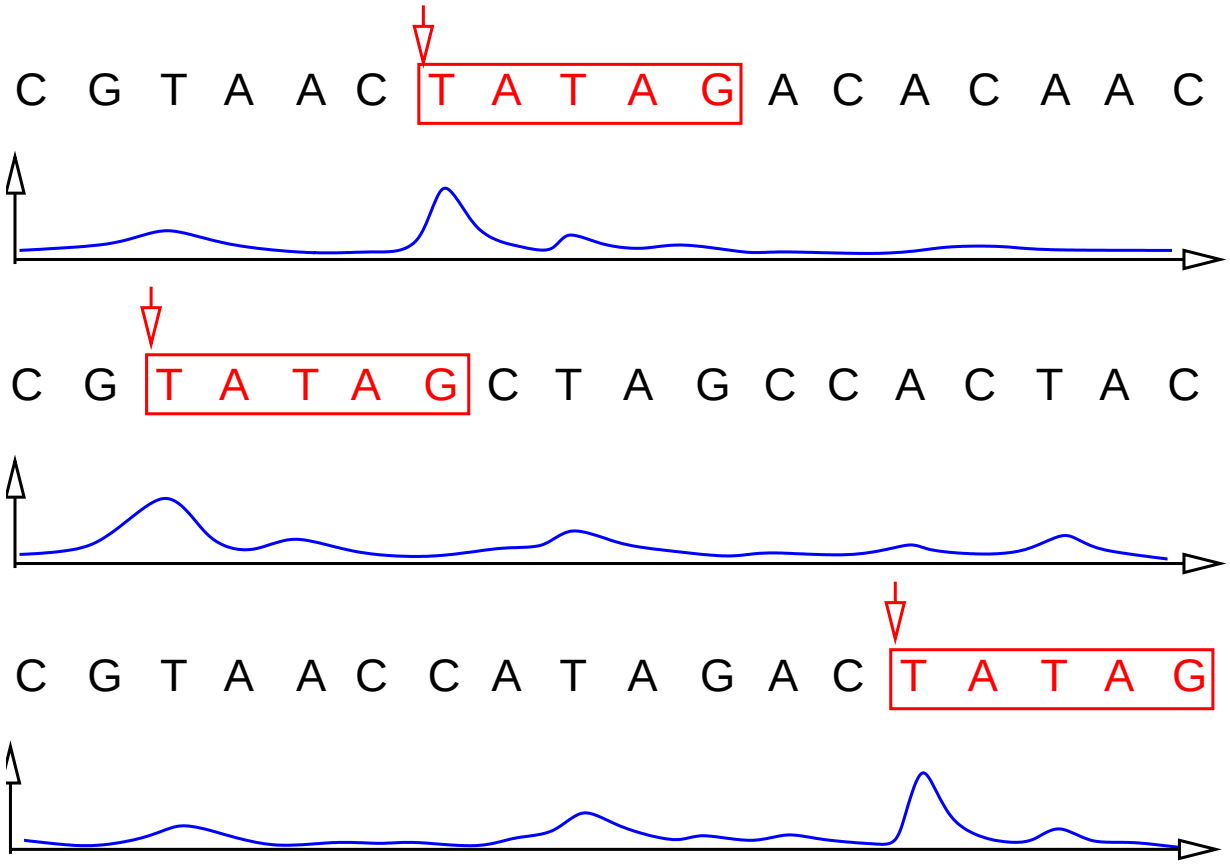
Counts =

	1	2	3	4	5
A	0	3	1	1	0
C	0	0	0	2	2
G	0	0	2	0	0
T	3	0	0	0	1



Counts =

	1	2	3	4	5
A	0	3	1	1	0
C	0	0	0	2	2
G	0	0	2	0	0
T	3	0	0	0	1



Counts =

	1	2	3	4	5
A	0	3	1	1	0
C	0	0	0	2	2
G	0	0	2	0	0
T	3	0	0	0	1

Sufficient statistics: $\mathbf{C}_{d,s}$

$$C_{d,s,k,l} = \delta(sequence_{s,a_{d,s}+k} = l)$$

Sufficient statistics: $\mathbf{C}_{d,s}$

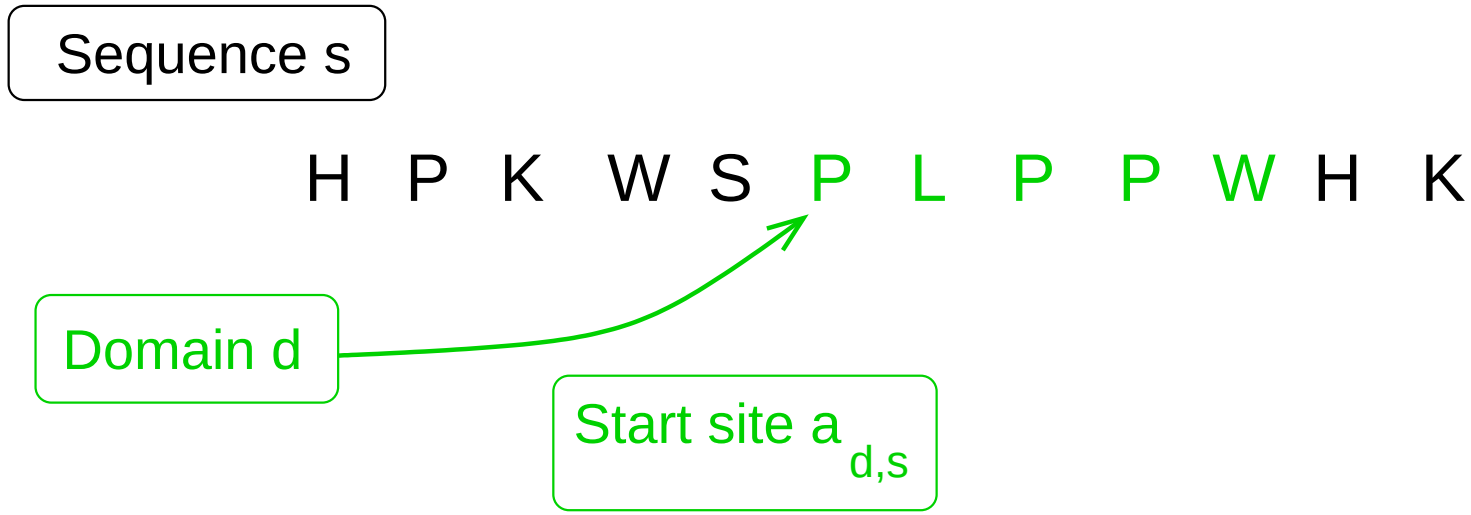
$$C_{d,s,k,l} = \delta(sequence_{s,a_{d,s}+k} = l)$$

Sequence s

H P K W S P L P P W H K

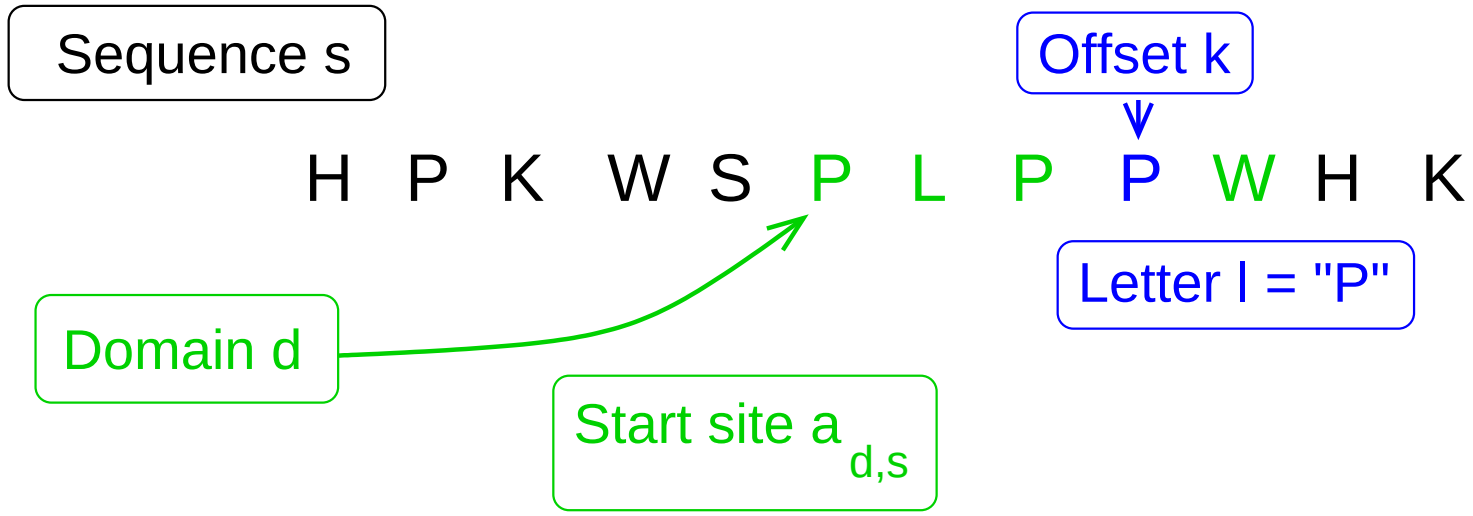
Sufficient statistics: $\mathbf{C}_{d,s}$

$$C_{d,s,k,l} = \delta(sequence_{s,a_{d,s}+k} = l)$$



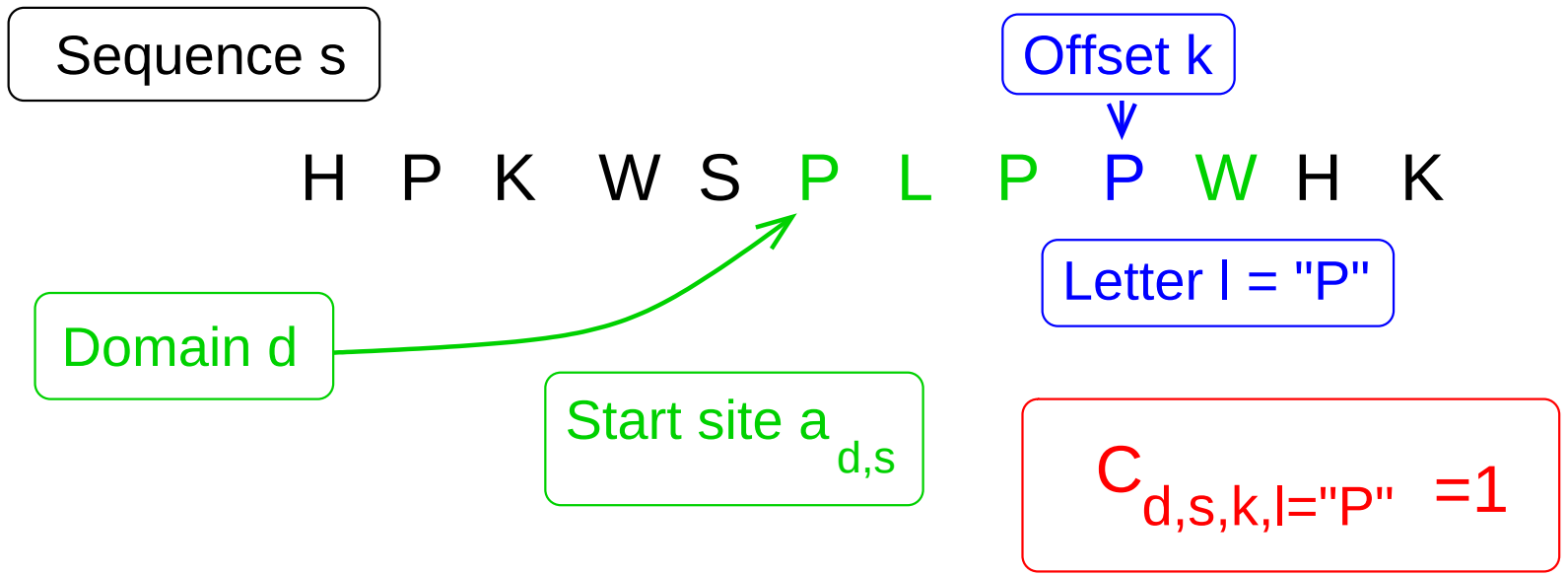
Sufficient statistics: $\mathbf{C}_{d,s}$

$$C_{d,s,k,l} = \delta(sequence_{s,a_{d,s}+k} = l)$$



Sufficient statistics: $\mathbf{C}_{d,s}$

$$C_{d,s,k,l} = \delta(sequence_{s,a_{d,s}+k} = l)$$



Sufficient statistics: $\mathbf{C}_{d,s}$

$$C_{d,s,k,l} = \delta(sequence_{s,a_{d,s}+k} = l)$$

In words: $C_{d,s,k,l}$ is 1 if the kth position of the binding motif in sequence s that binds to PRM domain d is amino acid l. Otherwise, it is zero.

Sufficient statistics: $\mathbf{C}_{d,s}$

$$C_{d,s,k,l} = \delta(sequence_{s,a_{d,s}+k} = l)$$

In words: $C_{d,s,k,l}$ is 1 if the kth position of the binding motif in sequence s that binds to PRM domain d is amino acid l. Otherwise, it is zero.

Problem:

There are too few binding peptide sequences (average: 9 sequences per domain) $\Rightarrow$ high estimation uncertainty.

BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i274–i282
DOI: 10.1093/bioinformatics/bth922

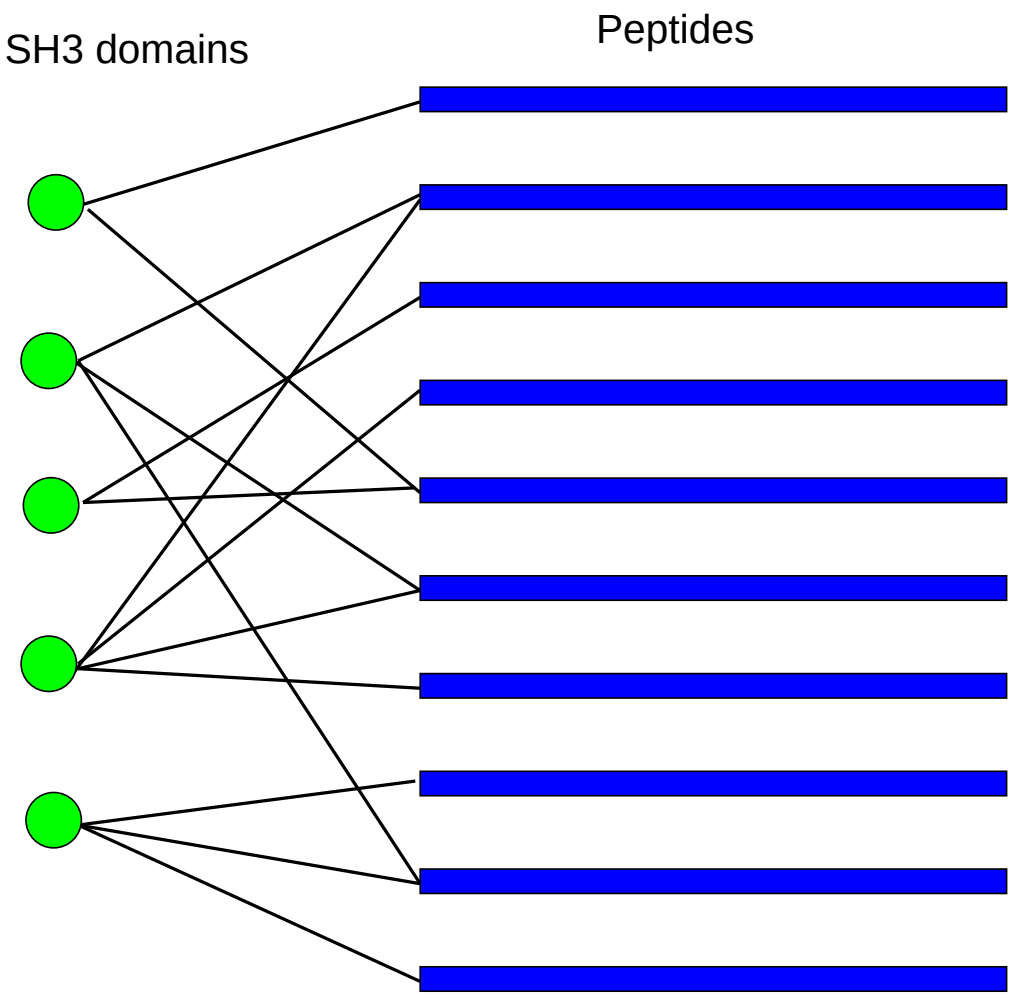
Predicting protein–peptide interactions via a network-based motif sampler

David J. Reiss* and Benno Schwikowski

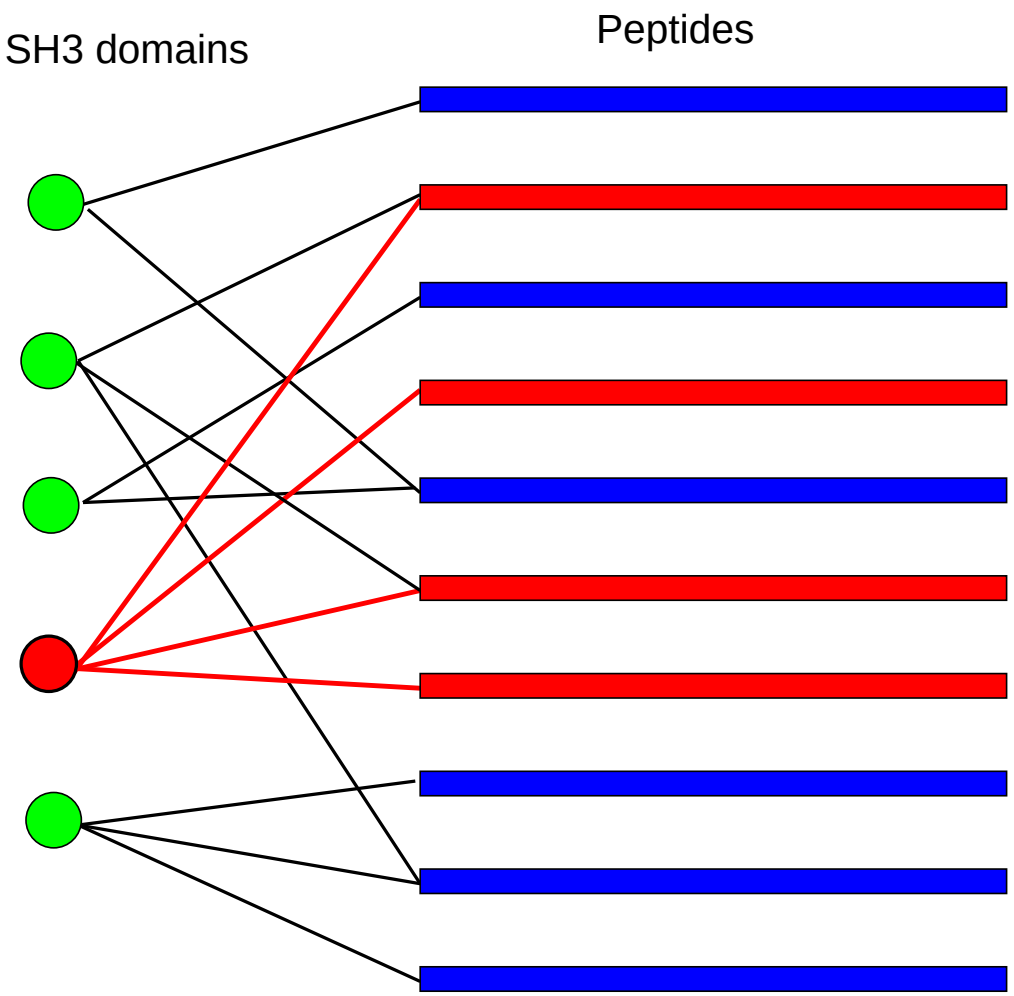
Institute for Systems Biology, 1441 North 34th street, Seattle, WA 98103-8904, USA

Received on January 15, 2004; accepted on March 1, 2004

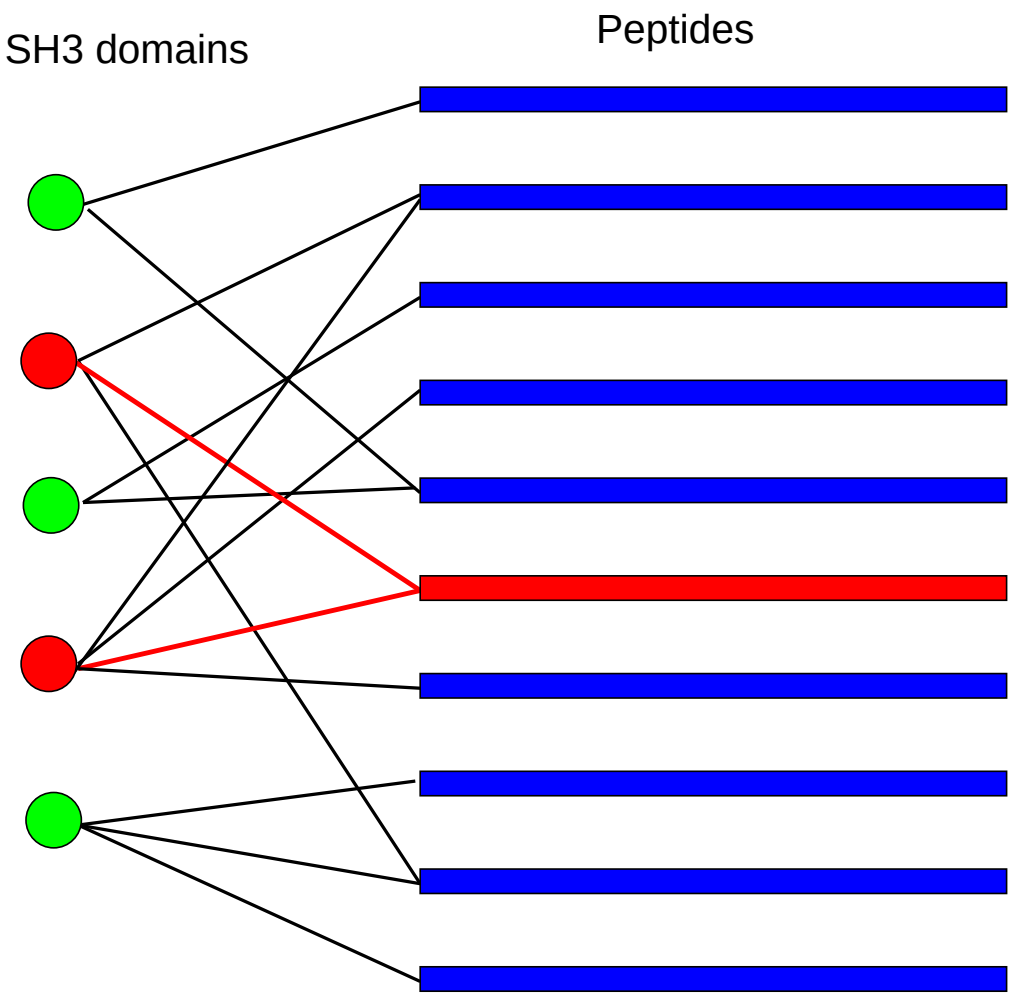
Modify the count matrix $\mathbf{C}_{d,s}$, using the network topology.



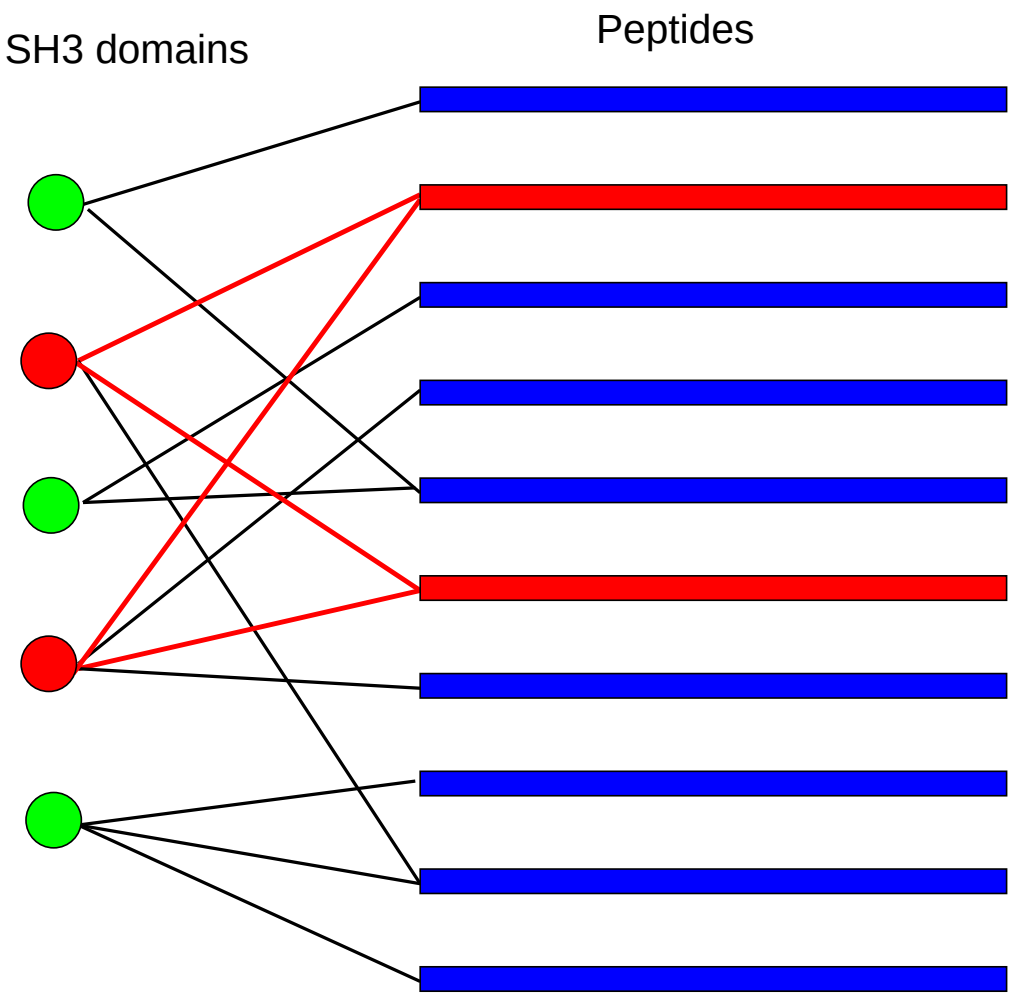
$$\tilde{\mathbf{C}}_{d,s}$$



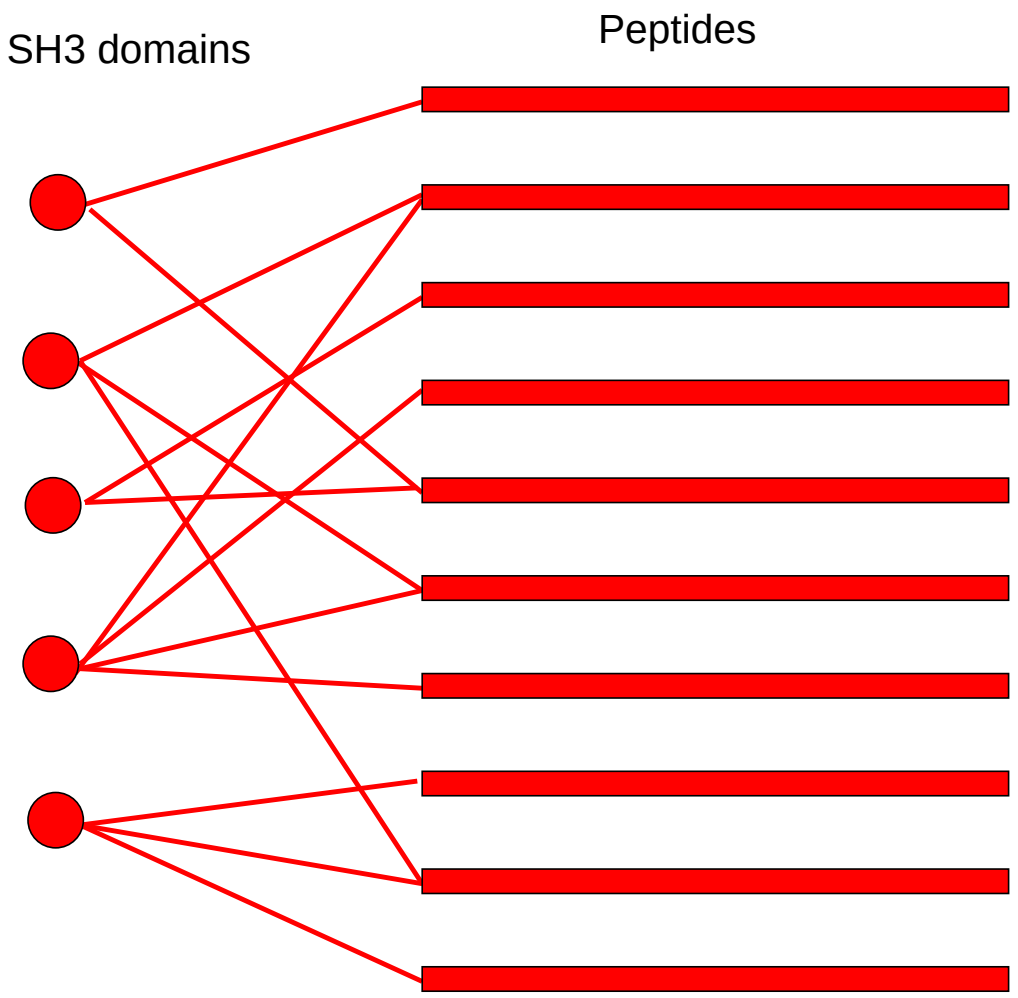
$$\tilde{\mathbf{C}}_{d,s} = \sum_s \varepsilon_{d,s} \mathbf{C}_{d,s}$$



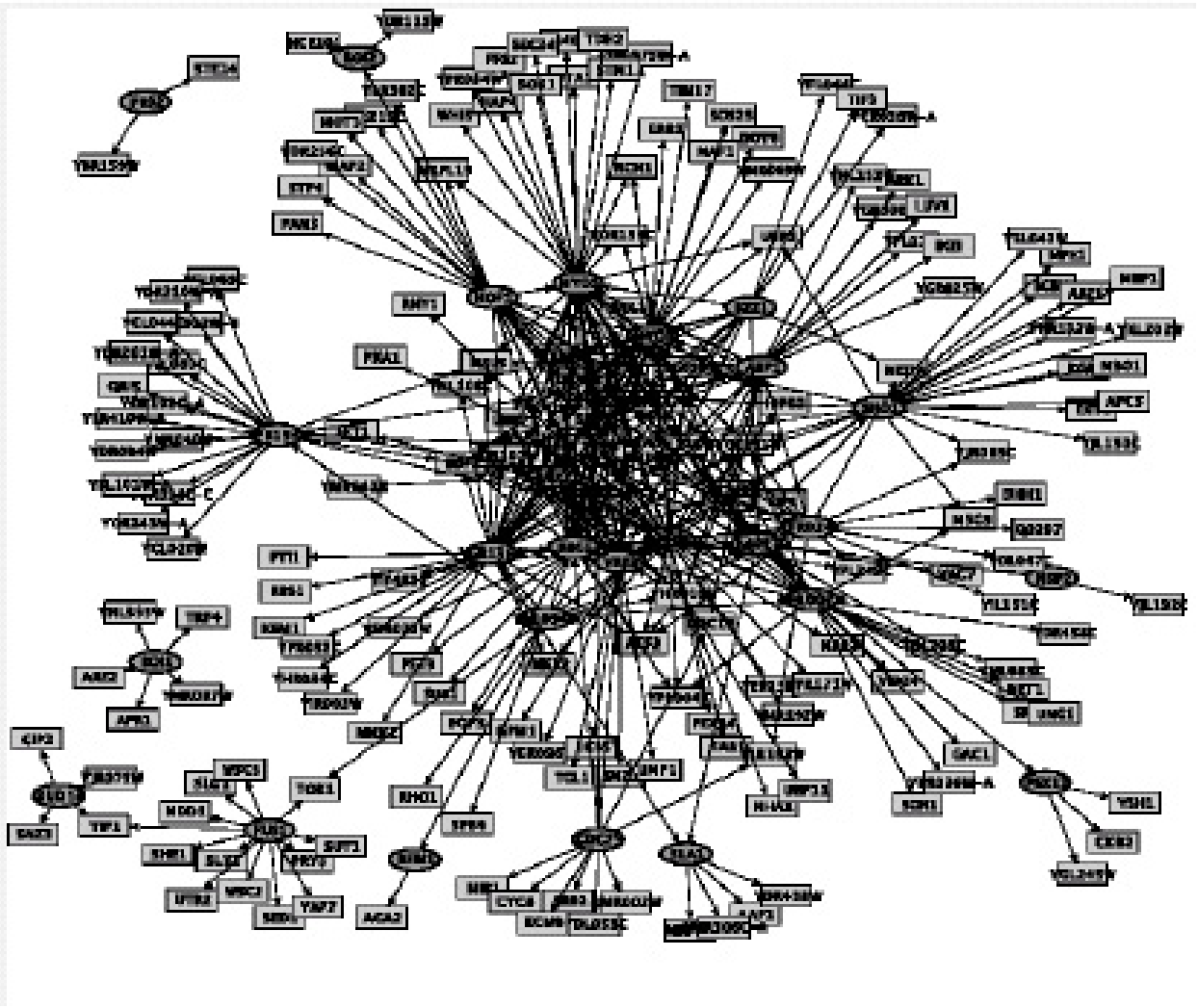
$$\tilde{\mathbf{C}}_{d,s} = \sum_s \varepsilon_{d,s} \mathbf{C}_{d,s} + \lambda_1 \sum_d \varepsilon_{d,s} \mathbf{C}_{d,s}$$



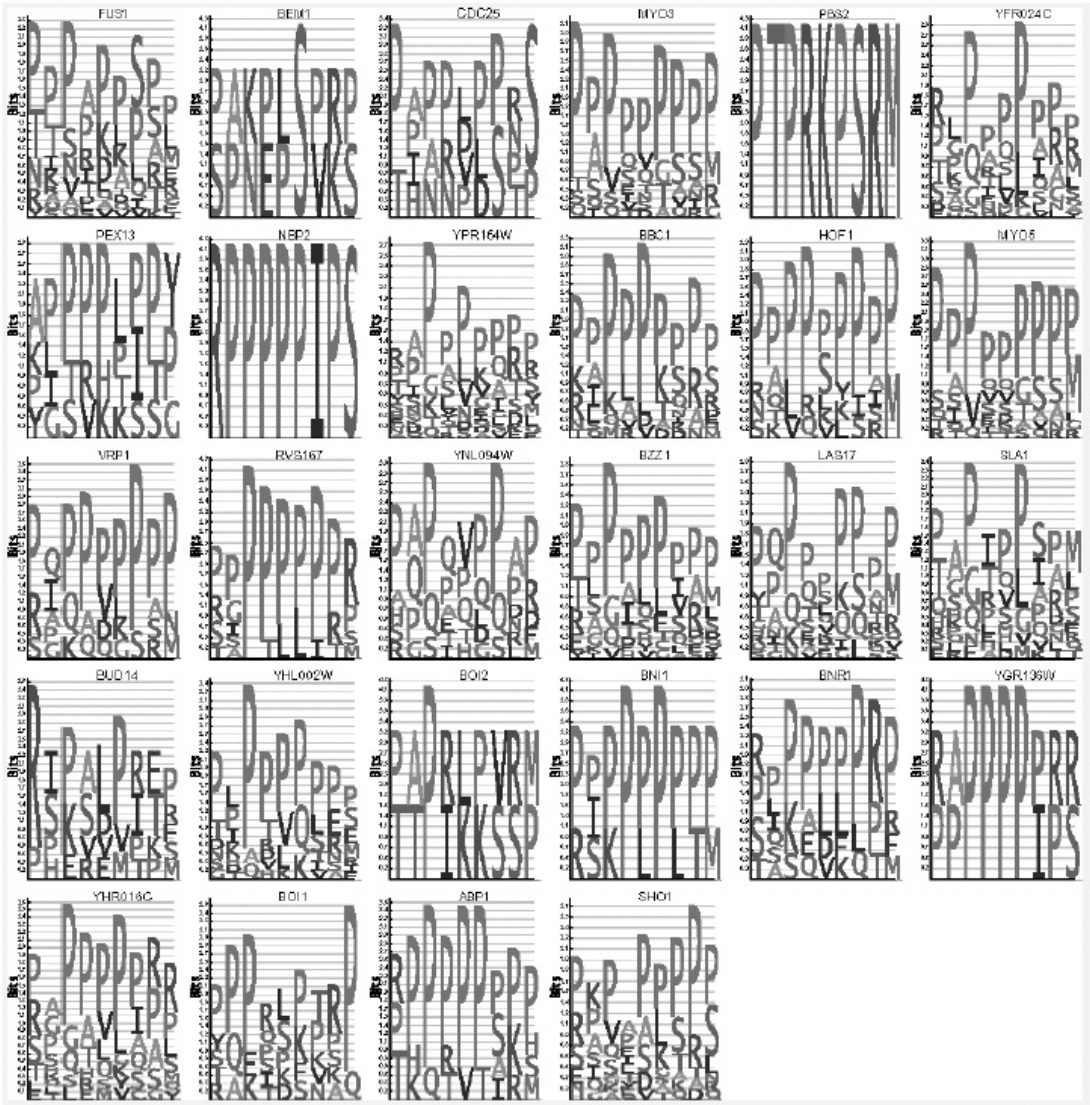
$$\tilde{\mathbf{C}}_{d,s} = \sum_s \varepsilon_{d,s} \mathbf{C}_{d,s} + \lambda_1 \sum_d \varepsilon_{d,s} \mathbf{C}_{d,s}$$



$$\tilde{\mathbf{C}}_{d,s} = \sum_s \varepsilon_{d,s}\mathbf{C}_{d,s} + \lambda_1 \sum_d \varepsilon_{d,s}\mathbf{C}_{d,s} + \lambda_2 \sum_s \sum_d \varepsilon_{d,s}\mathbf{C}_{d,s}$$



SH3 domain protein interaction network in *S. cerevisiae*; from Tong et al. (2002)


FUS1
BEM1
CDC25
MYO3
PBS2
YFR024C
PEX13
NBP2
YPR164W
BBC1
HOF1
MYO5
VRP1
RVS167
YNL094W
BZZ1
LAS17
SLA1
BUD14
YHL002W
BOI2
BNI1
BNR1
YGR136W
YHR018C
BOI1
ABP1
SHO1

Reiss & Schwikowski (2004)

$$\tilde{\mathbf{C}}_{d,s} = \sum_s \varepsilon_{d,s}\mathbf{C}_{d,s} + \lambda_1 \sum_d \varepsilon_{d,s}\mathbf{C}_{d,s} + \lambda_2 \sum_s \sum_d \varepsilon_{d,s}\mathbf{C}_{d,s}$$

Reiss & Schwikowski (2004)

$$\tilde{\mathbf{C}}_{d,s} = \sum_s \varepsilon_{d,s}\mathbf{C}_{d,s} + \lambda_1 \sum_d \varepsilon_{d,s}\mathbf{C}_{d,s} + \lambda_2 \sum_s \sum_d \varepsilon_{d,s}\mathbf{C}_{d,s}$$

Heuristic modification to make the model more discriminative: Give higher probability to sites that are distinct from non-binding motifs. New tuning parameter λ_3

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 5 2006, pages 532–540
doi:10.1093/bioinformatics/bti804

Sequence analysis

A regularized discriminative model for the prediction of protein–peptide interactions

Wolfgang P. Lehrach[1,2,*], Dirk Husmeier[2] and Christopher K. I. Williams[1]

[1]University of Edinburgh, Edinburgh EH1 2QL, UK and [2]Biomathematics and Statistics Scotland, Edinburgh EH9 3JZ, UK

Received on August 4, 2005; revised on November 23, 2005; accepted on November 25, 2005
Advance Access publication January 5, 2006
Associate Editor: Keith A Crandall

Binding sequence: R=1, motif starting at position m+1

$$P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) = \prod_{t=1}^{N} \theta_0(S_t) \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$

Binding sequence: R=1, motif starting anywhere

$$\begin{aligned} & P(S_1, S_2, \ldots, S_N | R = 1) \\ = & \sum_{m=0}^{N-W} P(start = m + 1) P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) \\ = & \prod_{t=1}^{N} \theta_0(S_t) \frac{1}{N - W + 1} \sum_{m=0}^{N-W} \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned}$$

Binding sequence: R=1, motif starting at position m+1

$$P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) = \prod_{t=1}^{N} \theta_0(S_t) \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$ Binding sequence: R=1, motif starting anywhere

$$\begin{aligned} & P(S_1, S_2, \ldots, S_N | R = 1) \\ = & \sum_{m=0}^{N-W} P(start = m + 1) P(S_1, S_2, \ldots, S_N | R = 1, start = m + 1) \\ = & \prod_{t=1}^{N} \theta_0(S_t) \frac{1}{N - W + 1} \sum_{m=0}^{N-W} \prod_{k=1}^{W} \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned}$$

Objective: Prediction of binding activity from sequence:

$$P(R = 1 | S_1, S_2, \ldots, S_N)$$

Apply Bayes rule:

$$
\begin{aligned}
P(R=1|S_1,S_2,\ldots,S_N) &= \frac{P(S_1,S_2,\ldots,S_N|R=1)P(R=1)}{P(S_1,S_2,\ldots,S_N)} \\
&= \frac{P(S_1,S_2,\ldots,S_N|R=1)P(R=1)}{P(S_1,S_2,\ldots,S_N|R=0)P(R=0)+P(S_1,S_2,\ldots,S_N|R=1)P(R=1)} \\
&= \left(1+\frac{P(R=0)P(S_1,S_2,\ldots,S_N|R=0)}{P(R=1)P(S_1,S_2,\ldots,S_N|R=1)}\right)^{-1} \\
&= \left(1+\left[\frac{P(R=1)}{P(R=0)}\frac{1}{(N-W+1)}\sum_{m=0}^{N-W}\prod_{k=1}^{W}\frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}\right]^{-1}\right)^{-1}
\end{aligned}
$$

Apply Bayes rule:

$$
\begin{aligned}
P(R=1|S_1,S_2,\ldots,S_N) &= \frac{P(S_1,S_2,\ldots,S_N|R=1)P(R=1)}{P(S_1,S_2,\ldots,S_N)} \\
&= \frac{P(S_1,S_2,\ldots,S_N|R=1)P(R=1)}{P(S_1,S_2,\ldots,S_N|R=0)P(R=0)+P(S_1,S_2,\ldots,S_N|R=1)P(R=1)} \\
&= \left(1+\frac{P(R=0)P(S_1,S_2,\ldots,S_N|R=0)}{P(R=1)P(S_1,S_2,\ldots,S_N|R=1)}\right)^{-1} \\
&= \left(1+\left[\frac{P(R=1)}{P(R=0)}\frac{1}{(N-W+1)}\sum_{m=0}^{N-W}\prod_{k=1}^{W}\frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}\right]^{-1}\right)^{-1}
\end{aligned}
$$

Define:

$$
w_k(l)=\log\frac{\psi_k(l)}{\theta_0(l)},\ w_0=\frac{P(R=1)}{P(R=0)},\ \mathrm{logit}(z)=\frac{1}{1+\exp(-z)}
$$

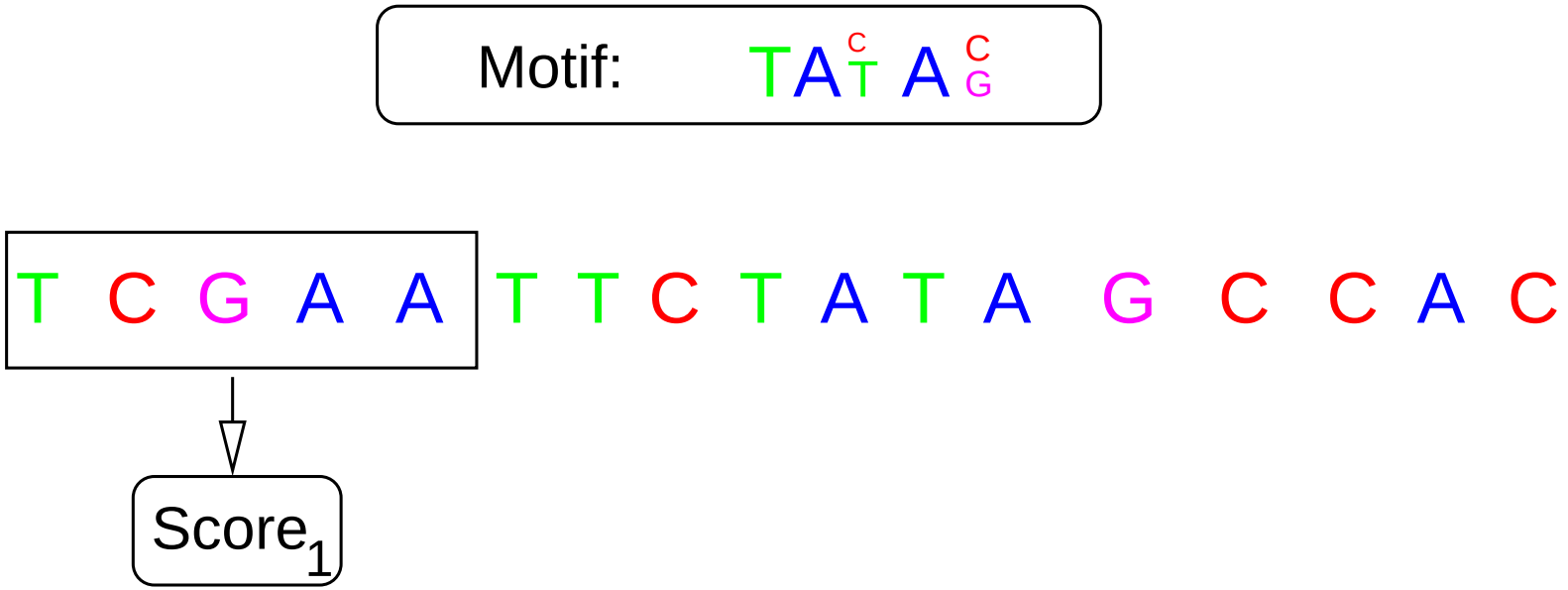
$$P(R=1|S_1,S_2,\ldots,S_N)$$

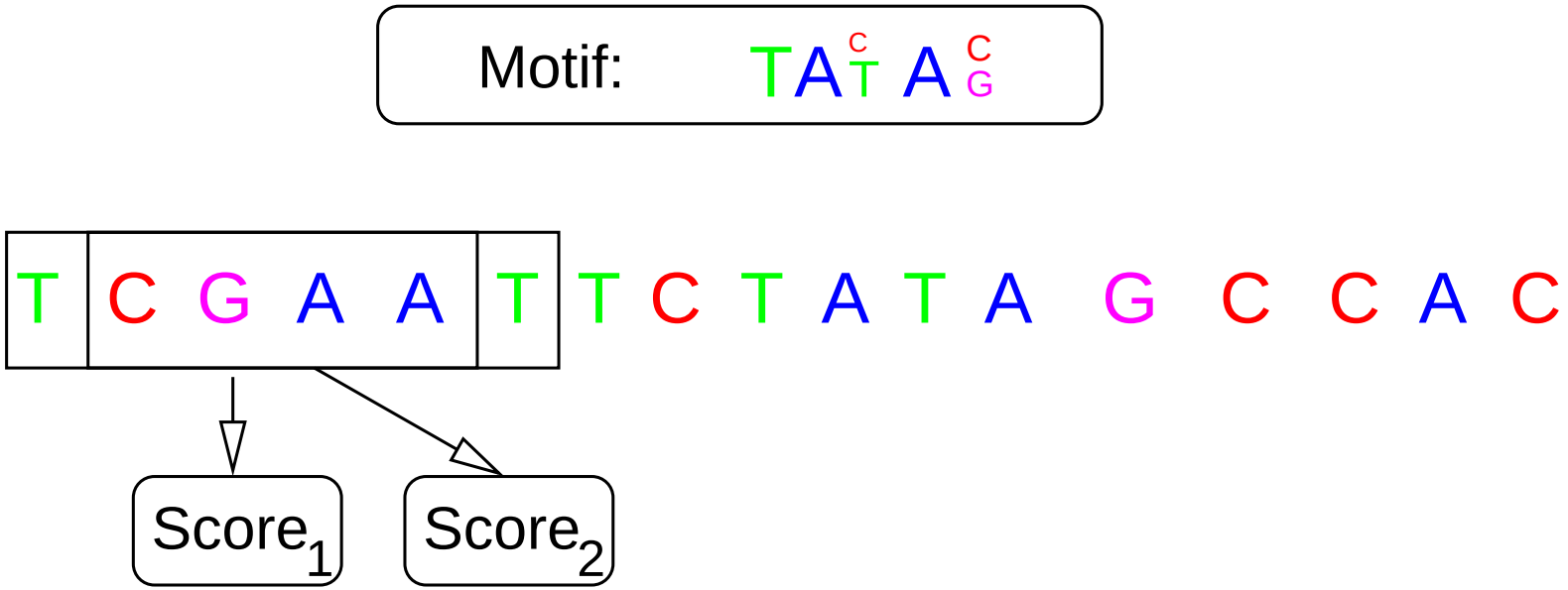
$$= \quad \text{logit}\left(\log\left[\frac{w_0}{N-W+1}\sum_{m=0}^{N-W}\exp\left(\sum_{k=1}^{W} w_k(S_{t+k})\right)\right]\right)$$

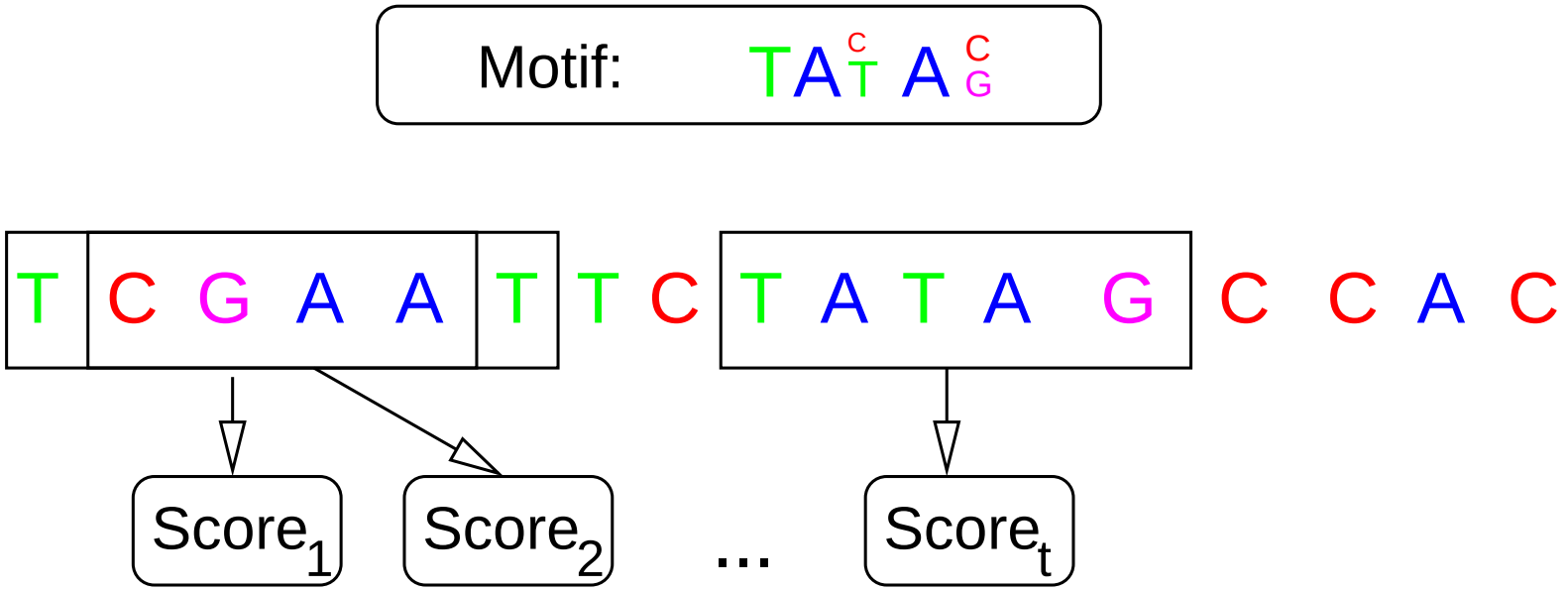
$4 \times W + 1$ parameters: $w_k(l)$, w_0

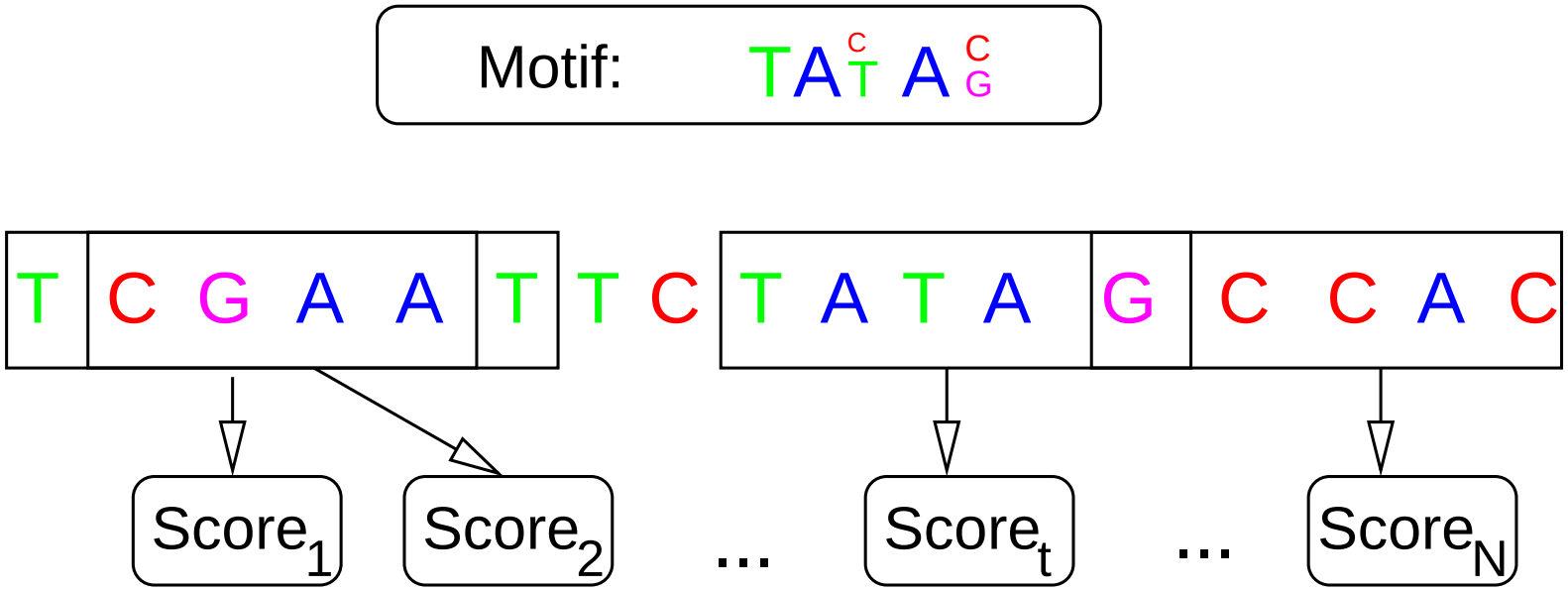
Motif: TA[C/T] A[C/G]

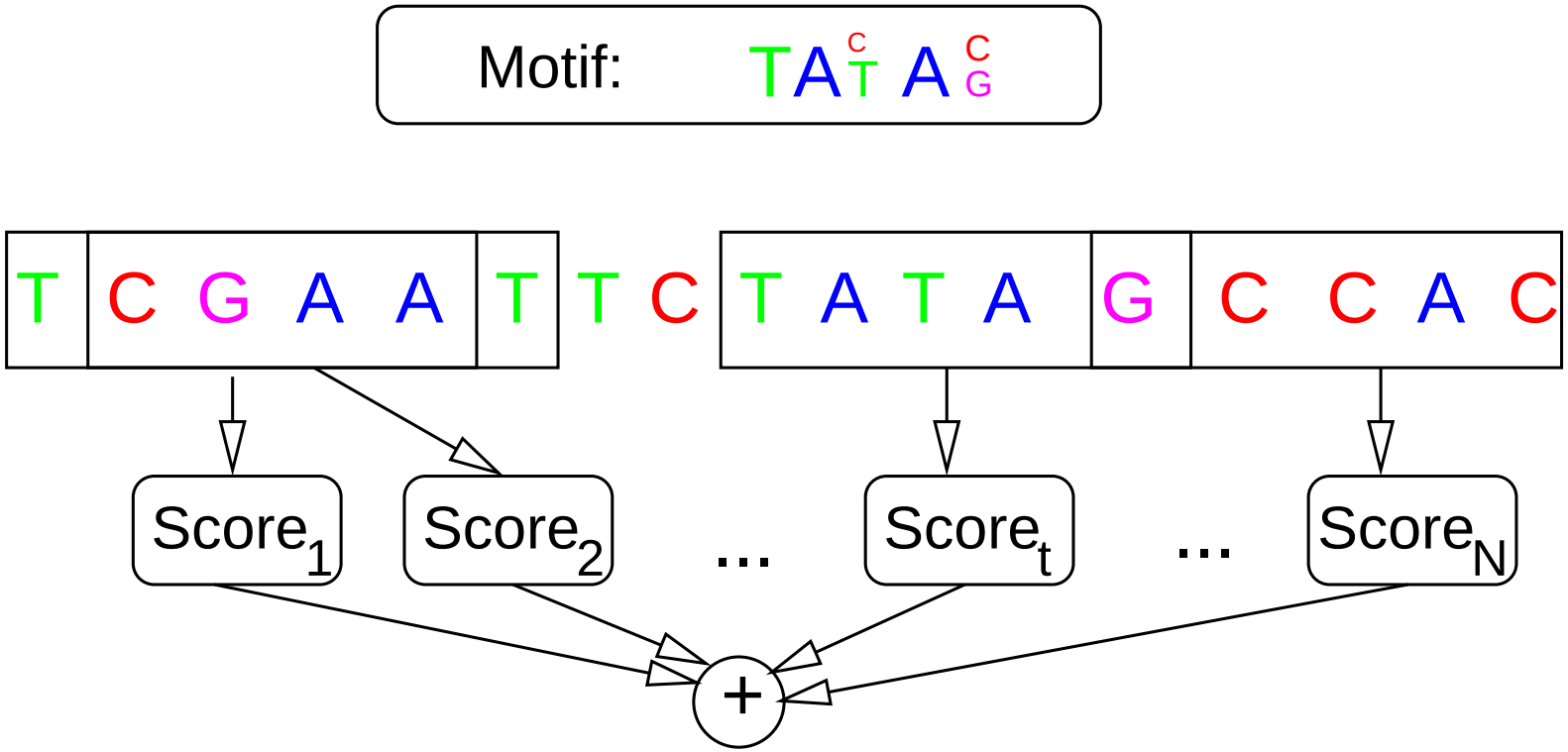
T C G A A T T C T A T A G C C A C

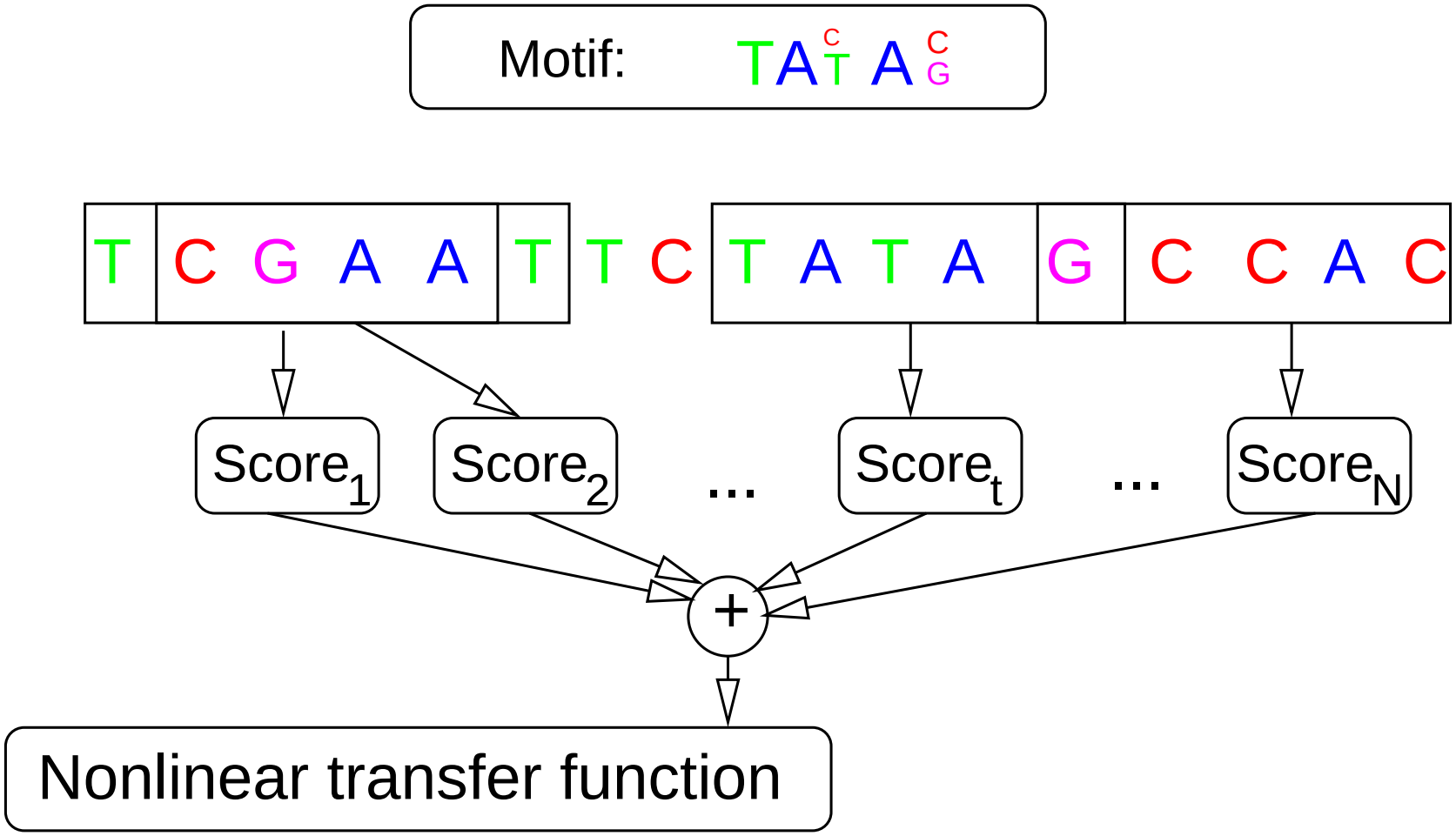

Motif:
TA C/T A C/G
T C G A A T T C T A T A G C C A C
Score1

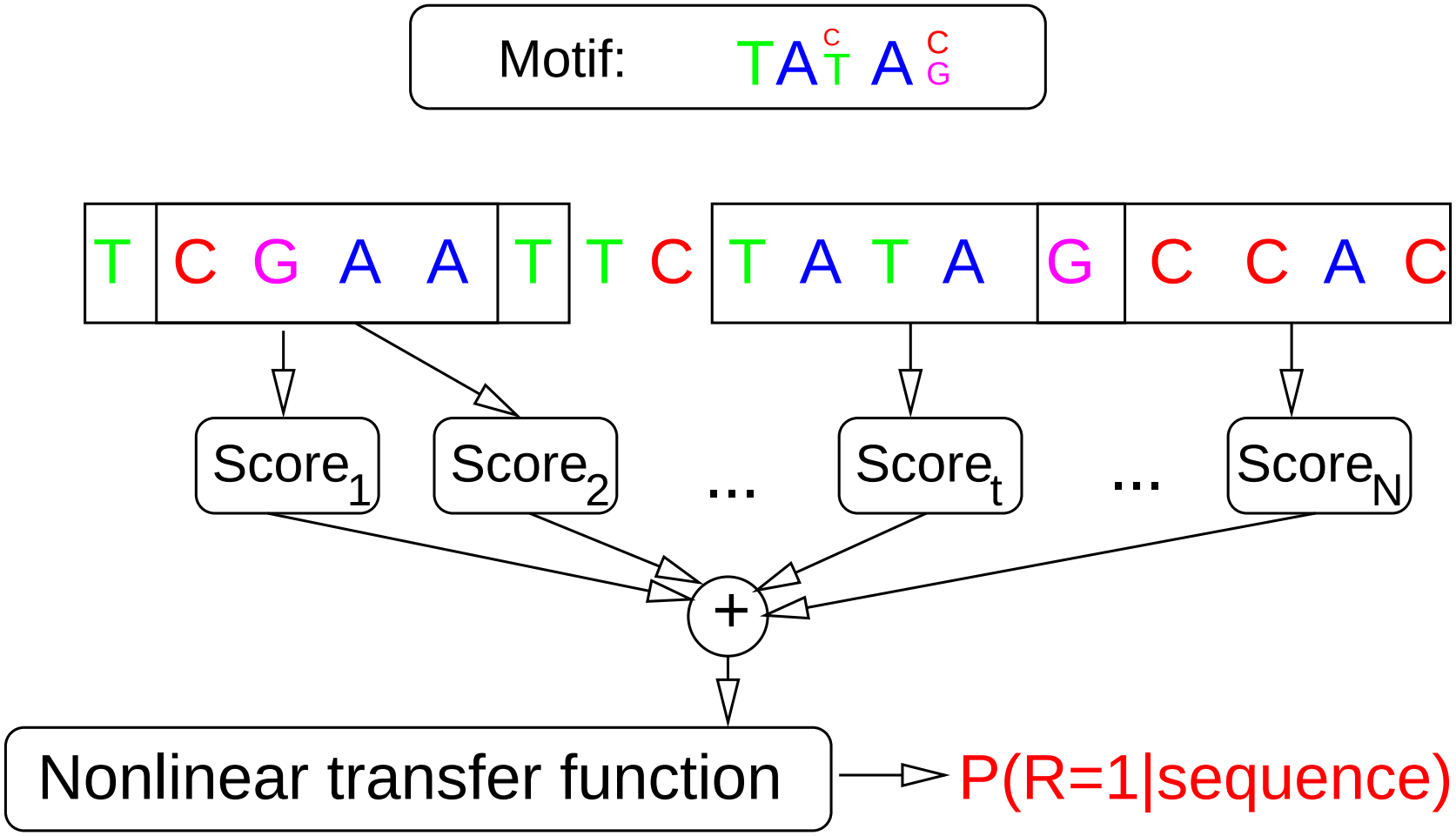

Motif: TA[C/T] A[C/G]
T C G A A T T C T A T A G C C A C
Score1
Score2


Motif: TA[C/T] A[C/G]
T C G A A T T C T A T A G C C A C
Score1
Score2
...
Scoret


Motif: TA[C/T] A[C/G]
T C G A A T T C T A T A G C C A C
Score1
Score2
...
Scoret
...
ScoreN


Motif: TA[C/T] A[C/G]
T C G A A T T C T A T A G C C A C
Score1
Score2
...
Scoret
...
ScoreN
+


Motif: TA[C/T] A[C/G]
T C G A A T T C T A T A G C C A C
Score1
Score2
...
Scoret
...
ScoreN
+
Nonlinear transfer function


Motif: TA[C/T] A[C/G]
T C G A A T T C T A T A G C C A C
Score1
Score2
...
Scoret
...
ScoreN
+
Nonlinear transfer function
P(R=1|sequence)

$$P(R=1|S_1,S_2,\ldots,S_N)$$

$$= \quad \text{logit}\left(\log\left[\frac{w_0}{N-W+1}\sum_{m=0}^{N-W}\exp\left(\sum_{k=1}^{W} w_k(S_{t+k})\right)\right]\right)$$

$4 \times W + 1$ parameters: $w_k(l)$, w_0

Parameter estimation

Data D: Set of sequences $\mathbf{x}_i$ with associated interaction indicators $R_i \in \{0, 1\}$

Model predicts an interaction R_i given the sequence $\mathbf{x}_i$:

$$y(\mathbf{x}_i, \mathbf{w}) \;=\; P(R_i = 1 | \mathbf{x}_i, \mathbf{w})$$

Parameter estimation

Data D: Set of sequences $\mathbf{x}_i$ with associated interaction indicators $R_i \in \{0, 1\}$

Model predicts an interaction R_i given the sequence $\mathbf{x}_i$:

$$y(\mathbf{x}_i, \mathbf{w}) = P(R_i = 1 | \mathbf{x}_i, \mathbf{w})$$

$$P(D|\mathbf{w}) = \prod_i y(\mathbf{x}_i, \mathbf{w})^{R_i} [1 - y(\mathbf{x}_i, \mathbf{w})]^{(1-R_i)}$$

$$\log P(D|\mathbf{w}) = \sum_i R_i \log y(\mathbf{x}_i, \mathbf{w}) + (1 - R_i) \log y(\mathbf{x}_i, \mathbf{w})$$

Parameter estimation

Maximum likelihood: $\mathrm{argmax}_{\mathbf{w}} P(D|\mathbf{w})$

Parameter estimation

Maximum likelihood: $\text{argmax}_{\mathbf{w}} P(D|\mathbf{w})$

Iterative optimisation scheme with gradient descent:

$$E_D(\mathbf{w}) \;=\; -\log P(D|\mathbf{w})$$

$$\Delta\mathbf{w} \;\propto\; -\nabla_{\mathbf{w}} E_D(\mathbf{w})$$

Parameter estimation

Maximum likelihood: $\text{argmax}_{\mathbf{w}} P(D|\mathbf{w})$

Iterative optimisation scheme with gradient descent:

$$E_D(\mathbf{w}) = -\log P(D|\mathbf{w})$$

$$\Delta\mathbf{w} \propto -\nabla_{\mathbf{w}} E_D(\mathbf{w})$$

Problem: overfitting!

Parameter estimation

Maximum likelihood: $\text{argmax}_{\mathbf{w}} P(D|\mathbf{w})$

Iterative optimisation scheme with gradient descent:

$$E_D(\mathbf{w}) \;=\; -\log P(D|\mathbf{w})$$

$$\Delta\mathbf{w} \;\propto\; -\nabla_{\mathbf{w}} E_D(\mathbf{w})$$

Problem: overfitting!

Regularisation:

$$P(\mathbf{w}) = \tfrac{1}{Z}\exp(-\alpha E_R(\mathbf{w}))$$

Maximum a posteriori:

$$\text{argmax}_{\mathbf{w}} P(\mathbf{w}|D)$$

Bayes rule: $P(\mathbf{w}|D) \propto P(D|\mathbf{w})P(\mathbf{w})$

$$\text{argmax}_{\mathbf{w}} [\log P(D|\mathbf{w}) + \log P(\mathbf{w})]$$

Maximum a posteriori:

$$\text{argmax}_{\mathbf{w}} P(\mathbf{w}|D)$$

Bayes rule: $P(\mathbf{w}|D) \propto P(D|\mathbf{w})P(\mathbf{w})$

$$\text{argmax}_{\mathbf{w}}[\log P(D|\mathbf{w}) + \log P(\mathbf{w})]$$

$$P(\mathbf{w}) \propto \exp[-\alpha E_R(\mathbf{w})]$$
$$P(D|\mathbf{w}) \propto \exp[-E_D(\mathbf{w})]$$

$$\Delta\mathbf{w} \propto -\nabla_{\mathbf{w}} E_D(\mathbf{w}) - \alpha \nabla_{\mathbf{w}} E_R(\mathbf{w})$$

Weight decay:

$$\Delta \mathbf{w} \propto -\nabla_{\mathbf{w}} E_D(\mathbf{w}) - \alpha \nabla_{\mathbf{w}} E_R(\mathbf{w})$$

Gaussian prior: $E_R(\mathbf{w}) = \mathbf{w}^2$

Laplacian prior: $E_R(\mathbf{w}) = |\mathbf{w}|$

Justification of regularisation

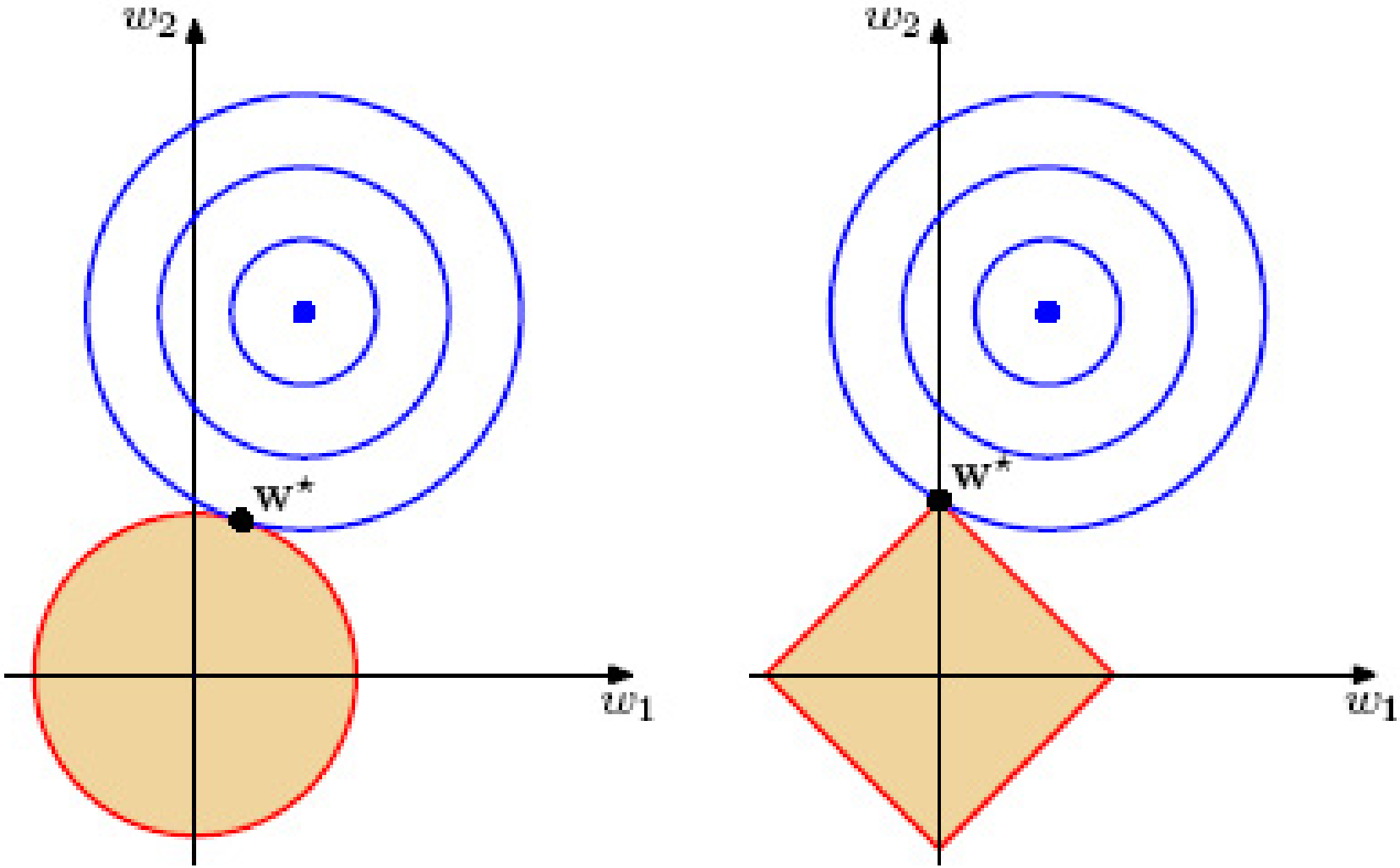
$$P(R=1|S_1,S_2,\ldots,S_N) = \frac{P(S_1,S_2,\ldots,S_N|R=1)P(R=1)}{P(S_1,S_2,\ldots,S_N)}$$

$$= \frac{P(S_1,S_2,\ldots,S_N|R=1)P(R=1)}{P(S_1,S_2,\ldots,S_N|R=0)P(R=0)+P(S_1,S_2,\ldots,S_N|R=1)P(R=1)}$$

$$= \left(1+\frac{P(R=0)P(S_1,S_2,\ldots,S_N|R=0)}{P(R=1)P(S_1,S_2,\ldots,S_N|R=1)}\right)^{-1}$$

$$= \left(1+\left[\frac{P(R=1)}{P(R=0)}\frac{1}{(N-W+1)}\sum_{m=0}^{N-W}\prod_{k=1}^{W}\frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}\right]^{-1}\right)^{-1}$$

Define: $w_k(l) = \log\frac{\psi_k(l)}{\theta_0(l)}$, $\quad w_0 = \log\frac{P(R=1)}{P(R=0)}$


w_2
w^*
w_1
w_2
w^*
w_1

Weight decay:

$$\Delta \mathbf{w} \propto -\nabla_{\mathbf{w}} E_D(\mathbf{w}) - \alpha \nabla_{\mathbf{w}} E_R(\mathbf{w})$$

Gaussian prior: $E_R(\mathbf{w}) = \mathbf{w}^2$

Laplacian prior: $E_R(\mathbf{w}) = |\mathbf{w}|$

The hyperparameter α can be integrated out analytically

$$P(\mathbf{w}) = \int_0^\infty P(\mathbf{w}|\alpha) P(\alpha) d\alpha$$

$$P(\mathbf{w}|\alpha) = \frac{\exp(-\alpha E_R)}{Z(\alpha)}$$

$$Z(\alpha) \propto \left(\frac{1}{\alpha}\right)^W$$

where W is the dimension of $\mathbf{w}$ (number of weights).

$$P(\mathbf{w}) \;=\; \int_0^\infty P(\mathbf{w}|\alpha)P(\alpha)d\alpha$$

Scale parameter: uninformative prior $P(\alpha) \propto \frac{1}{\alpha}$

$$
\begin{aligned}
P(\mathbf{w}) &= \int_0^\infty P(\mathbf{w}|\alpha)P(\alpha)d\alpha \\
&= C\int_0^\infty \exp(-\alpha E_R)\alpha^{W-1}d\alpha \\
&= CE_R^{-W}\int_0^\infty \exp(-\alpha E_R)(\alpha E_R)^{(W-1)}d(\alpha E_R) \\
&= CE_R^{-W}\int_0^\infty \exp(-u)u^{W-1}du \\
&= C\,E_R(\mathbf{w})^{-W}\,\Gamma(W)
\end{aligned}
$$

$$\log P(\mathbf{w}) \;=\; -W \log E_R(\mathbf{w}) + const$$

$$\nabla_{\mathbf{w}} \log P(\mathbf{w}) \;=\; -\frac{W}{E_R} \nabla_{\mathbf{w}} E_R(\mathbf{w})$$

Weight decay:

$$\Delta \mathbf{w} \;\propto\; -\nabla_{\mathbf{w}} E_D(\mathbf{w}) - \tilde{\alpha} \nabla_{\mathbf{w}} E_R(\mathbf{w}); \quad \tilde{\alpha} = \frac{W}{E_R}$$

Gaussian prior: $E_R(\mathbf{w}) = \mathbf{w}^2$

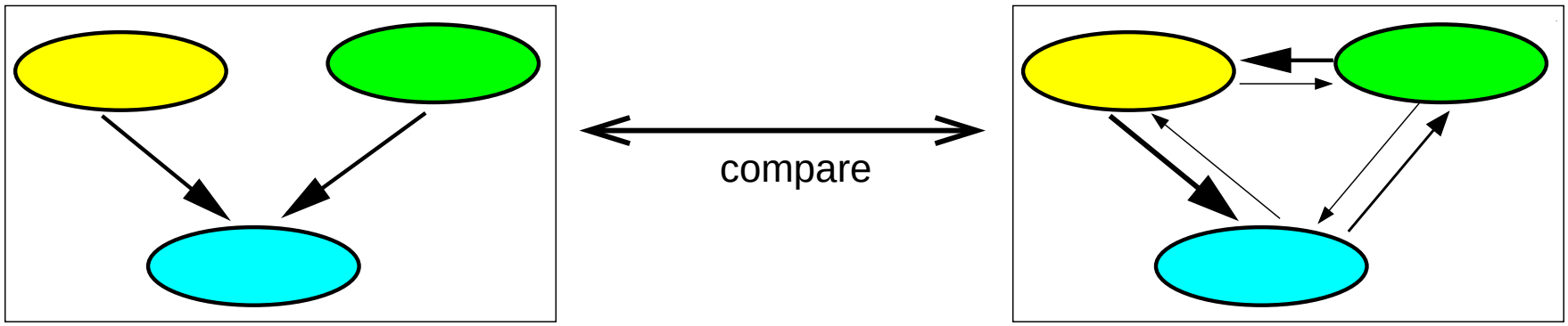
Laplacian prior: $E_R(\mathbf{w}) = |\mathbf{w}|$

Peter Williams (1995)

Bayesian regularisation and pruning
using a Laplacian prior

Neural Computation 7, 117–143

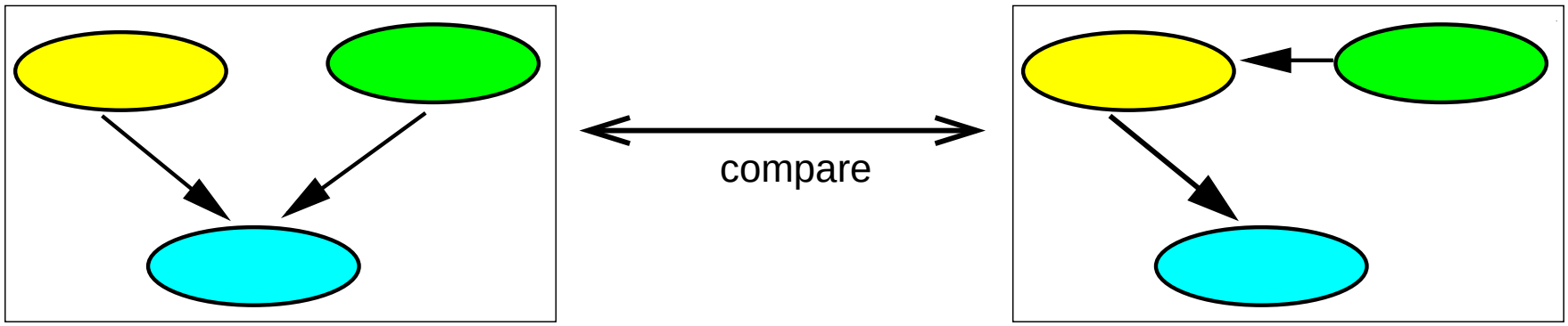
Evaluation



True network

Predicted network

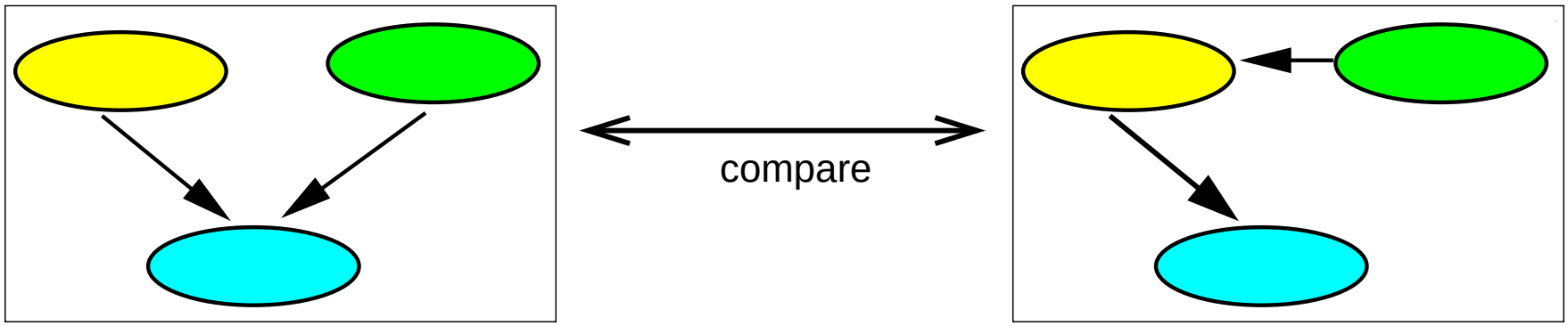
Probabilistic inference



True network

Predicted network

Thresholding



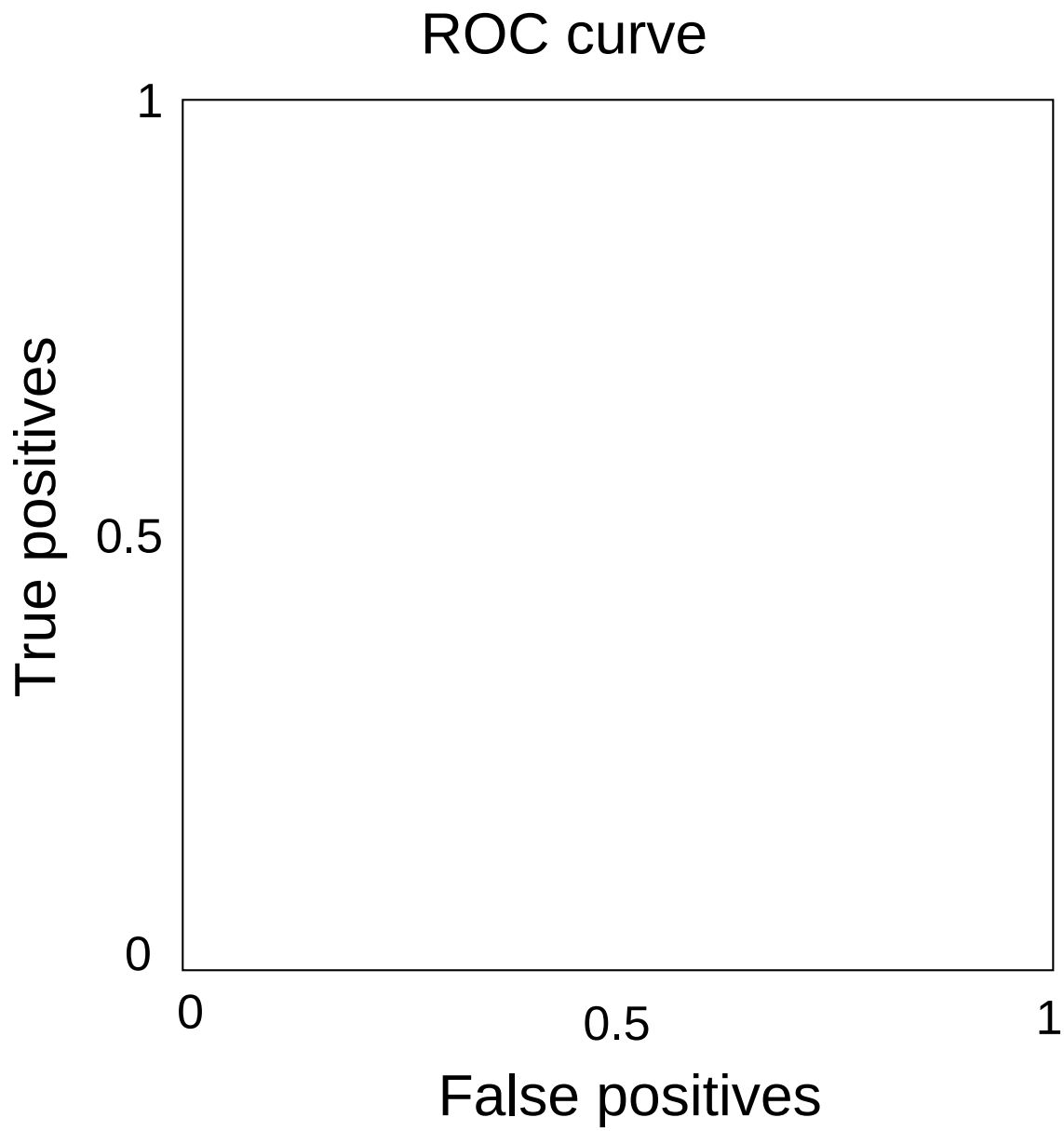
True network

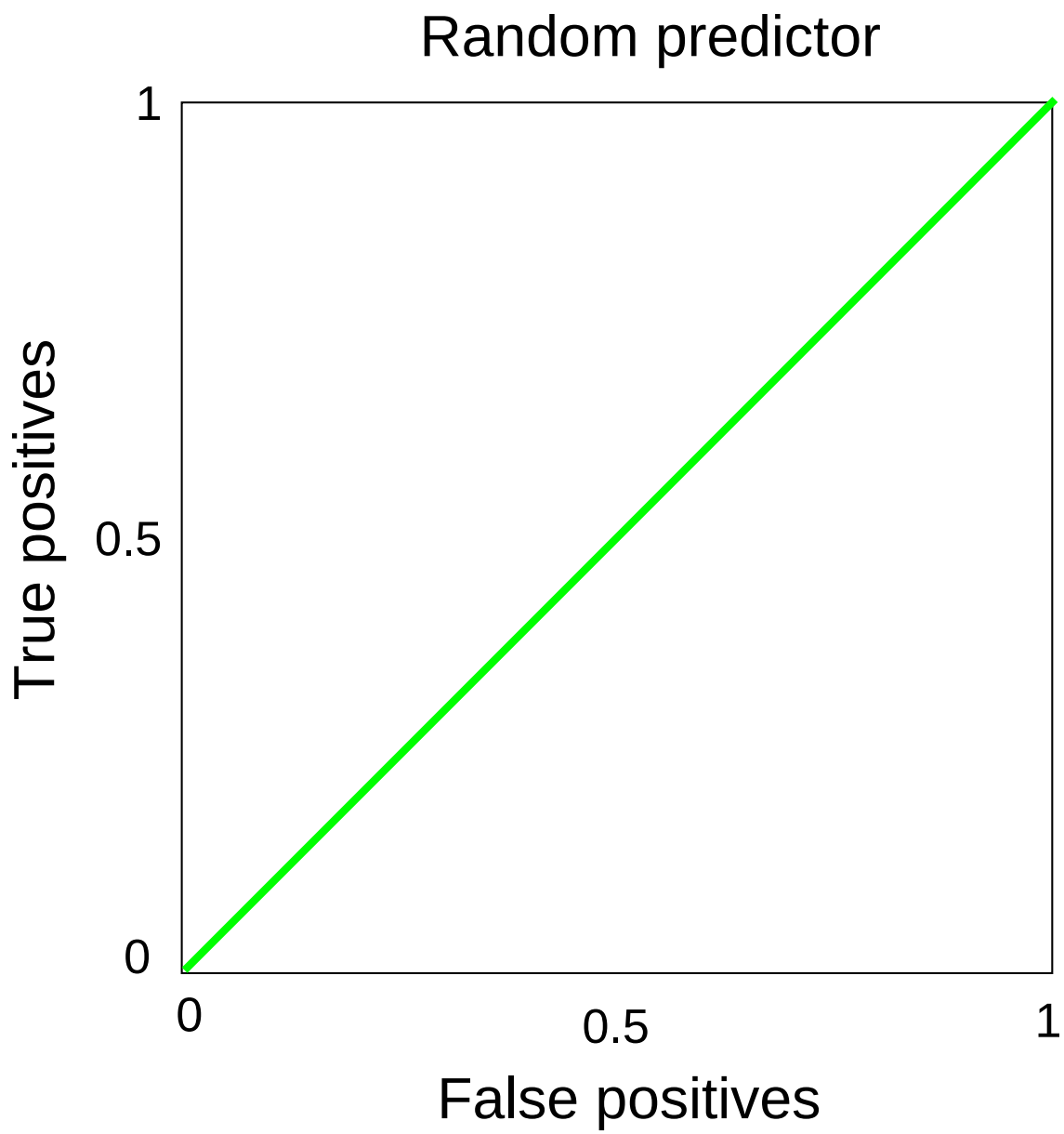
Predicted network

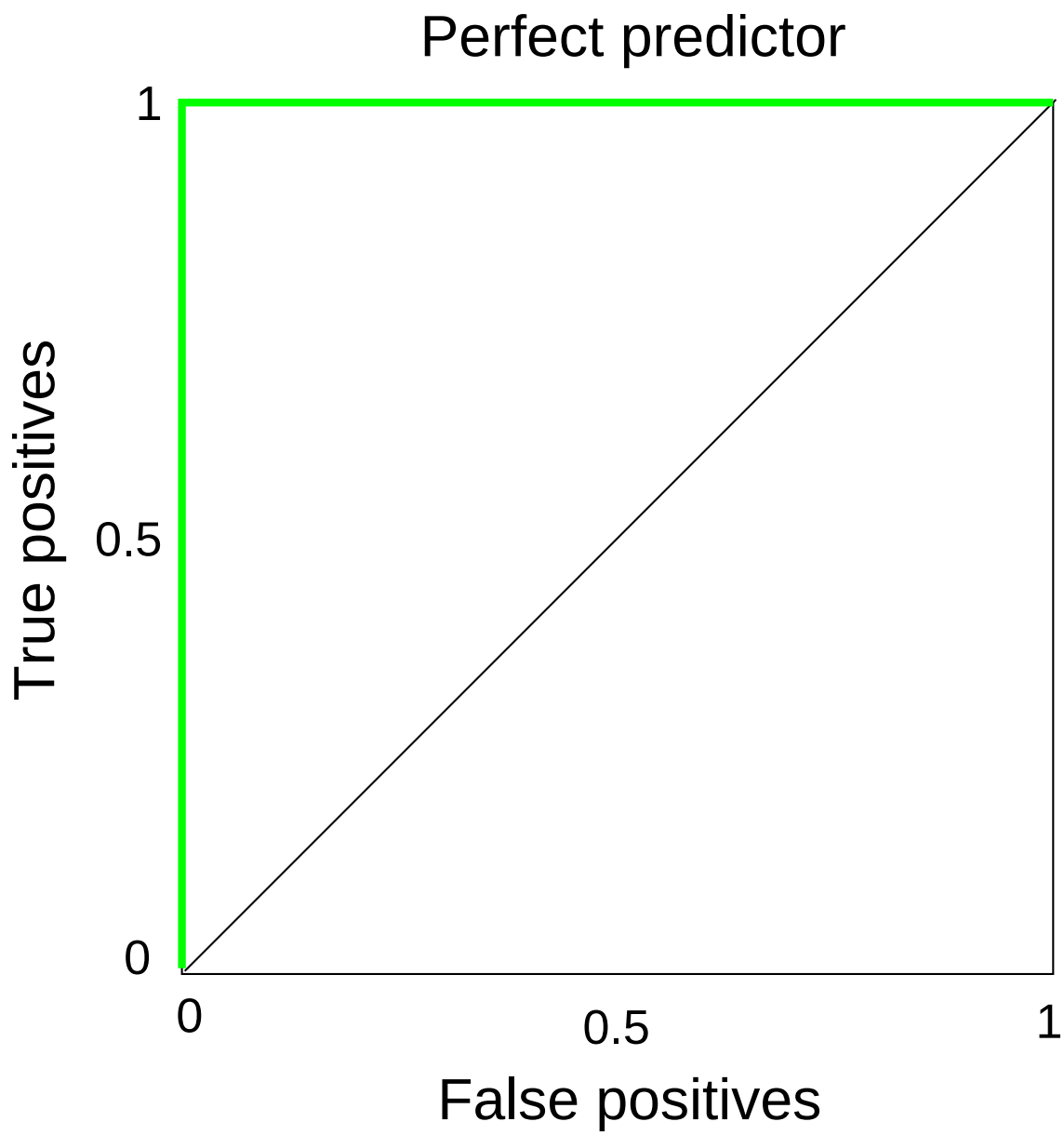
Thresholding

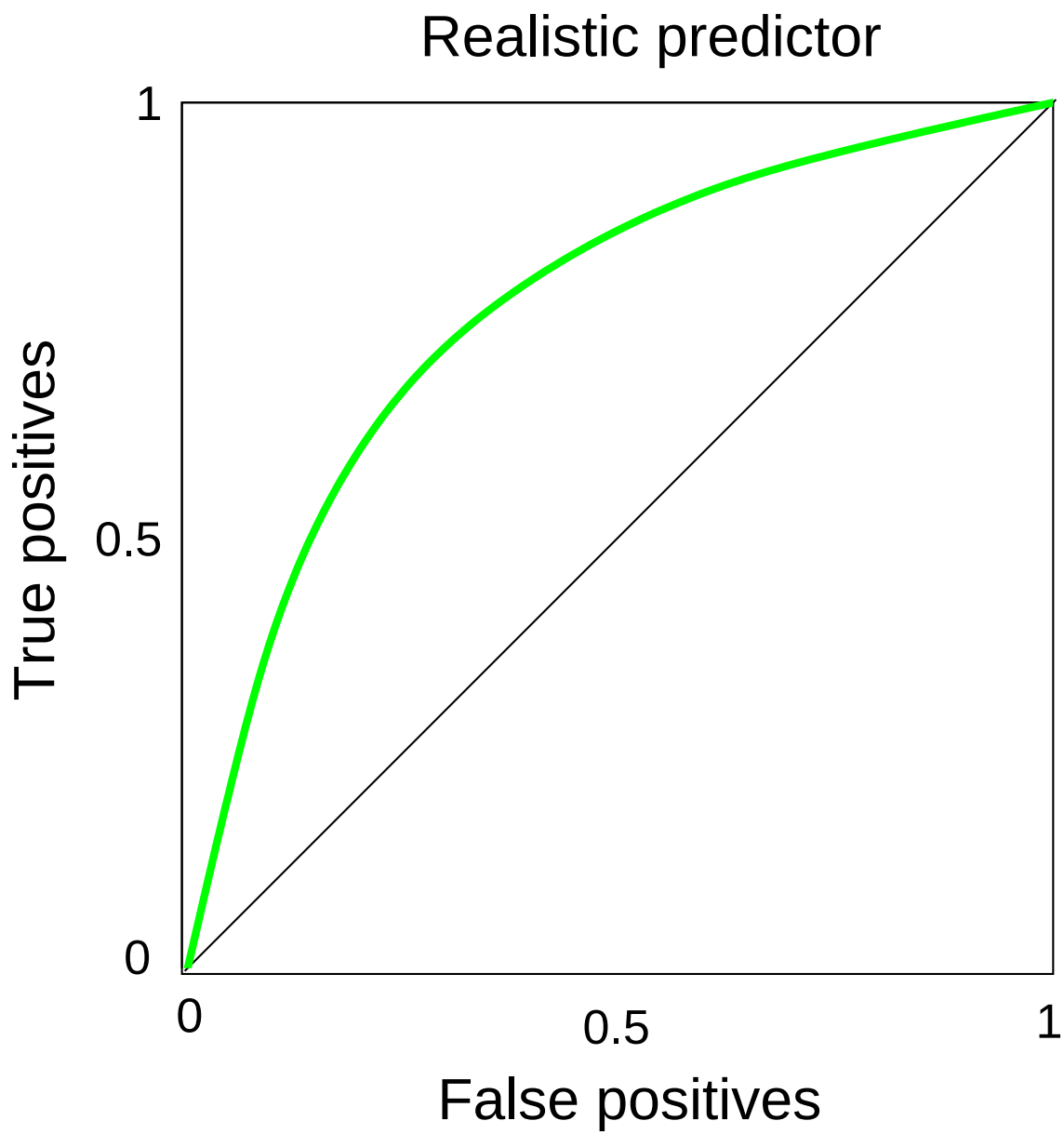
True positives

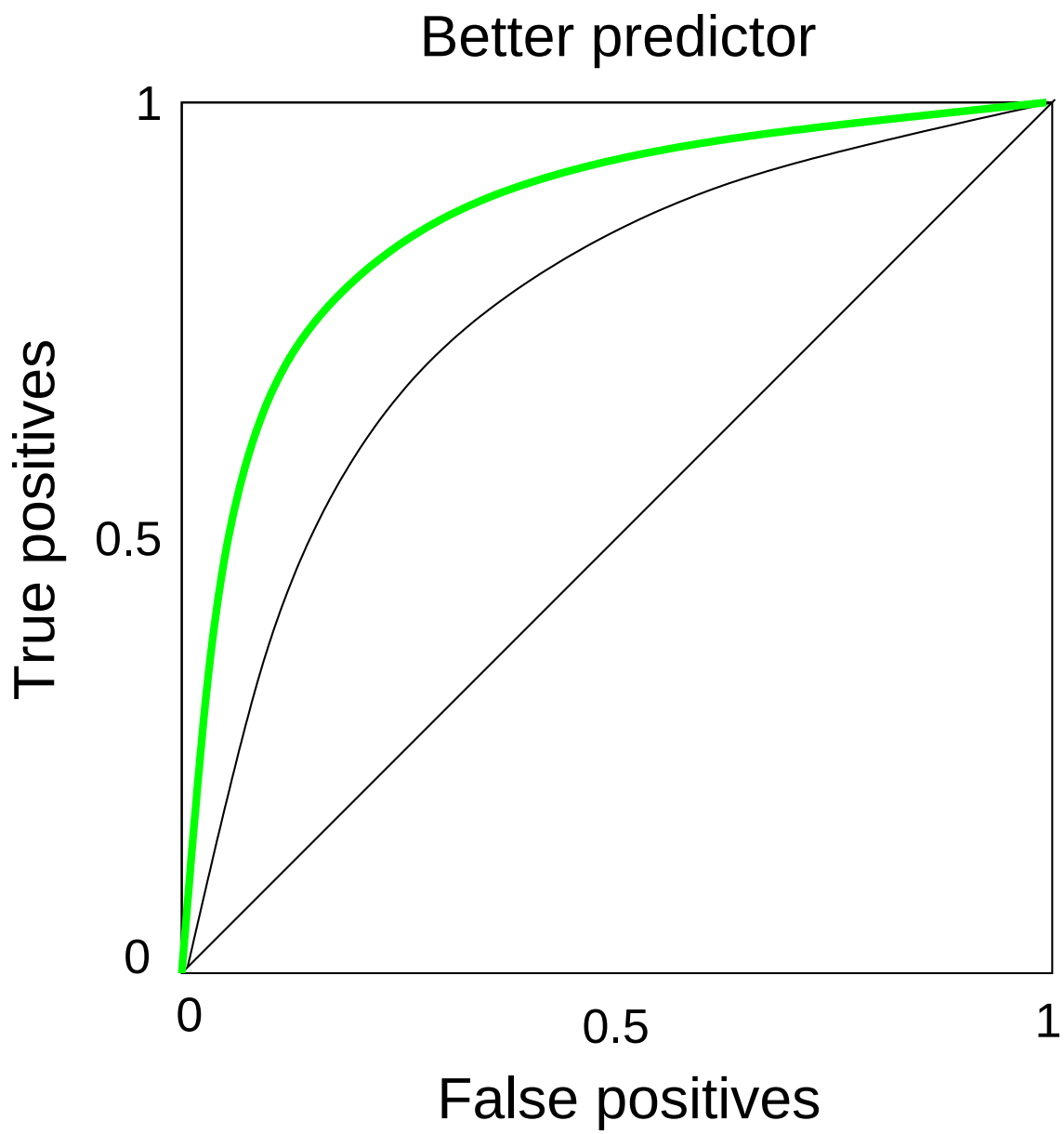
False positives

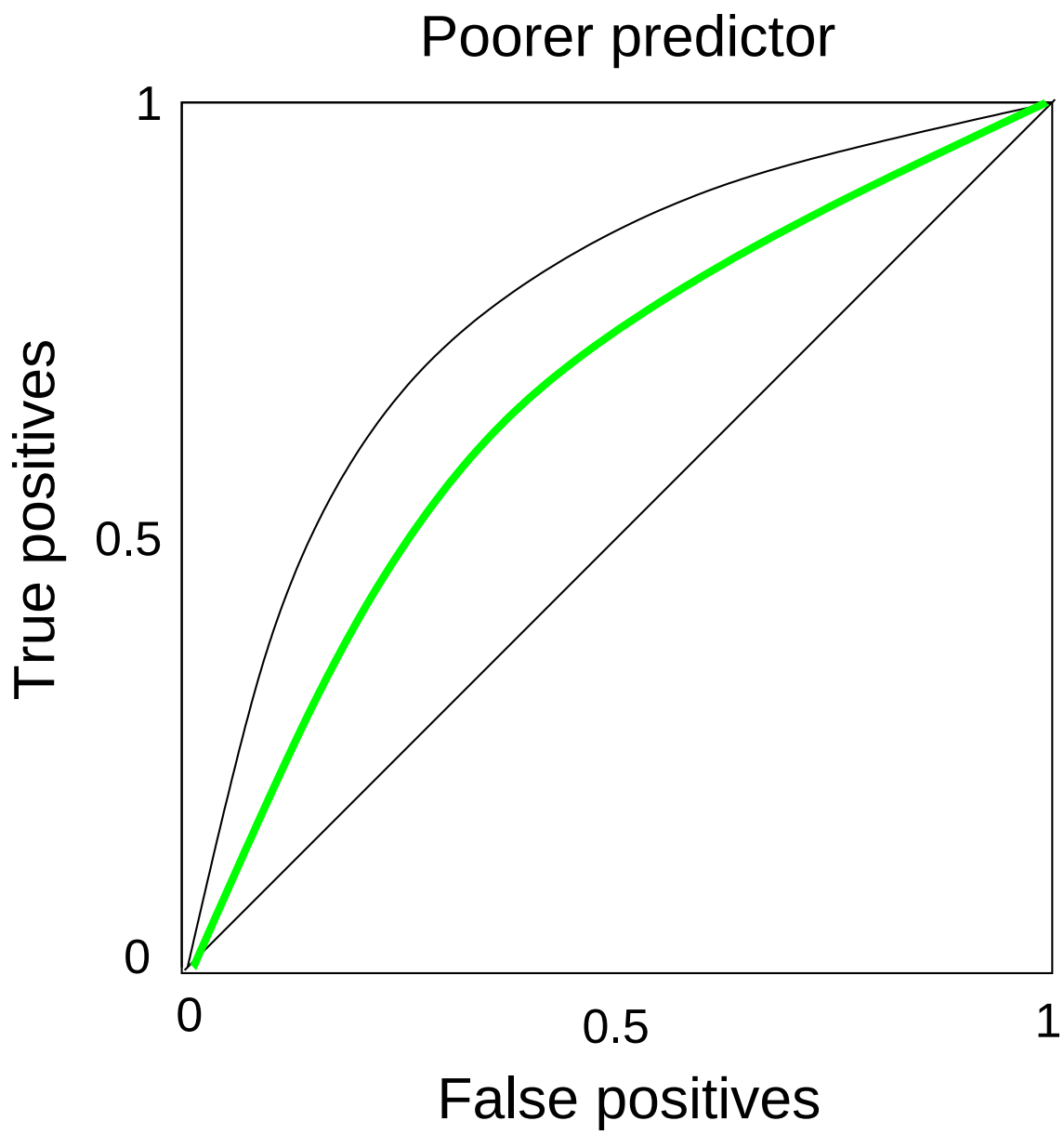

ROC curve
True positives
1
0.5
0
0
0.5
1
False positives


Random predictor
1
0.5
0
True positives
0
0.5
1
False positives


Perfect predictor
1
0.5
0
True positives
0
0.5
1
False positives


Realistic predictor
True positives
1
0.5
0
0
0.5
1
False positives


Better predictor
1
0.5
0
True positives
0
0.5
1
False positives

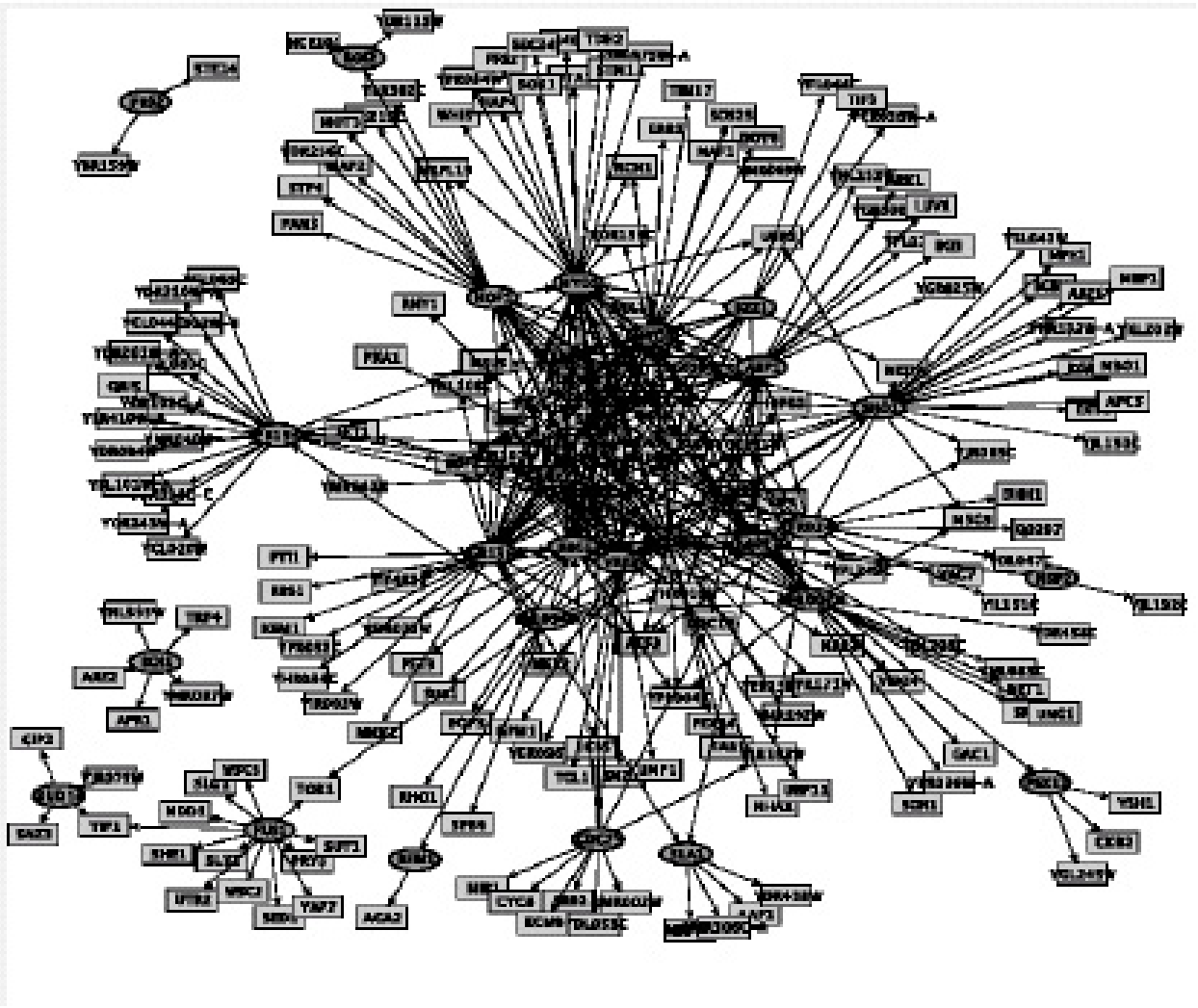

Poorer predictor
1
0.5
0
True positives
0
0.5
1
False positives

SH3 yeast two-hybrid interaction network

Tong et al. (2002), Science 295, 321-324

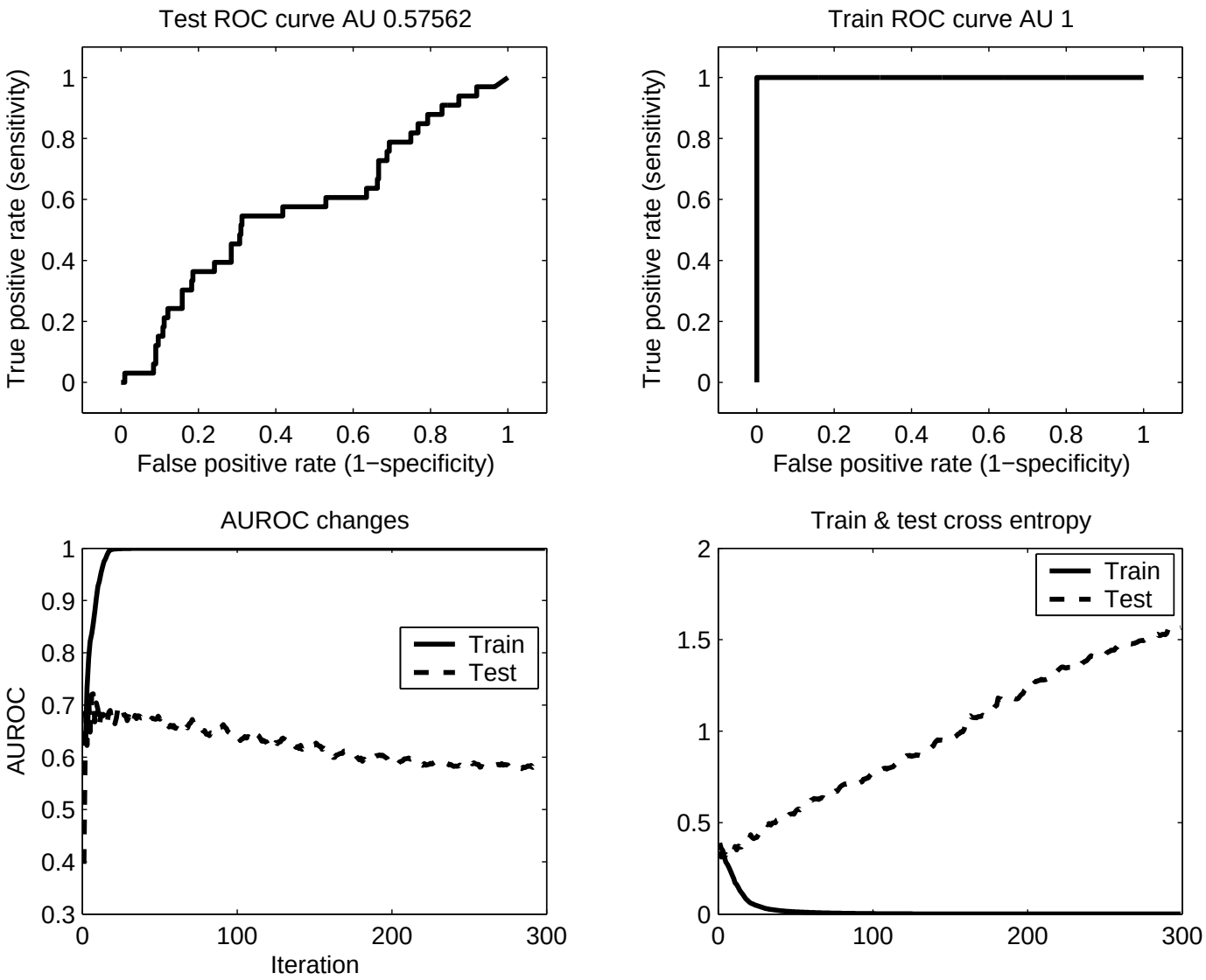
285 interactions between 28 SH3 proteins
and 143 binding peptides

9 binding partners per SH3 domain on average

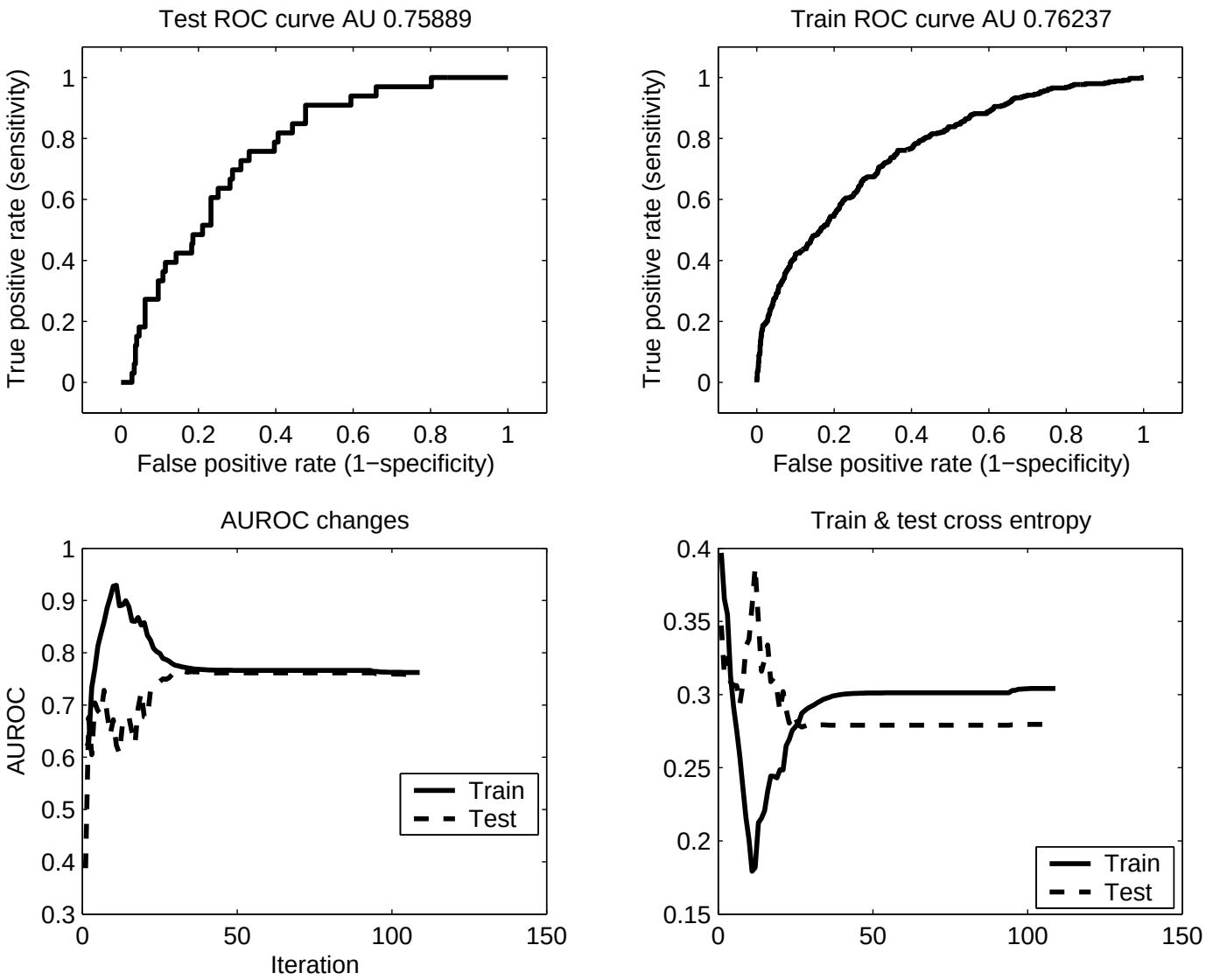


SH3 domain protein interaction network in *S. cerevisiae*; from Tong et al. (2002)

No regularisation


Test ROC curve AU 0.57562
True positive rate (sensitivity)
False positive rate (1−specificity)
Train ROC curve AU 1
True positive rate (sensitivity)
False positive rate (1−specificity)
AUROC changes
AUROC
Iteration
Train
Test
Train & test cross entropy
Train
Test

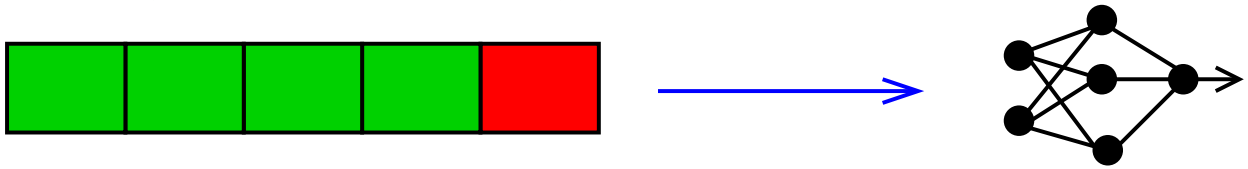
With regularisation


Test ROC curve AU 0.75889
True positive rate (sensitivity)
False positive rate (1−specificity)
Train ROC curve AU 0.76237
True positive rate (sensitivity)
False positive rate (1−specificity)
AUROC changes
AUROC
Iteration
Train
Test
Train & test cross entropy
Train
Test

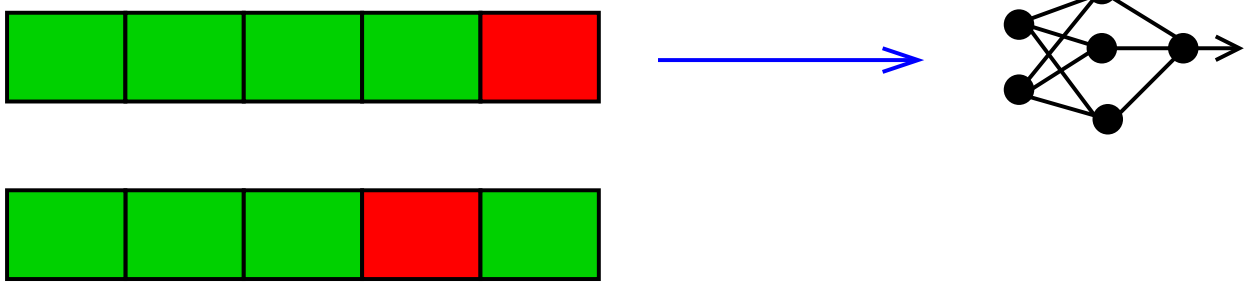
N-fold crossvalidation

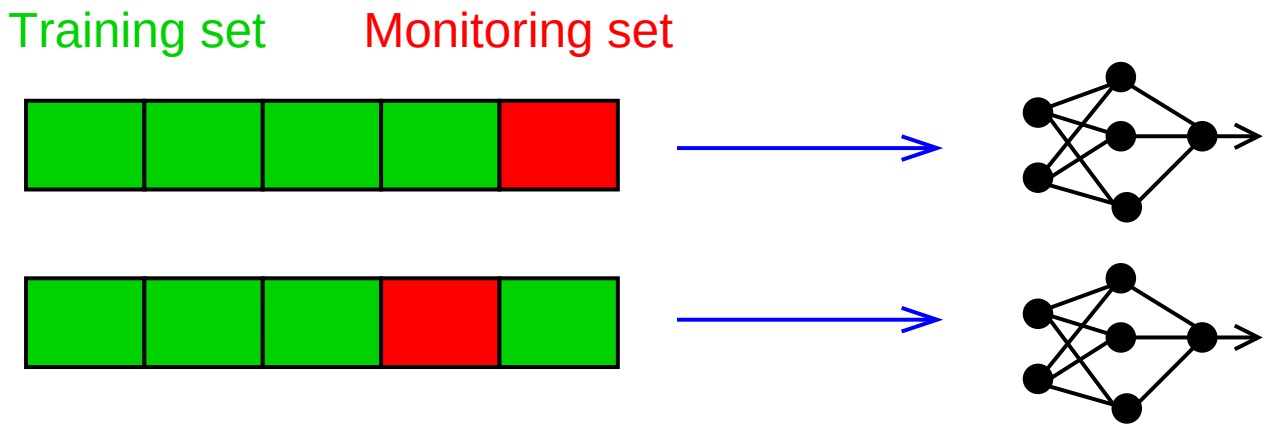
Training set Monitoring set

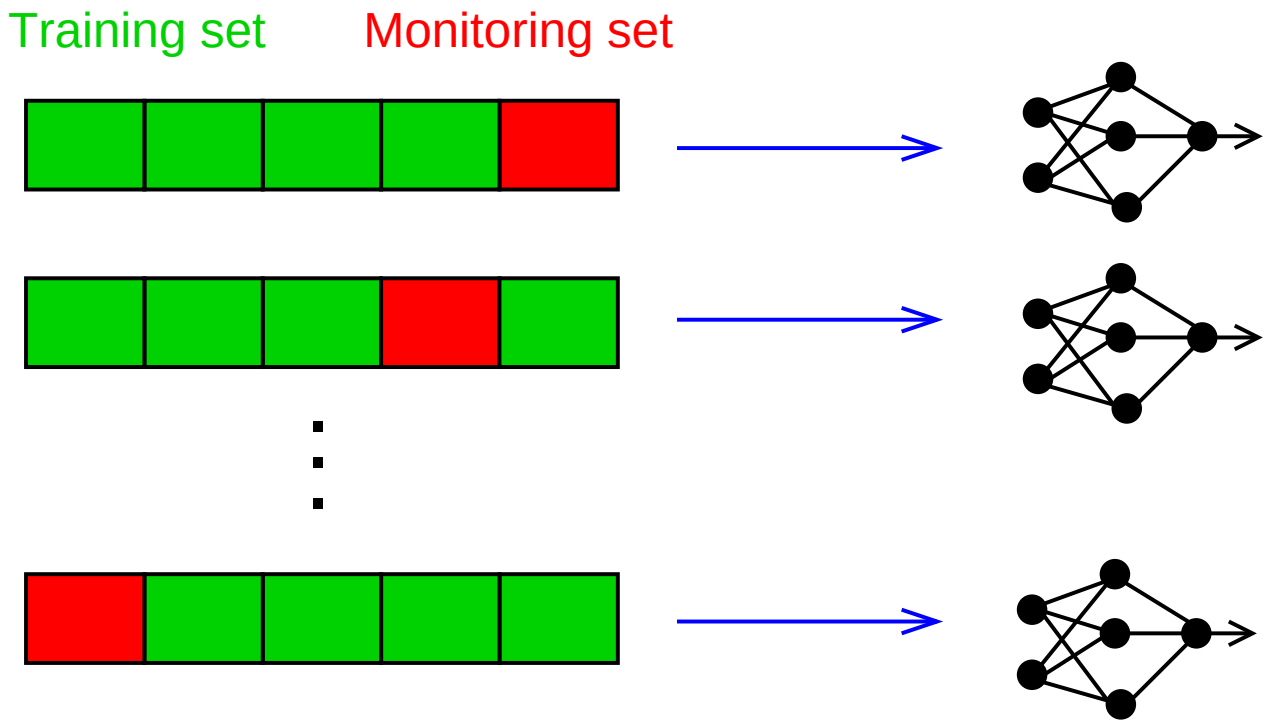
Training set Monitoring set



Training set Monitoring set




Training set
Monitoring set


Training set
Monitoring set

Tong et al. (2002), Science 295, 321-324.

SH3 domain proteins in *Saccharomyces cerevisiae*.

Yeast two-hybrid interaction network

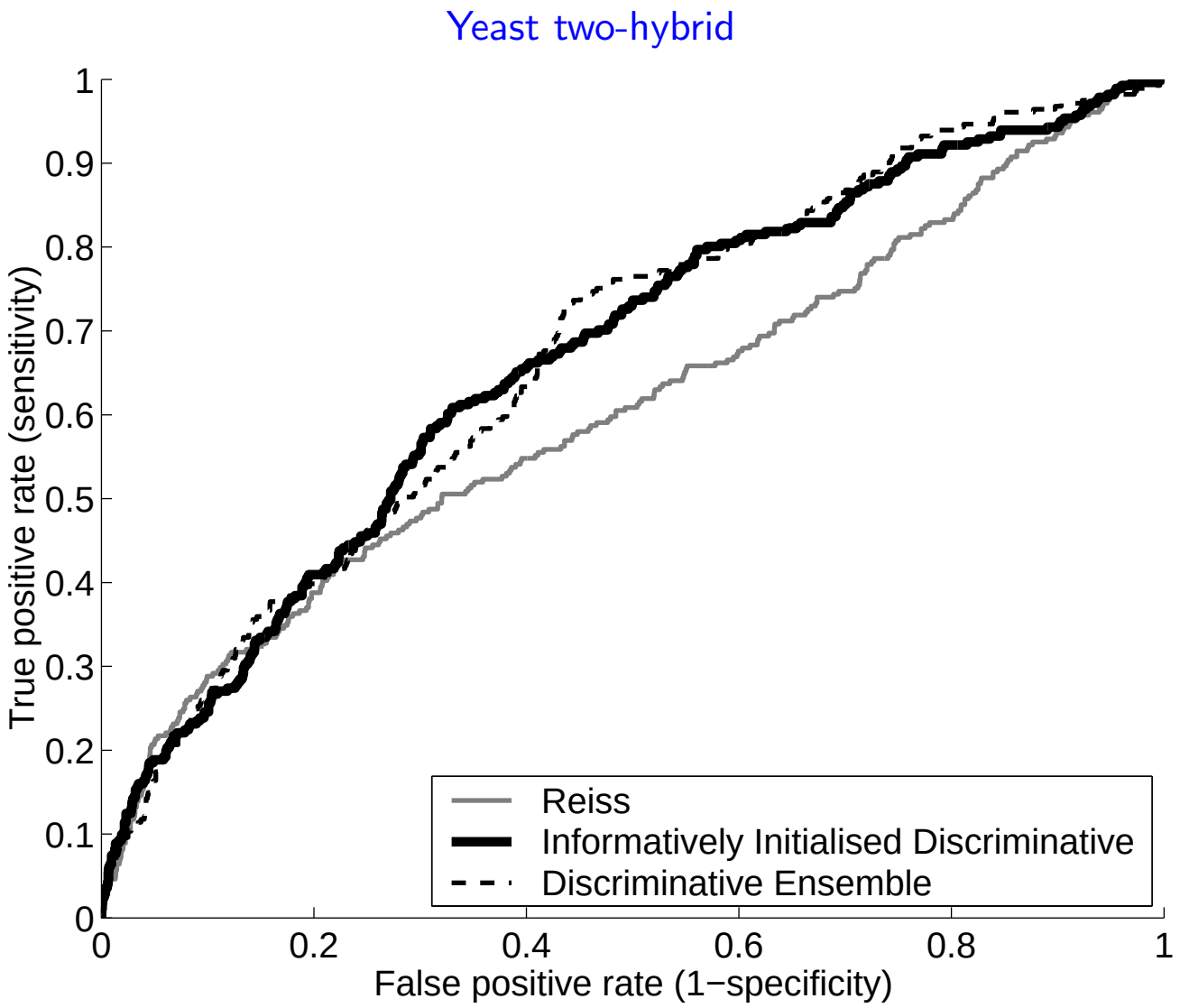
285 interactions between 28 SH3 proteins
and 143 binding peptides

Phage display interaction network

394 interactions between 28 SH3 proteins
and 178 binding peptides

Models compared in our study

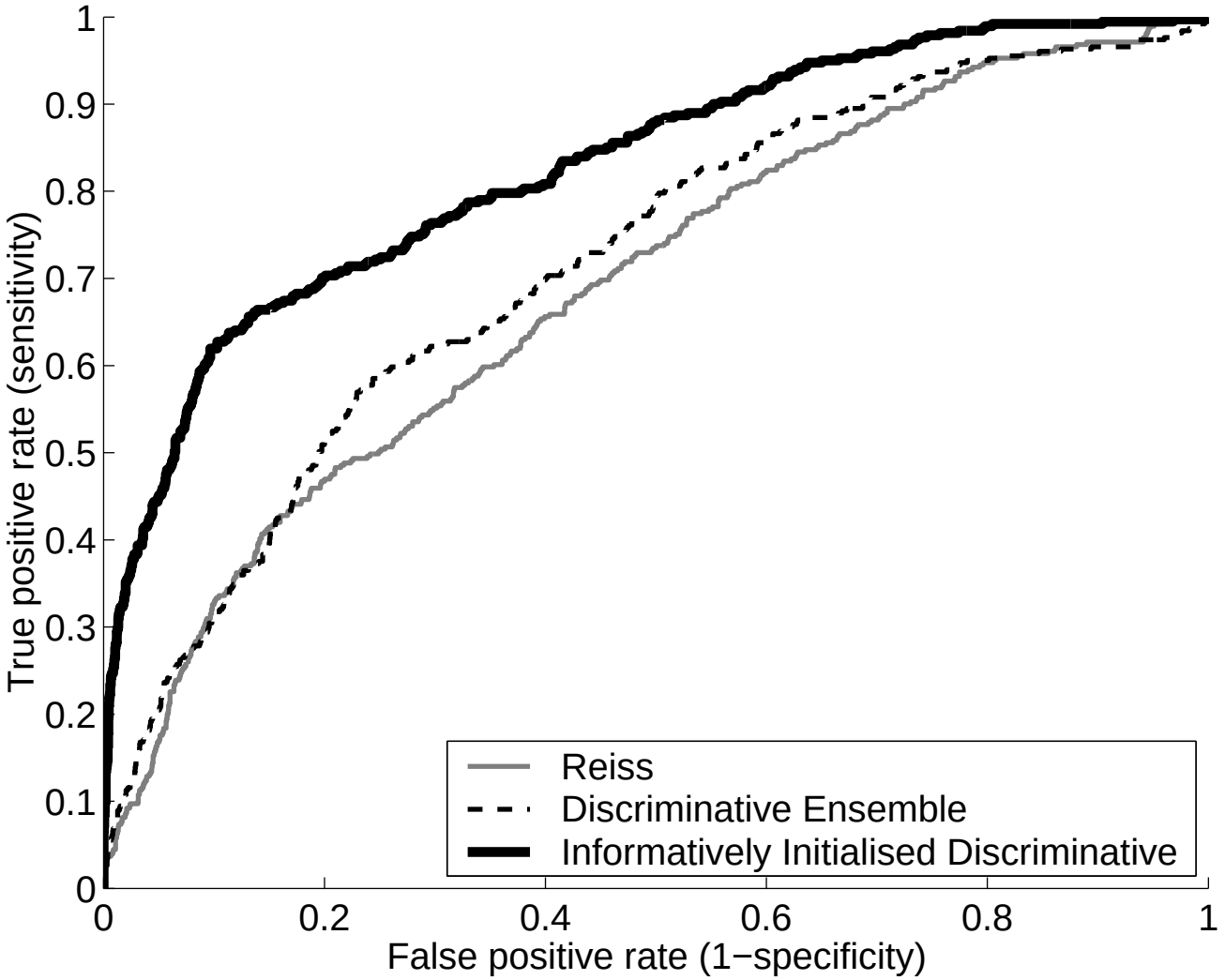
- Generative model of Reiss
- Discriminative model, informative initialisation
- Ensemble of discriminative models, random initialisations


Yeast two-hybrid
True positive rate (sensitivity)
False positive rate (1−specificity)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Reiss
Informatively Initialised Discriminative
Discriminative Ensemble

AUROC scores

Model	→	Generative (Reiss et al.)	Discriminative, informative init	Discriminative, ensemble
Yeast	AUROC	0.61	0.67	0.67

Phage display

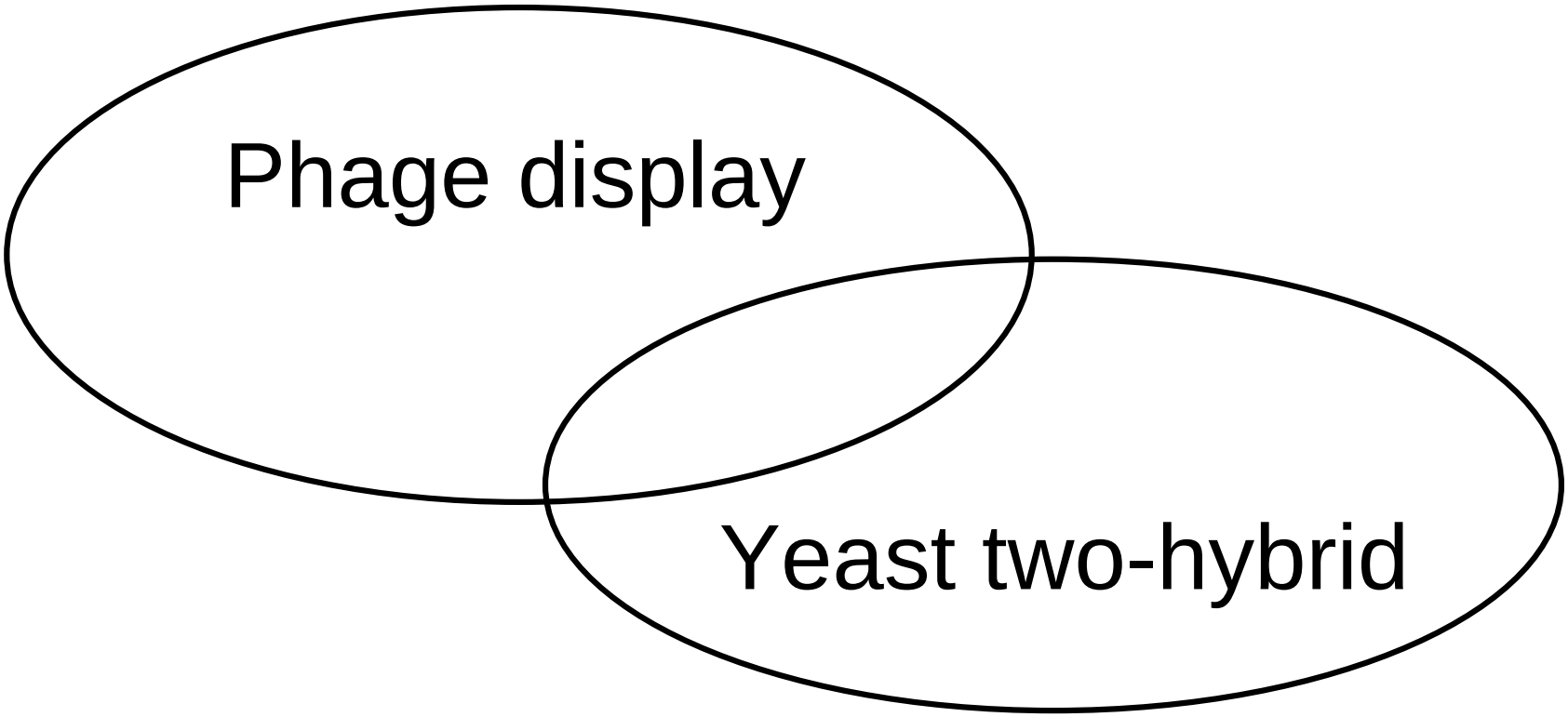


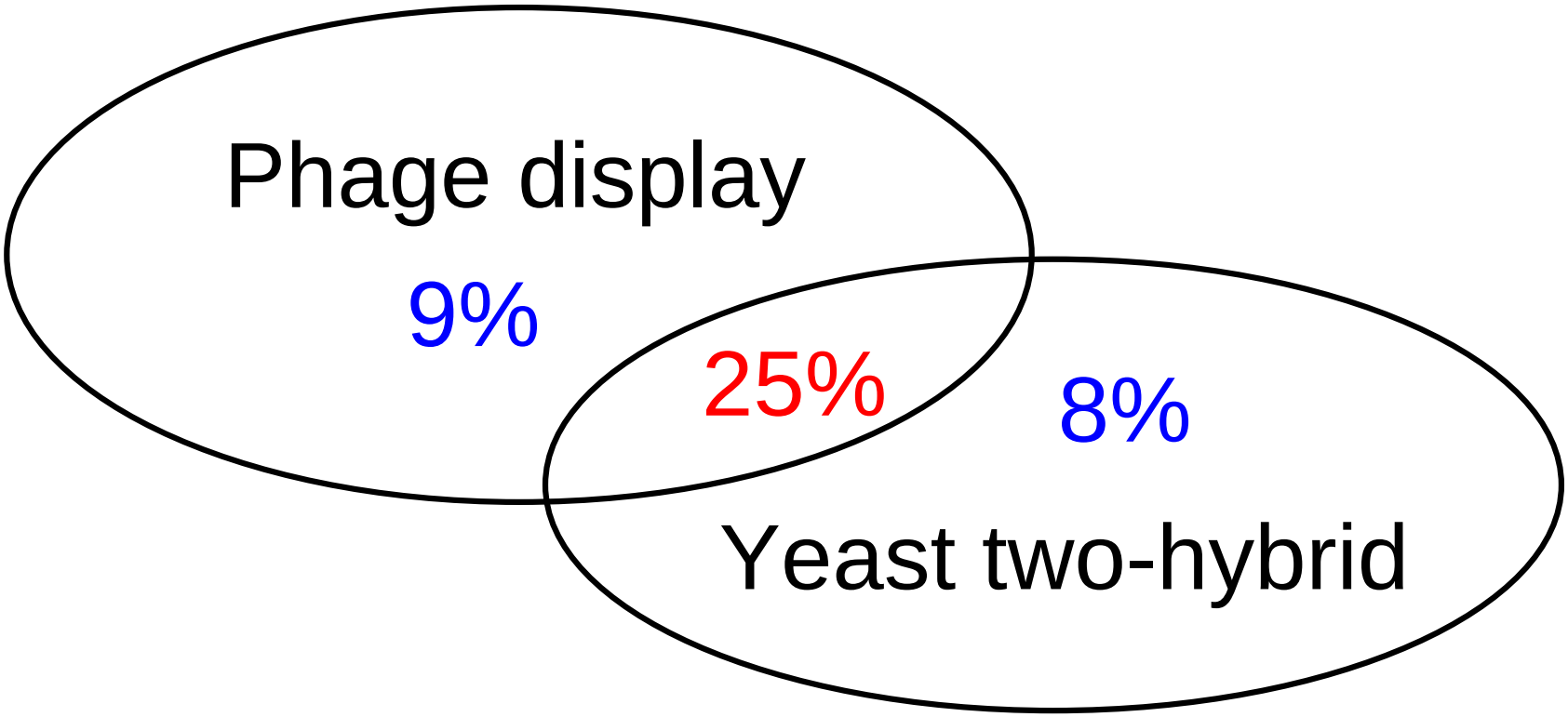
AUROC scores

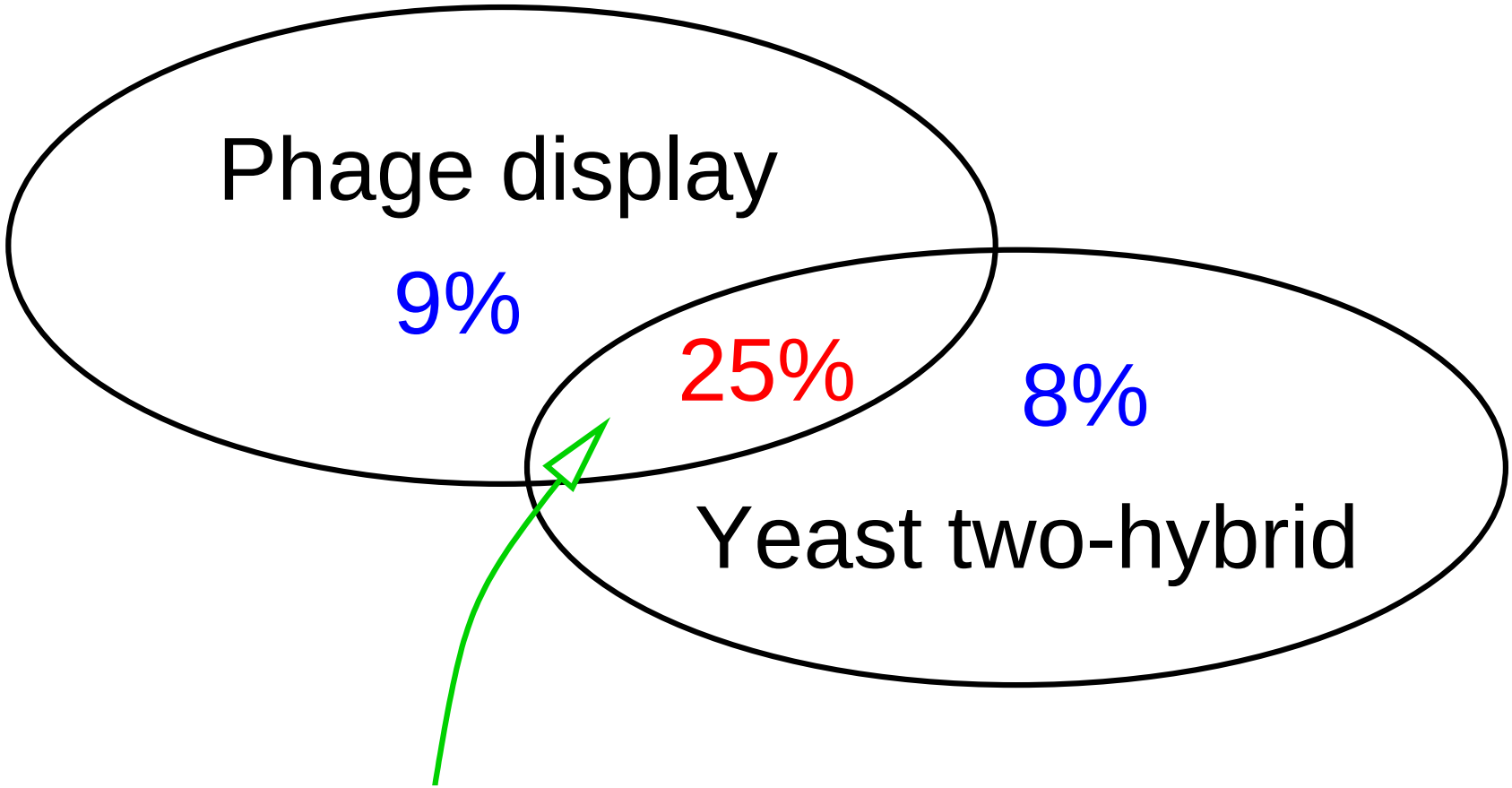
Model	→	Generative (Reiss et al.)	Discriminative, informative init	Discriminative, ensemble
Yeast	AUROC	0.61	0.67	0.67
Phage	AUROC	0.69	0.83	0.71

Biological validation

400 highest scoring interactions


Phage display
Yeast two-hybrid


Phage display
9%
25%
8%
Yeast two-hybrid



Enrichment for higher in silico scores,
filter for noisy high-throughput data.

Summary

- High-throughput interactomic data are noisy
  $\longrightarrow$ Complement data with in silico predictions.
- Generative probabilistic model of Reiss & Schwikowski (2004): Several user-defined tuning parameters
- Discriminative probabilistic model of Segal et al. (2003): Overfitting
- Regularisation with Laplacian prior (Williams 1995).