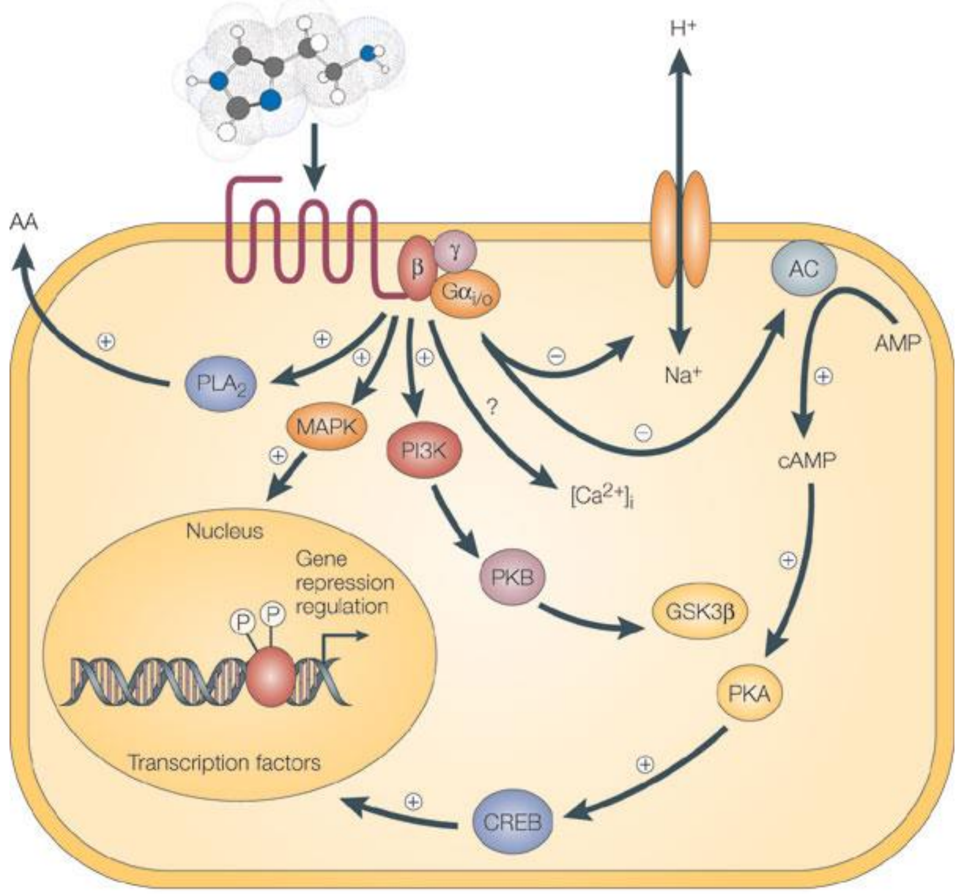
CISC 636 Computational Biology & Bioinformatics
(Fall 2016)

Predicting Protein-Protein Interactions

Background

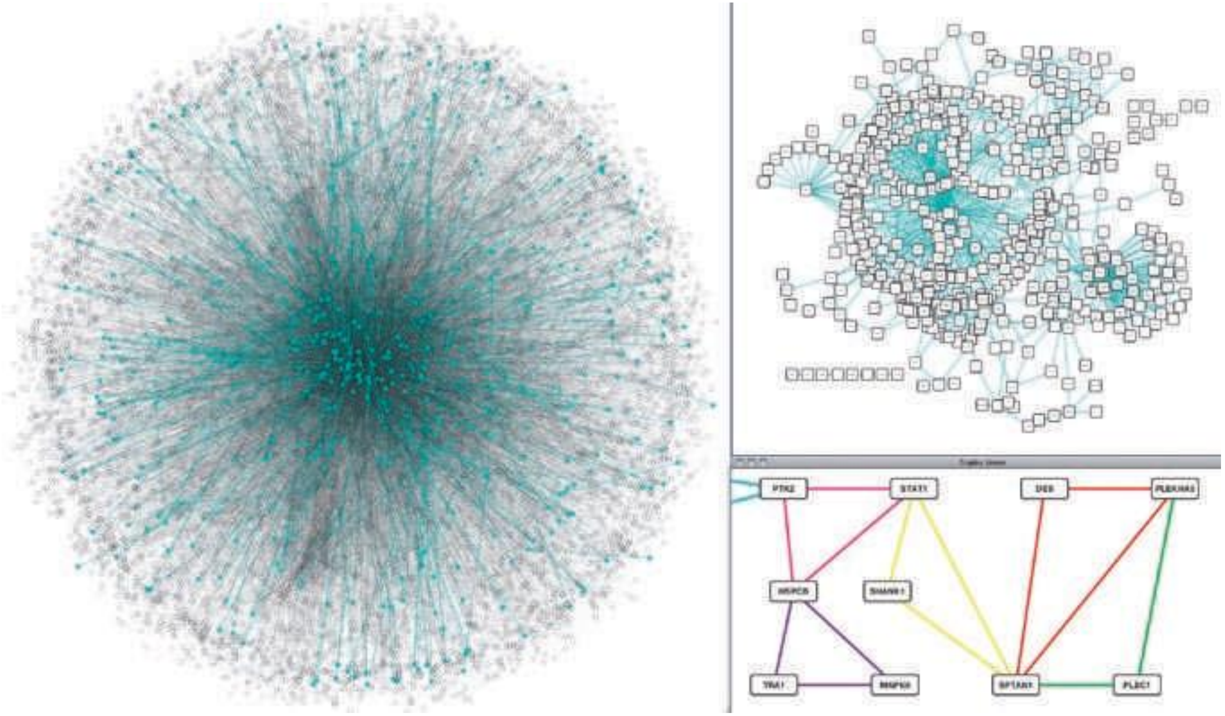
- Proteins do not function as isolated entities.
- Protein-Protein interaction is essential to cellular functions.
- When two proteins *interact*, it can mean:
  - They physically interact
  - They are enzymes catalyzing successive reactions in a pathway
  - One protein regulates expression of the other

Protein-Protein Interaction plays essential roles in cellular processes



Nature Reviews | Drug Discovery

PPI network reconstruction is a central task in systems biology



Given a pair of proteins:

1. **Do they interact? (identify *de novo* pathways, cross talk)**
2. **How do they interact, i.e., which amino acids are involved in interaction? (design mutants to modulate PPI)**

Data sources

- Yeast 2-hybrid system
- 2-D gel + MSMS
- Gene expression (DNA microarray)
- Localization data
- Phylogenetic profiles
- Structural information at binding sites
- Sequences?

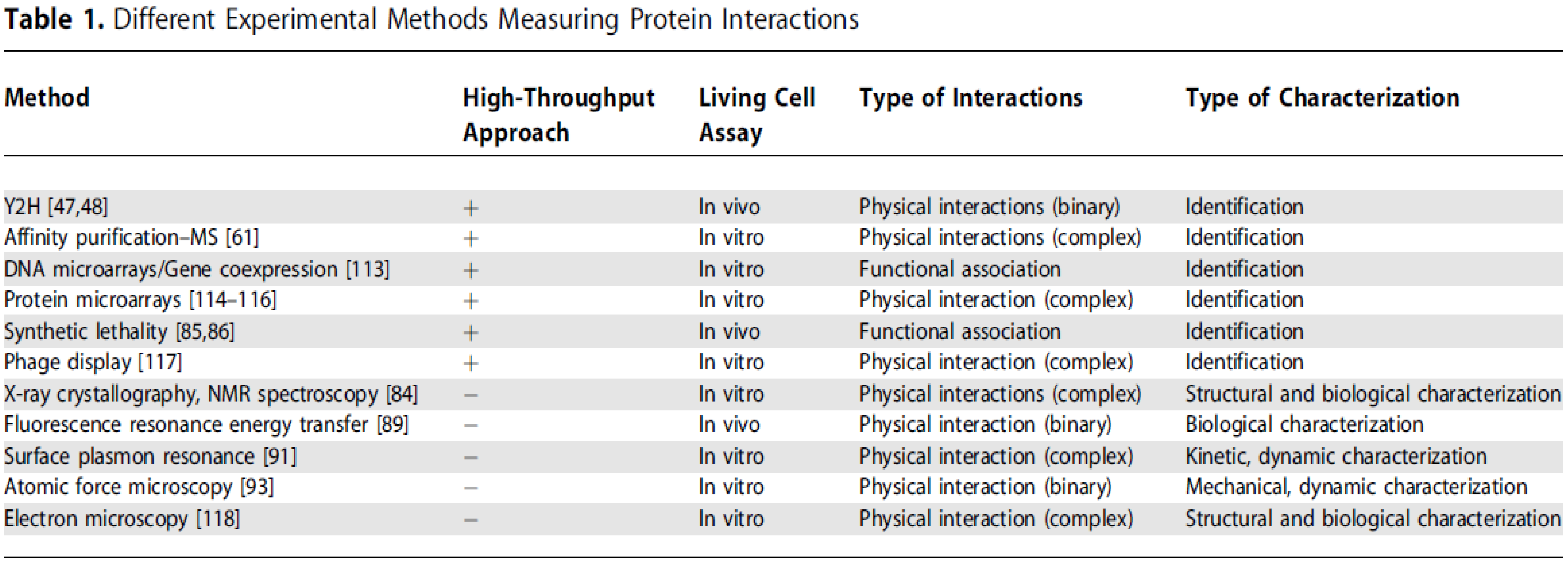


Table 1. Different Experimental Methods Measuring Protein Interactions

Method	High-Throughput Approach	Living Cell Assay	Type of Interactions	Type of Characterization
Y2H [47,48]	+	In vivo	Physical interactions (binary)	Identification
Affinity purification–MS [61]	+	In vitro	Physical interactions (complex)	Identification
DNA microarrays/Gene coexpression [113]	+	In vitro	Functional association	Identification
Protein microarrays [114–116]	+	In vitro	Physical interaction (complex)	Identification
Synthetic lethality [85,86]	+	In vivo	Functional association	Identification
Phage display [117]	+	In vitro	Physical interaction (complex)	Identification
X-ray crystallography, NMR spectroscopy [84]	–	In vitro	Physical interactions (complex)	Structural and biological characterization
Fluorescence resonance energy transfer [89]	–	In vivo	Physical interaction (binary)	Biological characterization
Surface plasmon resonance [91]	–	In vitro	Physical interaction (complex)	Kinetic, dynamic characterization
Atomic force microscopy [93]	–	In vitro	Physical interaction (binary)	Mechanical, dynamic characterization
Electron microscopy [118]	–	In vitro	Physical interaction (complex)	Structural and biological characterization

Shoemaker & Panchenko, 2007 PLoS Computational Biology

NATURE METHODS | VOL.6 NO.1 | JANUARY 2009 | 91

An experimentally derived confidence score for binary protein-protein interactions

Pascal Braun[1,2,8], Murat Tasan[3,8], Matija Dreze[1,2,4,8], Miriam Barrios-Rodiles[5], Irma Lemmens[6], Haiyuan Yu[1,2], Julie M Sahalie[1,2], Ryan R Murray[1,2], Luba Roncari[5], Anne-Sophie de Smet[6], Kavitha Venkatesan[1,2,7], Jean-François Rual[1,2,7], Jean Vandenhaute[4], Michael E Cusick[1,2], Tony Pawson[5], David E Hill[1,2], Jan Tavernier[6], Jeffrey L Wrana[5], Frederick P Roth[1,3] & Marc Vidal[1,2]

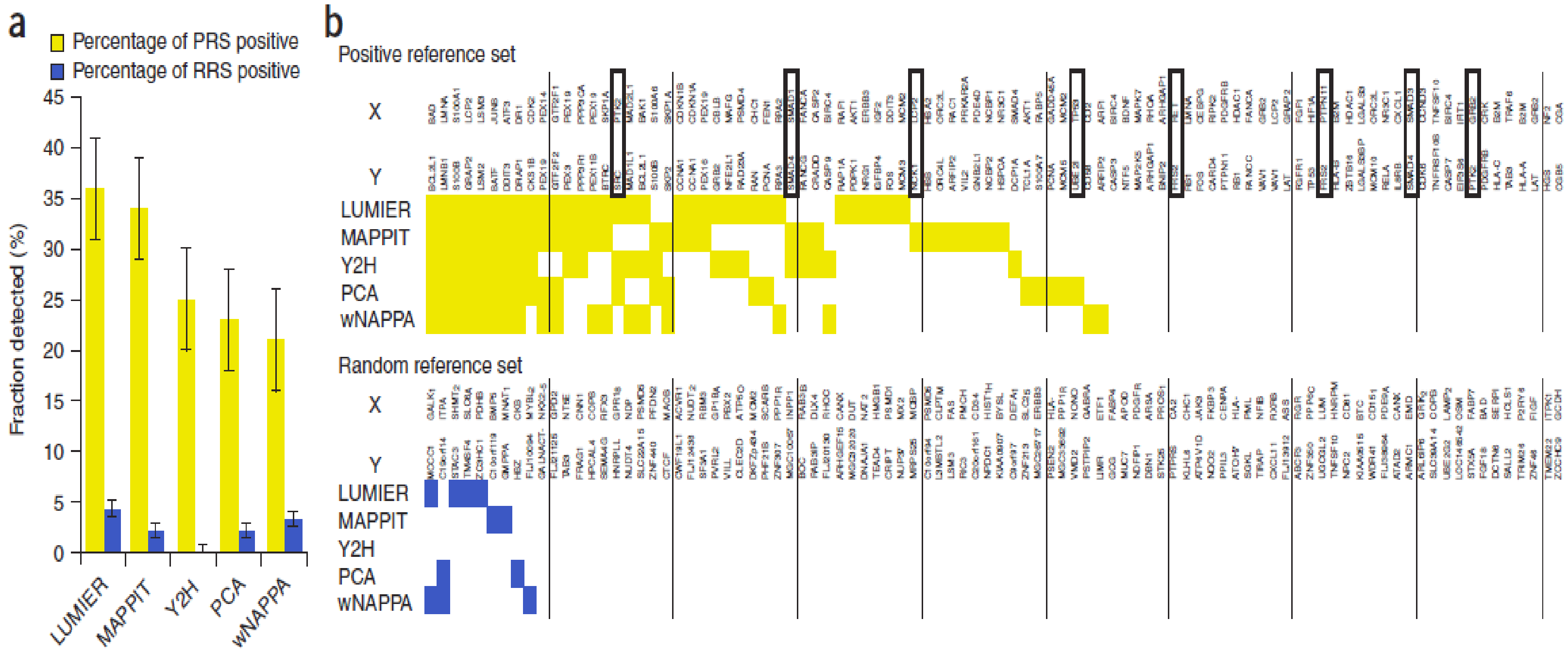


Figure 4 | Performance of assays against positive and random reference sets PRS and RRS. (**a**) Quantification of assay sensitivity and specificity, with s.e.m., using hsPRS-v1 and hsRRS-v1. (**b**) Detection of individual hsPRS-v1 and hsRRS-v1 pairs by the tool kit assays. Top panel: detected hsPRS-v1 pairs are indicated by yellow squares. Bottom: detected hsRRS-v1 pairs are indicated by blue squares. Phosphorylation-dependent interactions are boxed. Thresholds used for the assays can be found in Methods.

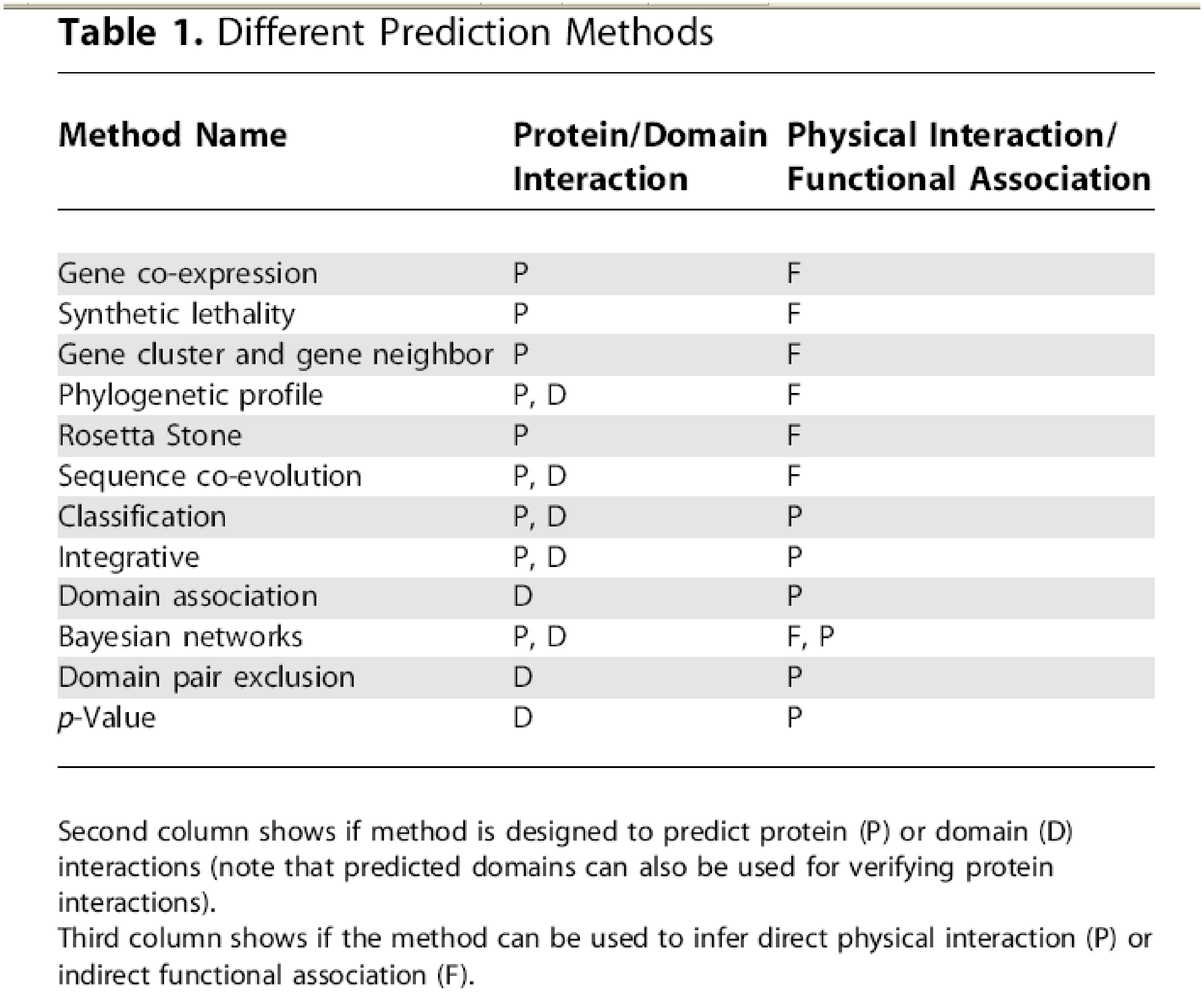


Table 1. Different Prediction Methods

Method Name	Protein/Domain Interaction	Physical Interaction/ Functional Association
Gene co-expression	P	F
Synthetic lethality	P	F
Gene cluster and gene neighbor	P	F
Phylogenetic profile	P, D	F
Rosetta Stone	P	F
Sequence co-evolution	P, D	F
Classification	P, D	P
Integrative	P, D	P
Domain association	D	P
Bayesian networks	P, D	F, P
Domain pair exclusion	D	P
p-Value	D	P

Second column shows if method is designed to predict protein (P) or domain (D) interactions (note that predicted domains can also be used for verifying protein interactions).
Third column shows if the method can be used to infer direct physical interaction (P) or indirect functional association (F).

Shoemaker & Panchenko, 2007 PLoS Computational Biology

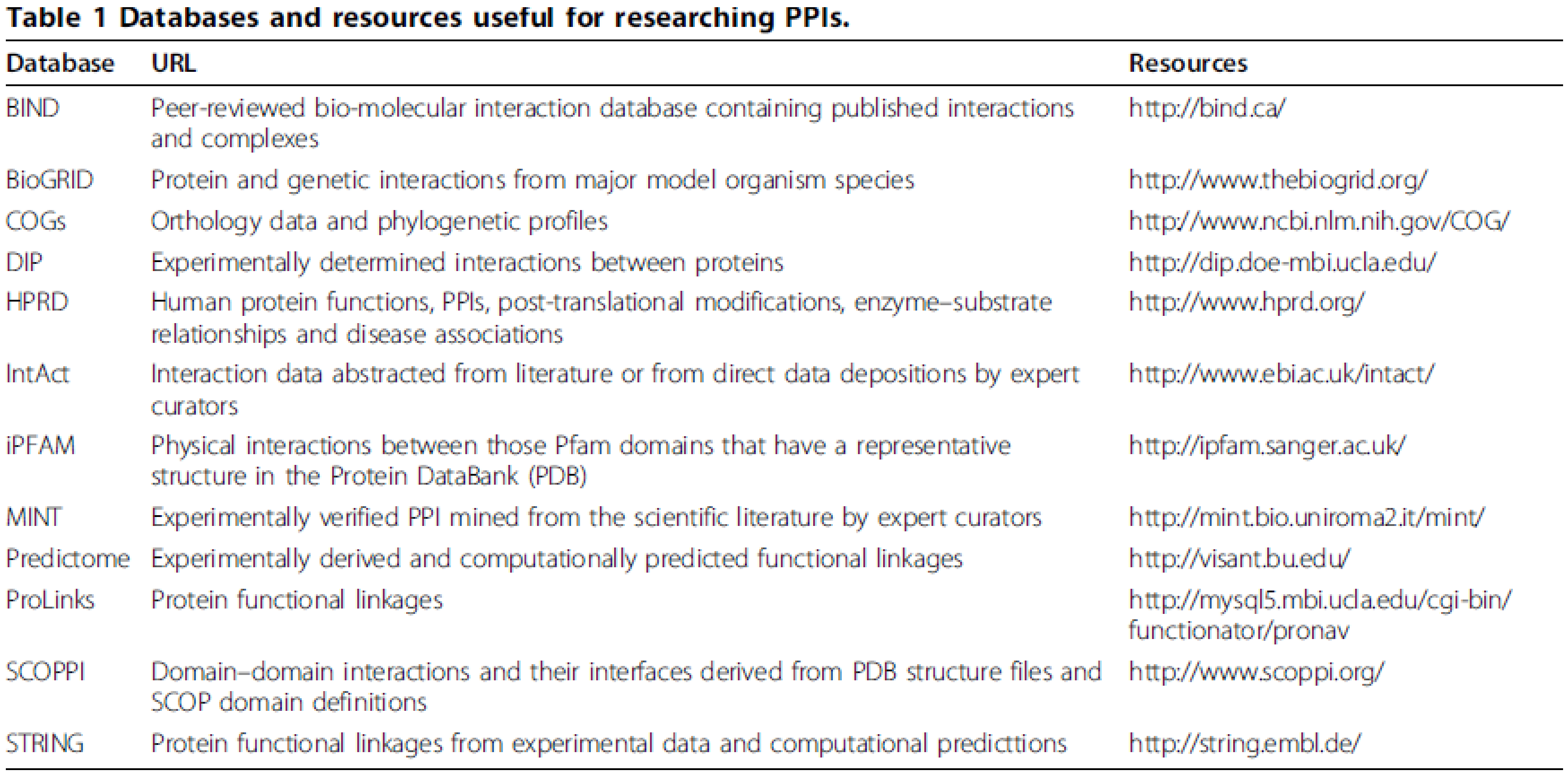
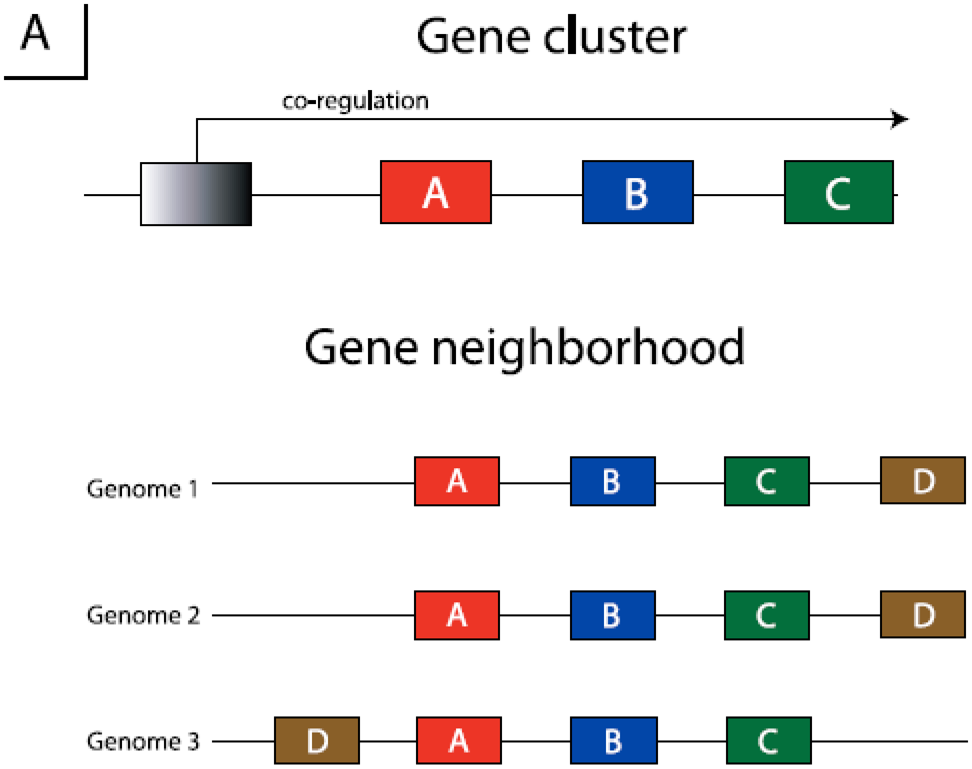


Table 1 Databases and resources useful for researching PPIs.

Database	URL	Resources
BIND	Peer-reviewed bio-molecular interaction database containing published interactions and complexes	http://bind.ca/
BioGRID	Protein and genetic interactions from major model organism species	http://www.thebiogrid.org/
COGs	Orthology data and phylogenetic profiles	http://www.ncbi.nlm.nih.gov/COG/
DIP	Experimentally determined interactions between proteins	http://dip.doe-mbi.ucla.edu/
HPRD	Human protein functions, PPIs, post-translational modifications, enzyme–substrate relationships and disease associations	http://www.hprd.org/
IntAct	Interaction data abstracted from literature or from direct data depositions by expert curators	http://www.ebi.ac.uk/intact/
iPFAM	Physical interactions between those Pfam domains that have a representative structure in the Protein DataBank (PDB)	http://ipfam.sanger.ac.uk/
MINT	Experimentally verified PPI mined from the scientific literature by expert curators	http://mint.bio.uniroma2.it/mint/
Predictome	Experimentally derived and computationally predicted functional linkages	http://visant.bu.edu/
ProLinks	Protein functional linkages	http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav
SCOPPI	Domain–domain interactions and their interfaces derived from PDB structure files and SCOP domain definitions	http://www.scoppi.org/
STRING	Protein functional linkages from experimental data and computational predicttions	http://string.embl.de/

Use of sequence information

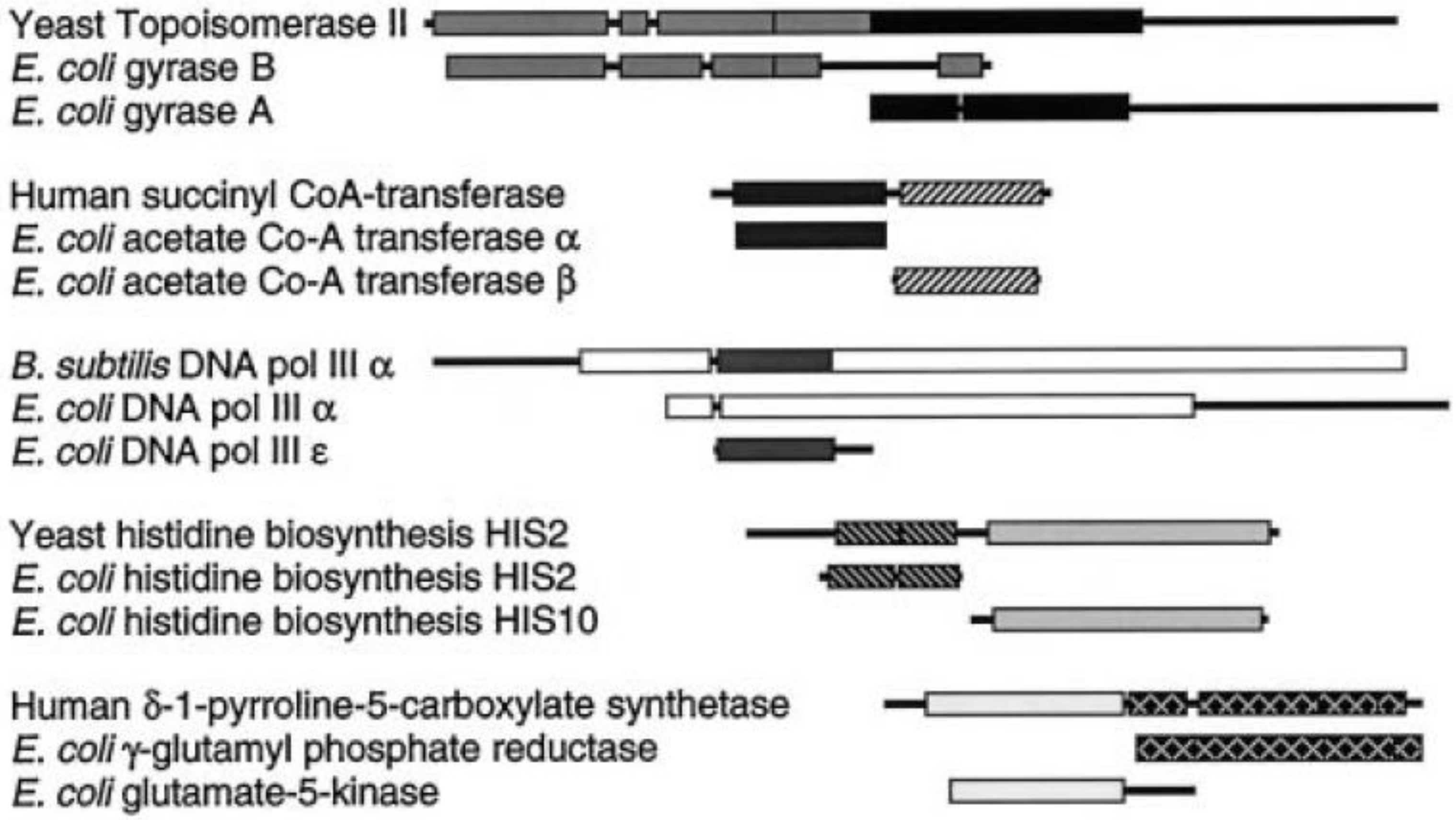
(gene cluster)



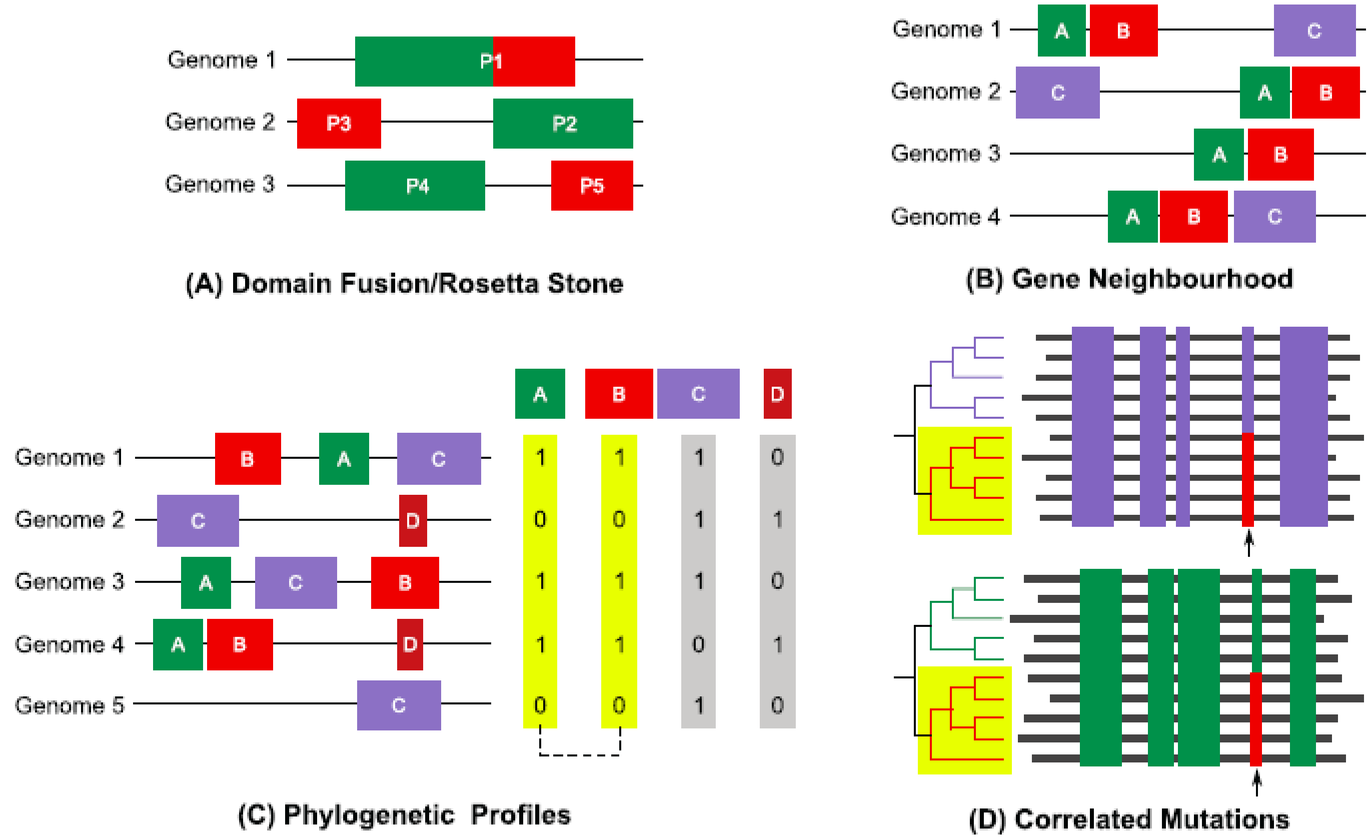
Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18: S329–S336

Use of sequence information

(Rosetta stone, Gene fusion)


Yeast Topoisomerase II
E. coli gyrase B
E. coli gyrase A
Human succinyl CoA-transferase
E. coli acetate Co-A transferase α
E. coli acetate Co-A transferase β
B. subtilis DNA pol III α
E. coli DNA pol III α
E. coli DNA pol III ε
Yeast histidine biosynthesis HIS2
E. coli histidine biosynthesis HIS2
E. coli histidine biosynthesis HIS10
Human δ-1-pyrroline-5-carboxylate synthetase
E. coli γ-glutamyl phosphate reductase
E. coli glutamate-5-kinase

Marcotte et al, Science (1999)



	A	B	C	D
Genome 1	1	1	1	0
Genome 2	0	0	1	1
Genome 3	1	1	1	0
Genome 4	1	1	0	1
Genome 5	0	0	1	0

(A) Domain Fusion/Rosetta Stone

(B) Gene Neighbourhood

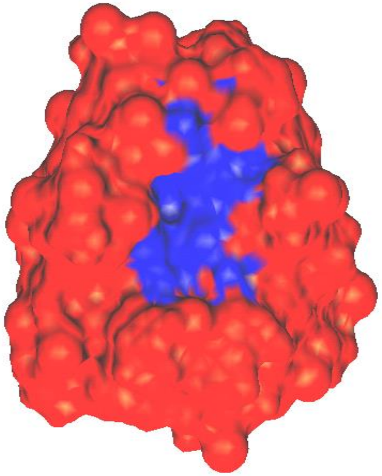
(C) Phylogenetic Profiles

(D) Correlated Mutations

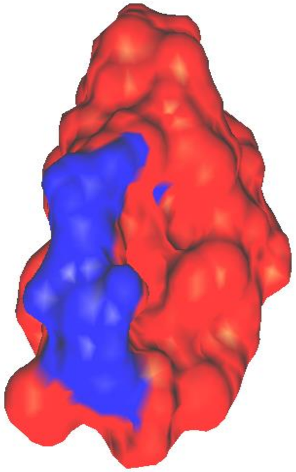
Figure 1 Prediction of functional linkages between proteins, based on different methods. **(A) Method of domain fusion**. The figure shows proteins predicted to interact by the Rosetta stone method (domain fusion). Each protein is shown schematically with boxes representing domains. Proteins P2 and P3 in Genomes 2 and 3 are predicted to interact because their homologues are fused in the first genome. **(B) Gene neighbourhood**. The figure shows four hypothetical genomes, containing one or more of the genes A, B and C. Since the genes A and B are co-localised in multiple genomes (1–4), they are likely to be functionally linked with one another. **(C) Phylogenetic profiles**. The figure shows five hypothetical genomes, each containing one or more of the proteins A, B, C and D. The presence or absence of each protein is indicated by 1 or 0, respectively, in the phylogenetic profiles given on the right. Identical profiles are highlighted — proteins A and B are functionally linked (dotted line), whereas proteins C and D, which have different phylogenetic profiles (shown in grey) are not likely to be functionally linked. **(D) Correlated mutations**. The alignments of two protein families are shown; conserved residues in either alignment are shown in the same colour (blue and green). Correlated mutations in either alignment (coloured red) are indicated by arrow marks. Common sub-trees of the phylogenetic trees are highlighted in yellow. The presence of correlated mutations in each family suggests that the corresponding sites may be involved in mediating interactions between the proteins from each family.

Use of structural information

Structural Compatibility



Trypsin inhibitor



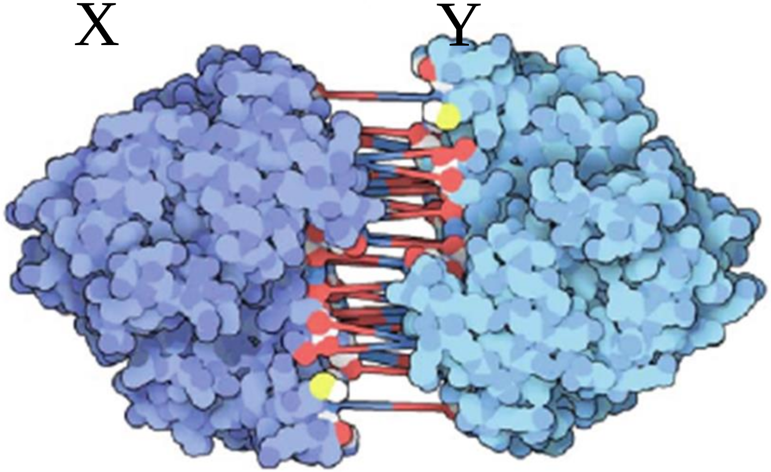
Thermitase

Proteins Interact via Domains

Chemical bonds are formed between amino acids across interface at two interacting proteins.

Protein X

Protein Y

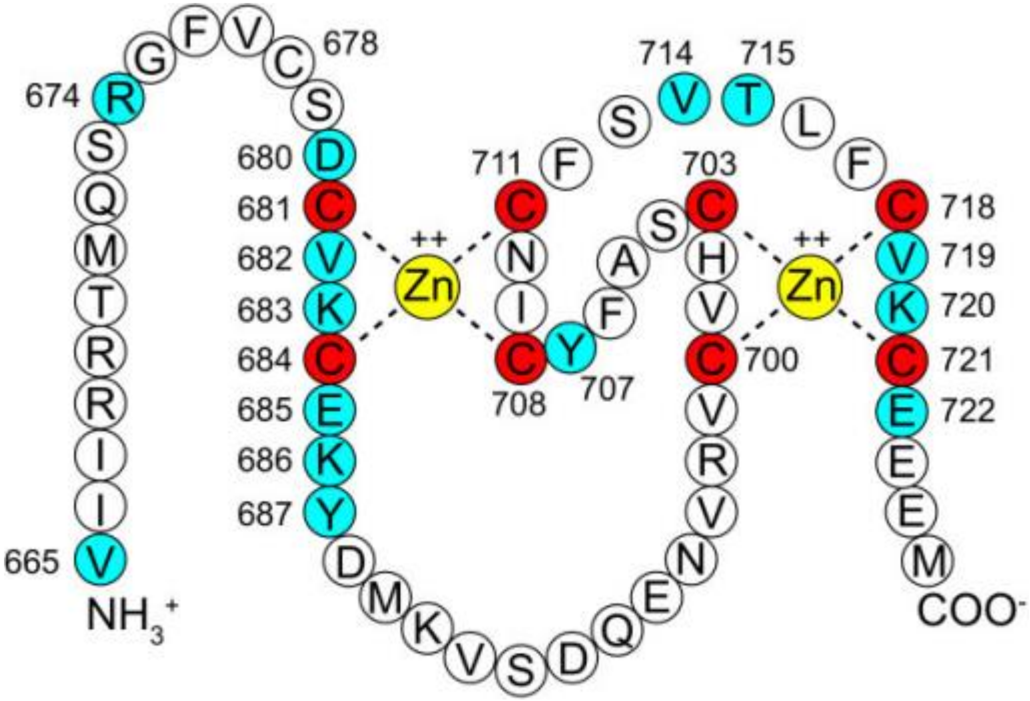


Residues at interface tend to be more conserved due to selection pressure during evolution.

Domain A

Domain B

Not all residues in domain directly participate in interaction



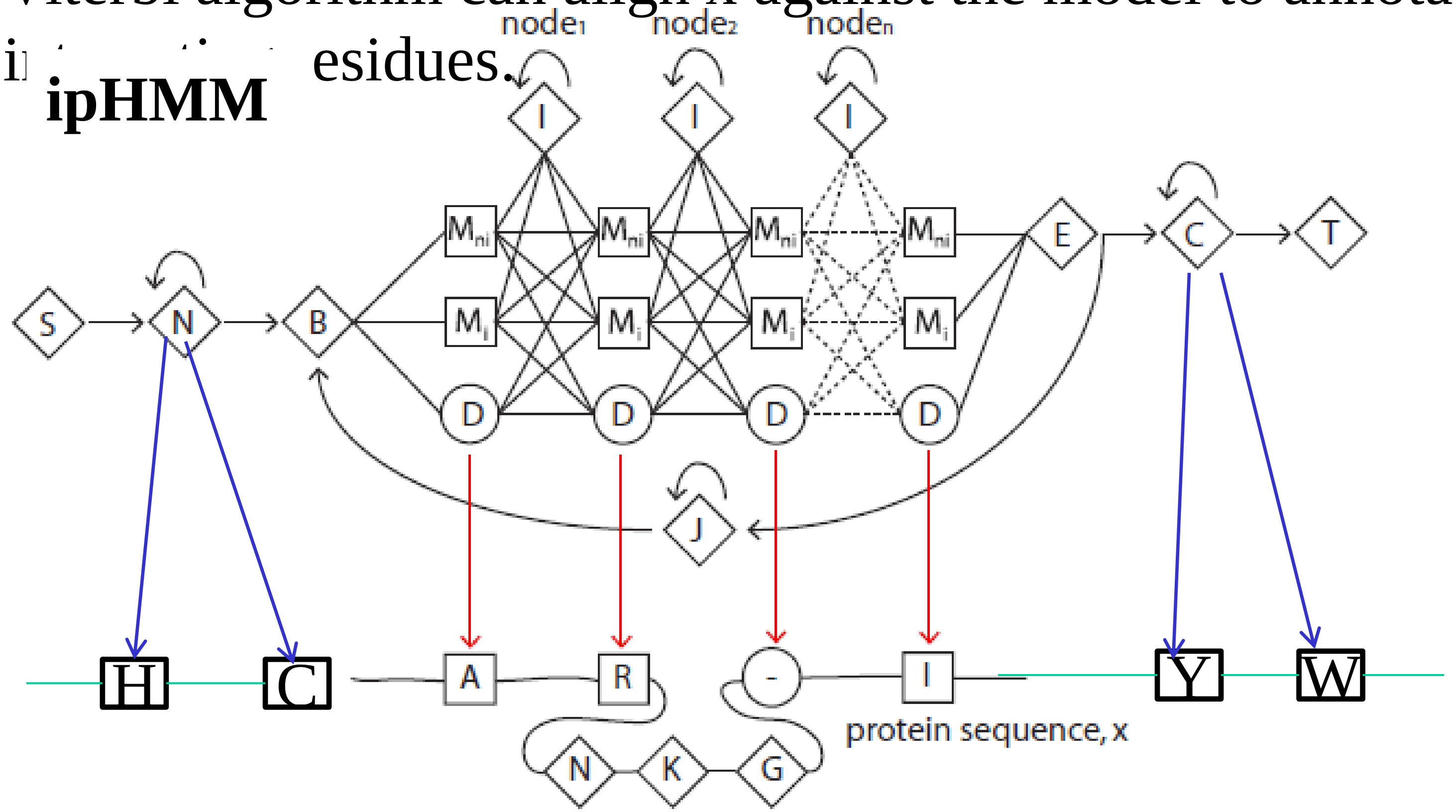
RING domain: cysteines residues in red interact with Zn++ ions to stabilize the ring finger structure; residues colored blue are on the interacting interface.

Wilson et al, BMC Dev. Biol. (2011)

Profile Hidden Markov Models capturing interaction

- P(x|θ) : probability that sequence x contains a domain described by the model θ.
- Viterbi algorithm can align x against the model to annotate i esidues.

ipHMM



Friedrich et al, *Bioinformatics* 2006

From Domain to Domain-Domain Interaction

Given a pair of proteins:

1. **Do they interact?**
2. **How do they interact, i.e., which amino acids are involved in interaction?**

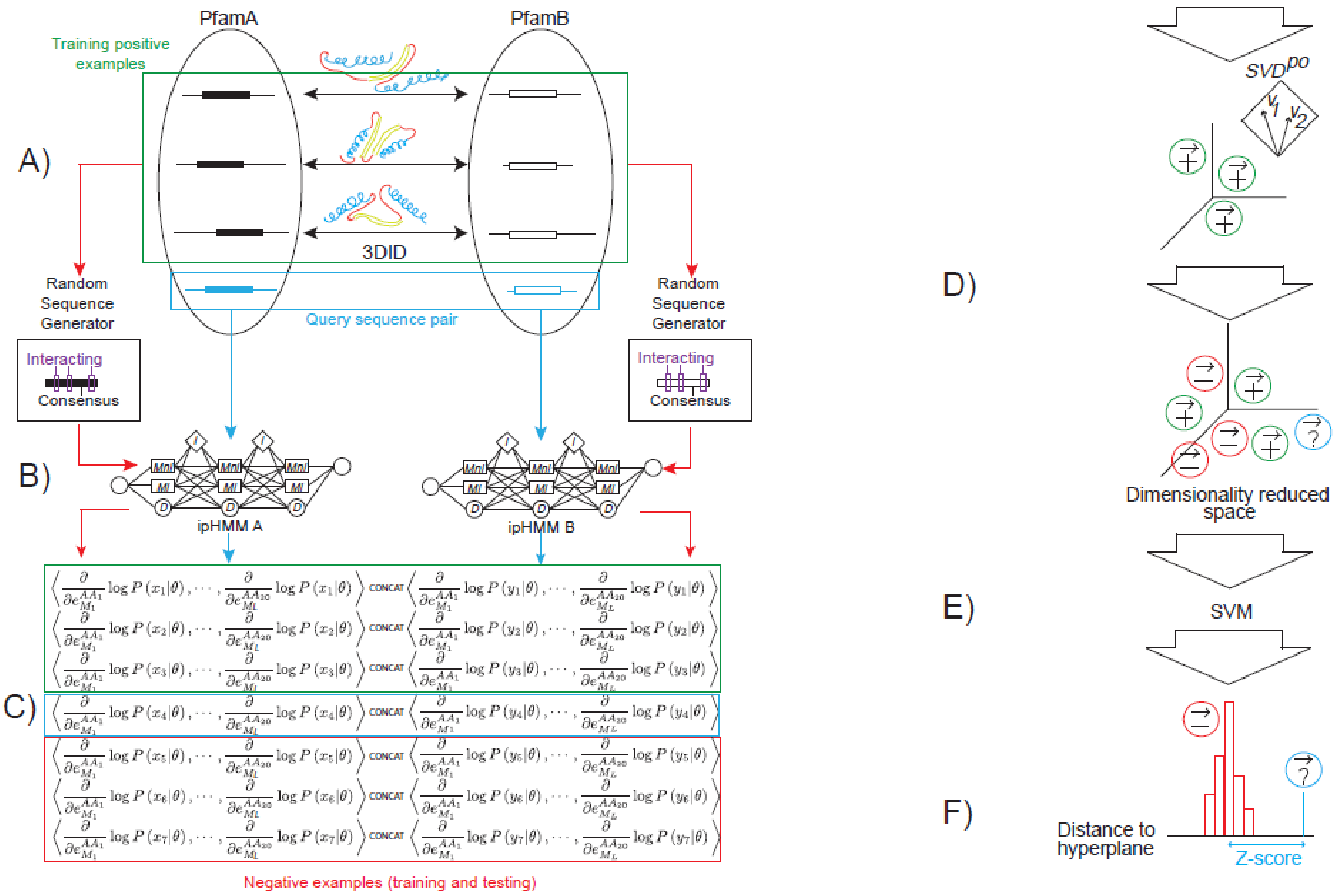
Simple Solution:

i. Query the sequences against domain databases like Pfam.

ii. If Protein X contains domain A, Protein Y contains domain B, and it is known that domain A interacts with domain B, then Protein X interact with Protein Y.

How reliable is the prediction?
If $P(X|A) = 0.8$, $P(Y|B) = 0.8$, probability X and Y interact via domains A and B is $P(X|A)\cdot P(Y|B)$ $= 0.8 \times 0.8 = 0.64$.

From Domain to Domain-Domain Interaction to Protein-Protein Interaction



$$\left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(x_1|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{10}}} \log P(x_1|\theta) \right\rangle \text{CONCAT} \left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(y_1|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(y_1|\theta) \right\rangle$$

$$\left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(x_2|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(x_2|\theta) \right\rangle \text{CONCAT} \left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(y_2|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(y_2|\theta) \right\rangle$$

$$\left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(x_3|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(x_3|\theta) \right\rangle \text{CONCAT} \left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(y_3|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(y_3|\theta) \right\rangle$$

$$\left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(x_4|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(x_4|\theta) \right\rangle \text{CONCAT} \left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(y_4|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(y_4|\theta) \right\rangle$$

$$\left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(x_5|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(x_5|\theta) \right\rangle \text{CONCAT} \left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(y_5|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(y_5|\theta) \right\rangle$$

$$\left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(x_6|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(x_6|\theta) \right\rangle \text{CONCAT} \left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(y_6|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(y_6|\theta) \right\rangle$$

$$\left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(x_7|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(x_7|\theta) \right\rangle \text{CONCAT} \left\langle \frac{\partial}{\partial e_{M_1}^{AA_1}} \log P(y_7|\theta), \cdots, \frac{\partial}{\partial e_{M_L}^{AA_{20}}} \log P(y_7|\theta) \right\rangle$$

Gonzalez & Liao, *BMC Bioinformatics* 2010

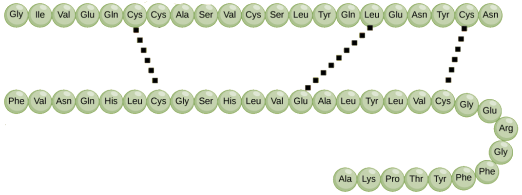
Results: Fisher+SVD+SVM vs *InterPreTS*

Category	Domain A	Domain B	# of distinct complexes	*InterPreTS* (avg. Z-score)	Fisher+SVD+SVM (avg. Z-score)
Signaling	RAS	Rho GAP	5	1.87	30.95
	RAS	Rho GDI	4	2.36	14.64
	G-alpha	Guanylate-cyc	15	3.70	22.95
Cytokines-Receptors	FGF	ig	6	1.01	24.55
	FGF	I-set	10	1.51	21.22
Peptidases-Inhibitors	Kringle	Trypsin	4	1.72	31.53
	Squash	Trypsin	9	1.28	10.23
	Kazal 2	Trypsin	4	0.73	30.64
	Peptidase M10	TIMP	6	0.61	31.35

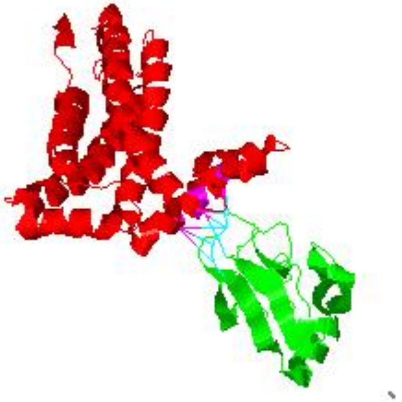
Given a pair of proteins:

1. **Do they interact?**
2. **How do they interact, i.e., which amino acids are involved in interaction?**

What is residue contact matrix?



Gly Ile Val Glu Gln Cys Cys Ala Ser Val Cys Ser Leu Tyr Gln Leu Glu Asn Tyr Cys Asn
Phe Val Asn Gln His Leu Cys Gly Ser His Leu Val Glu Ala Leu Tyr Leu Val Cys Gly Glu Arg Gly Phe
Ala Lys Pro Thr Tyr Phe

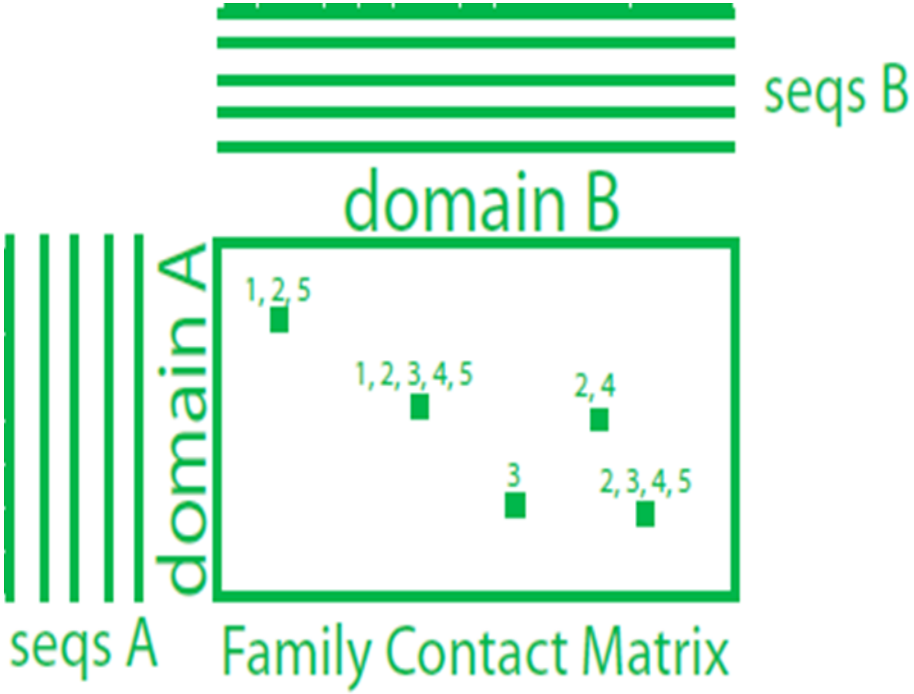


Sequence B

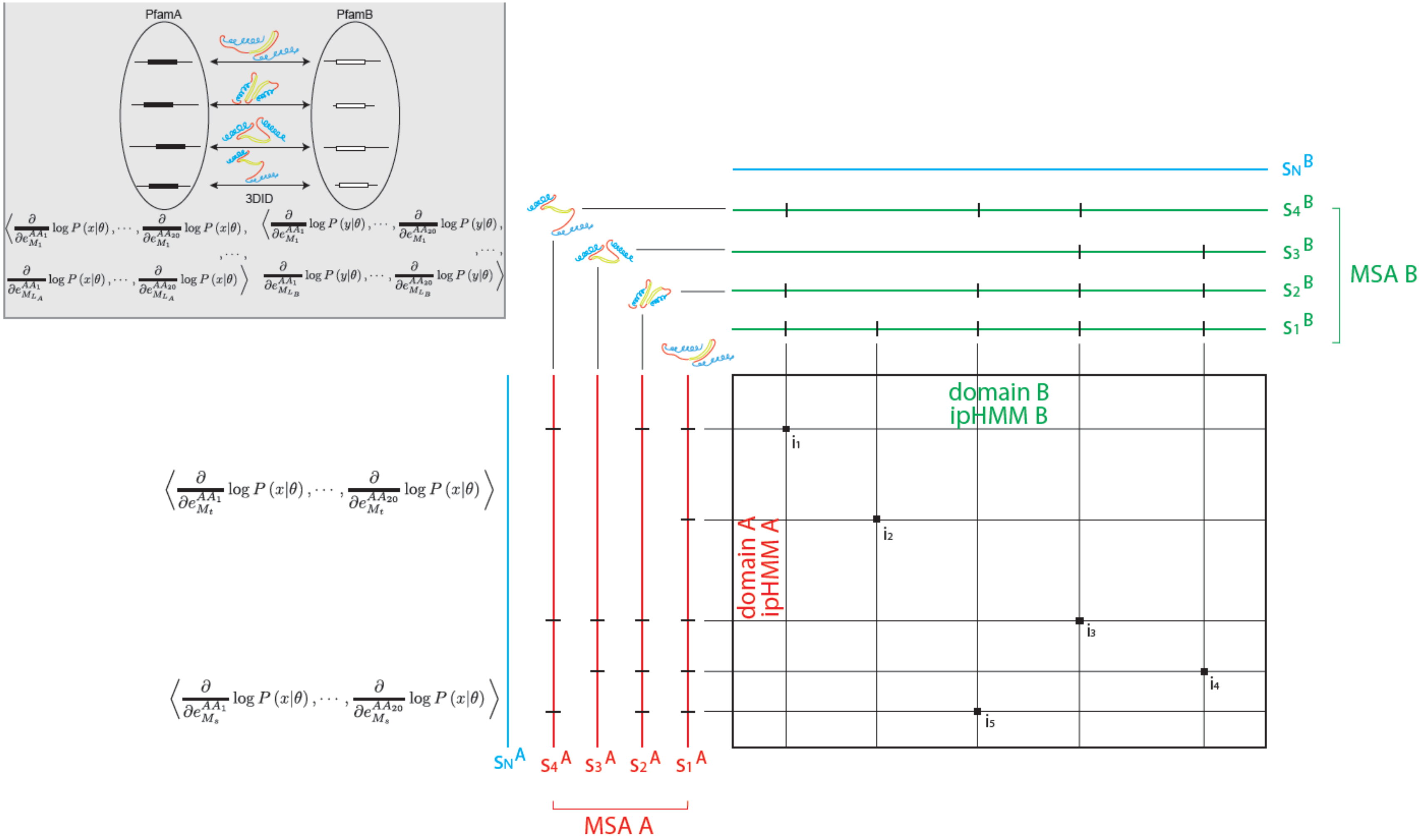
Sequence A

	B_1	B_2	B_3	B_4	B_5	…	B_n
A_1	0	1	0	0	0	…	0
A_2	0	0	0	0	0	…	0
A_3	0	0	0	0	0	…	0
A_5	0	0	0	0	1	…	0
…	…	…	…	…	…	…	…
A_m	0	0	0	0	0	…	0

Predicting residue contact matrix for a pair of interacting proteins

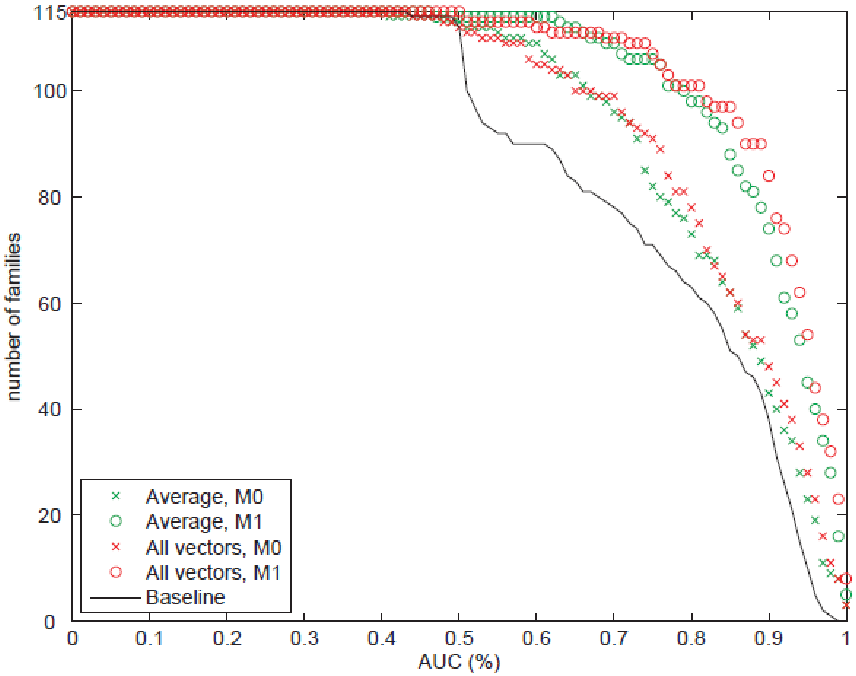


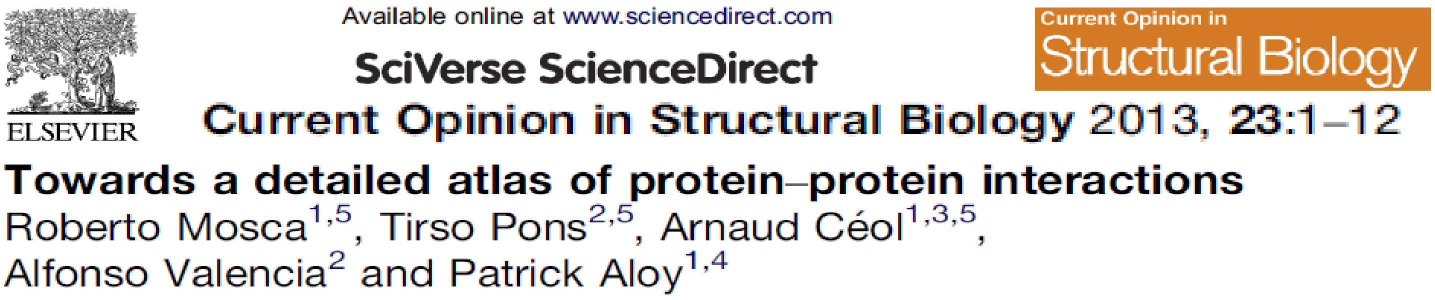
Gonzalez, Liao, Wu, *Bioinformatics* 2013


PfamA
PfamB
3DID
$\left\langle \frac{\partial}{\partial e^{AA_1}_{M_1}} \log P(x|\theta), \cdots, \frac{\partial}{\partial e^{AA_{20}}_{M_1}} \log P(x|\theta), \cdots, \frac{\partial}{\partial e^{AA_1}_{M_{L_A}}} \log P(x|\theta), \cdots, \frac{\partial}{\partial e^{AA_{20}}_{M_{L_A}}} \log P(x|\theta) \right\rangle$
$\left\langle \frac{\partial}{\partial e^{AA_1}_{M_1}} \log P(y|\theta), \cdots, \frac{\partial}{\partial e^{AA_{20}}_{M_1}} \log P(y|\theta), \cdots, \frac{\partial}{\partial e^{AA_1}_{M_{L_B}}} \log P(y|\theta), \cdots, \frac{\partial}{\partial e^{AA_{20}}_{M_{L_B}}} \log P(y|\theta) \right\rangle$
$S_N{}^B$
$S_4{}^B$
$S_3{}^B$
$S_2{}^B$
$S_1{}^B$
MSA B
domain B
ipHMM B
i_1
$\left\langle \frac{\partial}{\partial e^{AA_1}_{M_t}} \log P(x|\theta), \cdots, \frac{\partial}{\partial e^{AA_{20}}_{M_t}} \log P(x|\theta) \right\rangle$
i_2
domain A
ipHMM A
i_3
i_4
$\left\langle \frac{\partial}{\partial e^{AA_1}_{M_s}} \log P(x|\theta), \cdots, \frac{\partial}{\partial e^{AA_{20}}_{M_s}} \log P(x|\theta) \right\rangle$
i_5
$S_N{}^A$
$S_4{}^A$
$S_3{}^A$
$S_2{}^A$
$S_1{}^A$
MSA A

Results: LOO cross-validation, 115 DDIs

	Sensitivity	Specificity	AUC
Average	M_{ni}:84.46% M_i:71.75%	M_{ni}:69.54% M_i:84.81%	M_{ni}:81.01% M_i:89.35%
All vectors	M_{ni}:84.10% M_i:59.90%	M_{ni}:66.53% M_i:82.62%	M_{ni}:83.33% M_i:91.20%
Baseline	56.92%	78.22%	78.16%





Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Structural Biology

ELSEVIER

Current Opinion in Structural Biology 2013, 23:1–12

Towards a detailed atlas of protein–protein interactions

Roberto Mosca[1,5], Tirso Pons[2,5], Arnaud Céol[1,3,5],
Alfonso Valencia[2] and Patrick Aloy[1,4]

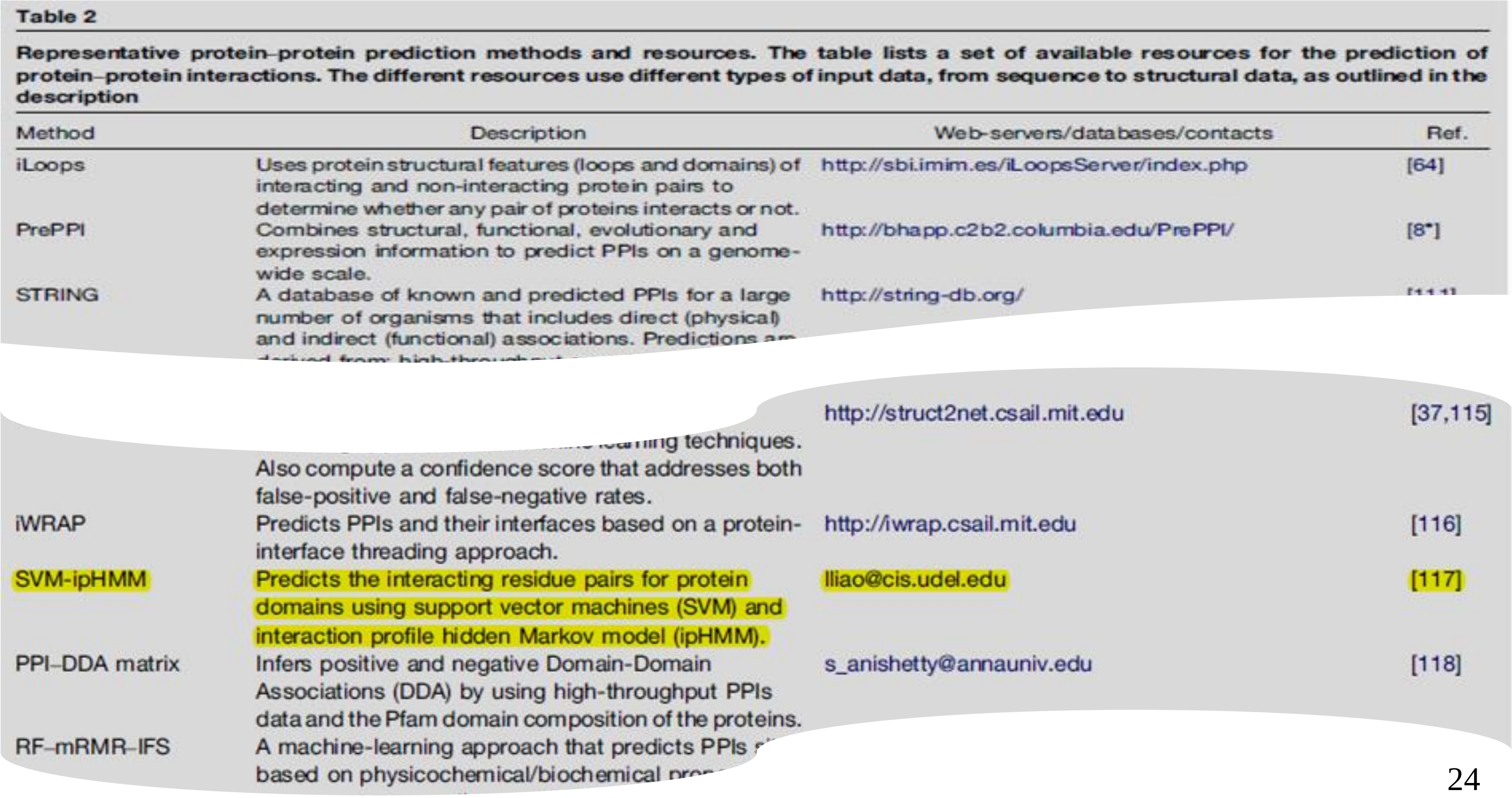
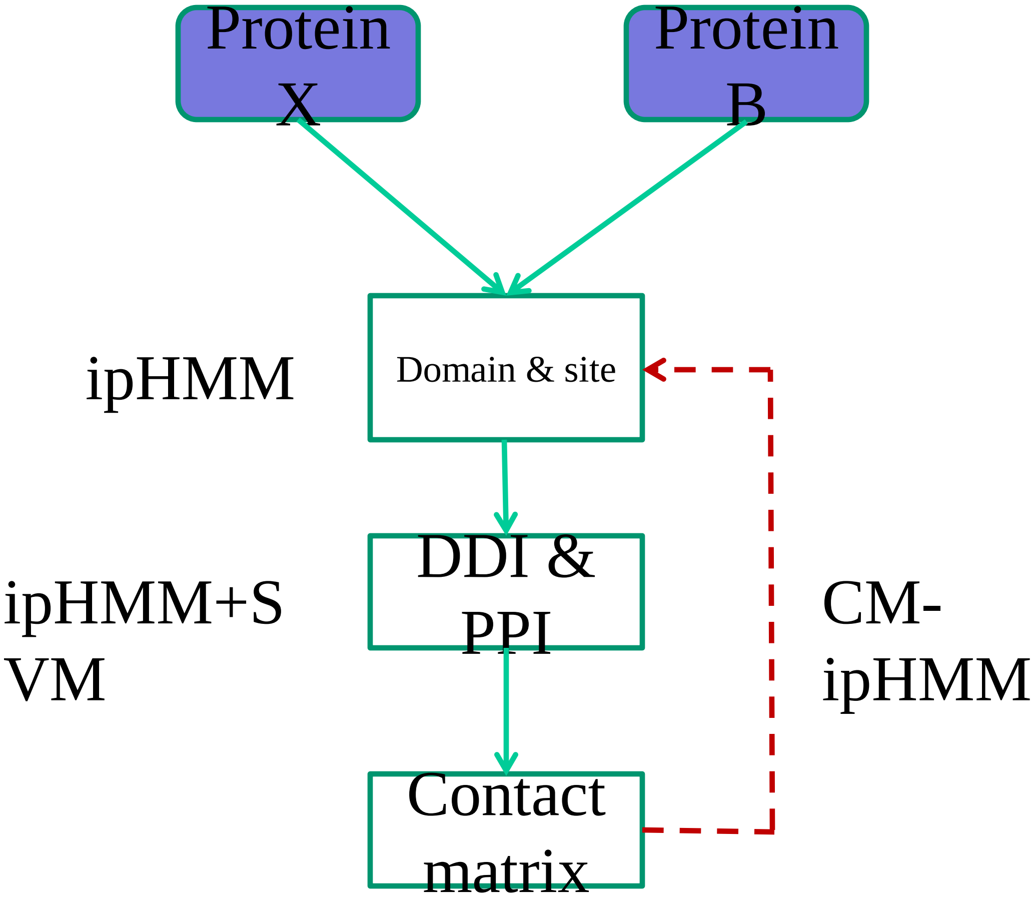


Table 2

Representative protein–protein prediction methods and resources. The table lists a set of available resources for the prediction of protein–protein interactions. The different resources use different types of input data, from sequence to structural data, as outlined in the description

Method	Description	Web-servers/databases/contacts	Ref.
iLoops	Uses protein structural features (loops and domains) of interacting and non-interacting protein pairs to determine whether any pair of proteins interacts or not.	http://sbi.imim.es/iLoopsServer/index.php	[64]
PrePPI	Combines structural, functional, evolutionary and expression information to predict PPIs on a genome-wide scale.	http://bhapp.c2b2.columbia.edu/PrePPI/	[8*]
STRING	A database of known and predicted PPIs for a large number of organisms that includes direct (physical) and indirect (functional) associations. Predictions are	http://string-db.org/	[111]
	learning techniques. Also compute a confidence score that addresses both false-positive and false-negative rates.	http://struct2net.csail.mit.edu	[37,115]
iWRAP	Predicts PPIs and their interfaces based on a protein-interface threading approach.	http://iwrap.csail.mit.edu	[116]
SVM-ipHMM	Predicts the interacting residue pairs for protein domains using support vector machines (SVM) and interaction profile hidden Markov model (ipHMM).	lliao@cis.udel.edu	[117]
PPI–DDA matrix	Infers positive and negative Domain-Domain Associations (DDA) by using high-throughput PPIs data and the Pfam domain composition of the proteins.	s_anishetty@annauniv.edu	[118]
RF–mRMR–IFS	A machine-learning approach that predicts PPIs based on physicochemical/biochemical		

Knowledge Leverage and Integration for Better Learning



Protein X

Protein B

ipHMM

Domain & site

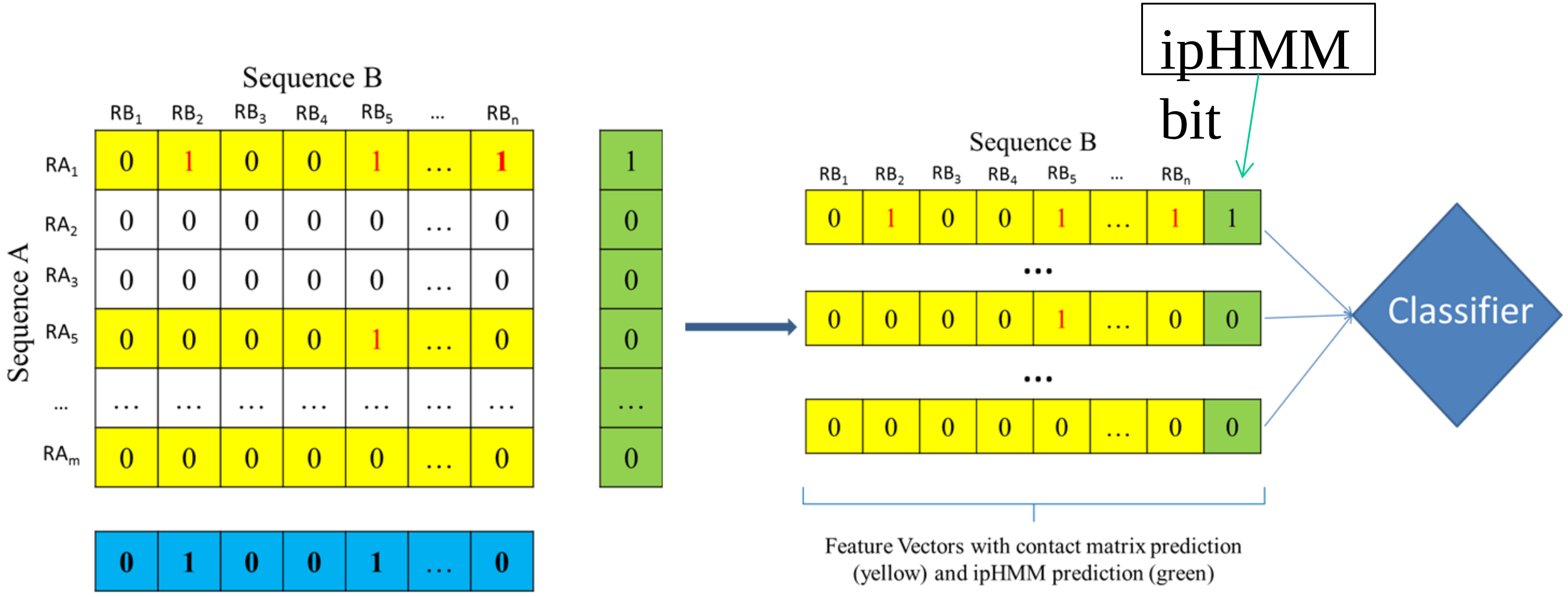
ipHMM+SVM

DDI & PPI

CM-ipHMM

Contact matrix

Method:



Sequence A	RB_1	RB_2	RB_3	RB_4	RB_5	...	RB_n	
RA_1	0	1	0	0	1	…	1	1
RA_2	0	0	0	0	0	…	0	0
RA_3	0	0	0	0	0	…	0	0
RA_5	0	0	0	0	1	…	0	0
...	…	…	…	…	…	…	…	…
RA_m	0	0	0	0	0	…	0	0
	0	1	0	0	1	…	0	

RB_1	RB_2	RB_3	RB_4	RB_5	...	RB_n	
0	1	0	0	1	…	1	1
•••							
0	0	0	0	1	…	0	0
•••							
0	0	0	0	0	…	0	0

Feature Vectors with contact matrix prediction (yellow) and ipHMM prediction (green)

Integrated machine learning classifier with contact matrix prediction and ipHMM site prediction.

Classifier: Logistic Regression

$$\Pr(Y=1|X_1,...,X_k)=\frac{1}{1+\exp[-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k)]}$$

Results

The data set contains 72 DDI families collected from 3DID. Each has 10 ~ 20 member sequences, with domain length < 150 residues.

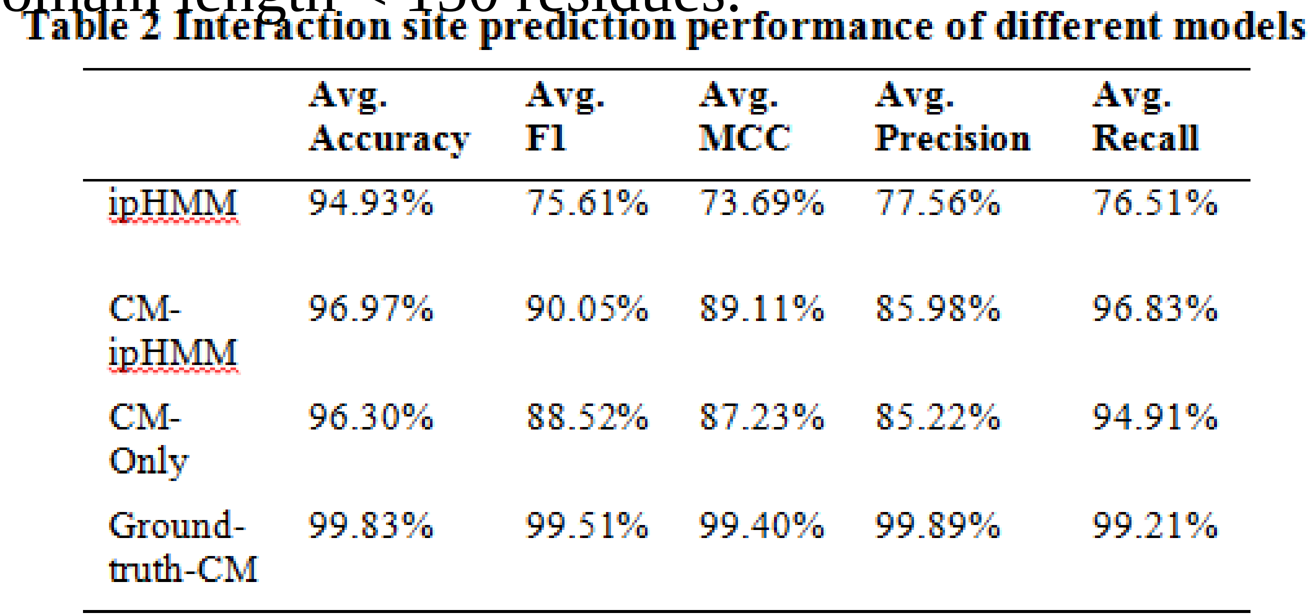
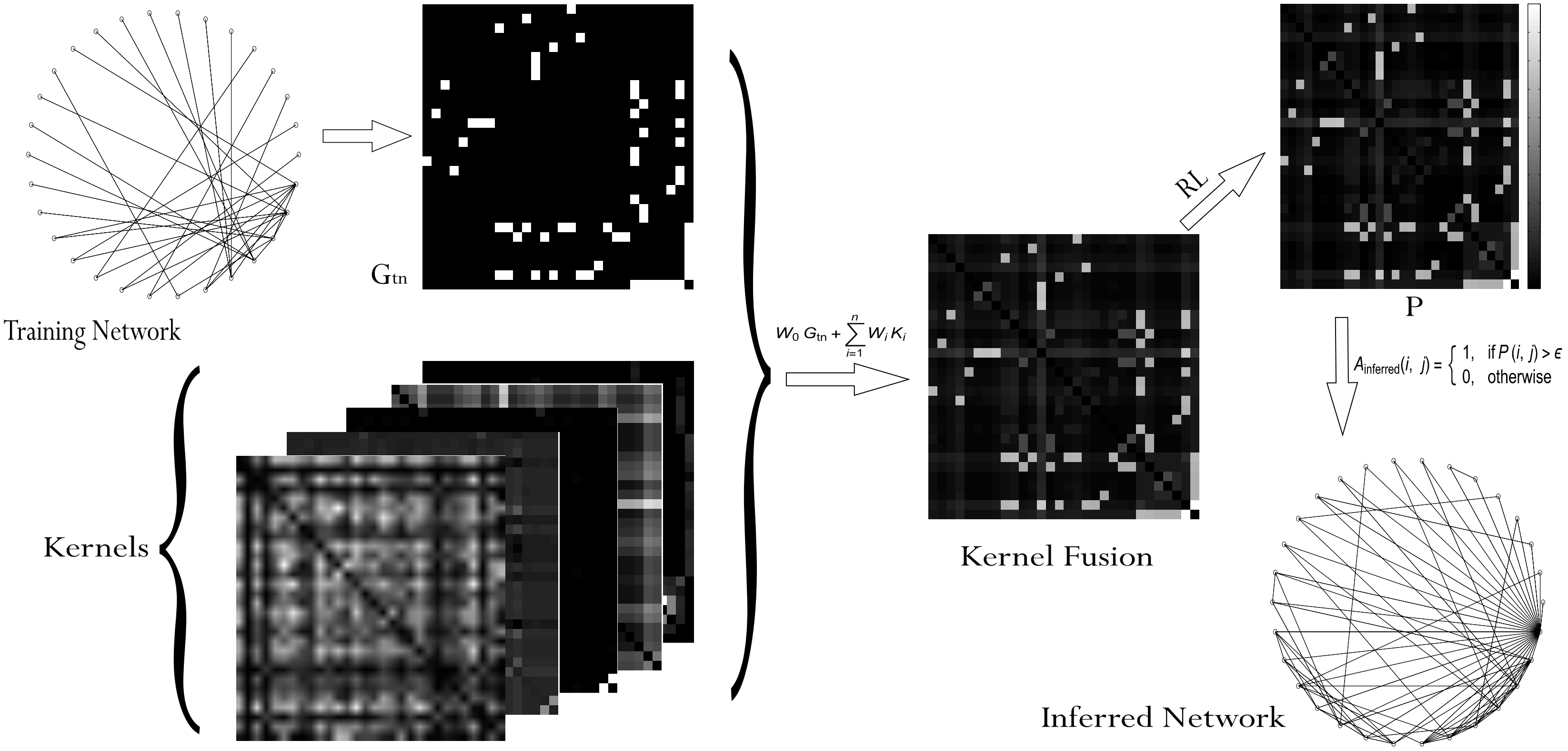


Table 2 Interaction site prediction performance of different models

	Avg. Accuracy	Avg. F1	Avg. MCC	Avg. Precision	Avg. Recall
ipHMM	94.93%	75.61%	73.69%	77.56%	76.51%
CM-ipHMM	96.97%	90.05%	89.11%	85.98%	96.83%
CM-Only	96.30%	88.52%	87.23%	85.22%	94.91%
Ground-truth-CM	99.83%	99.51%	99.40%	99.89%	99.21%

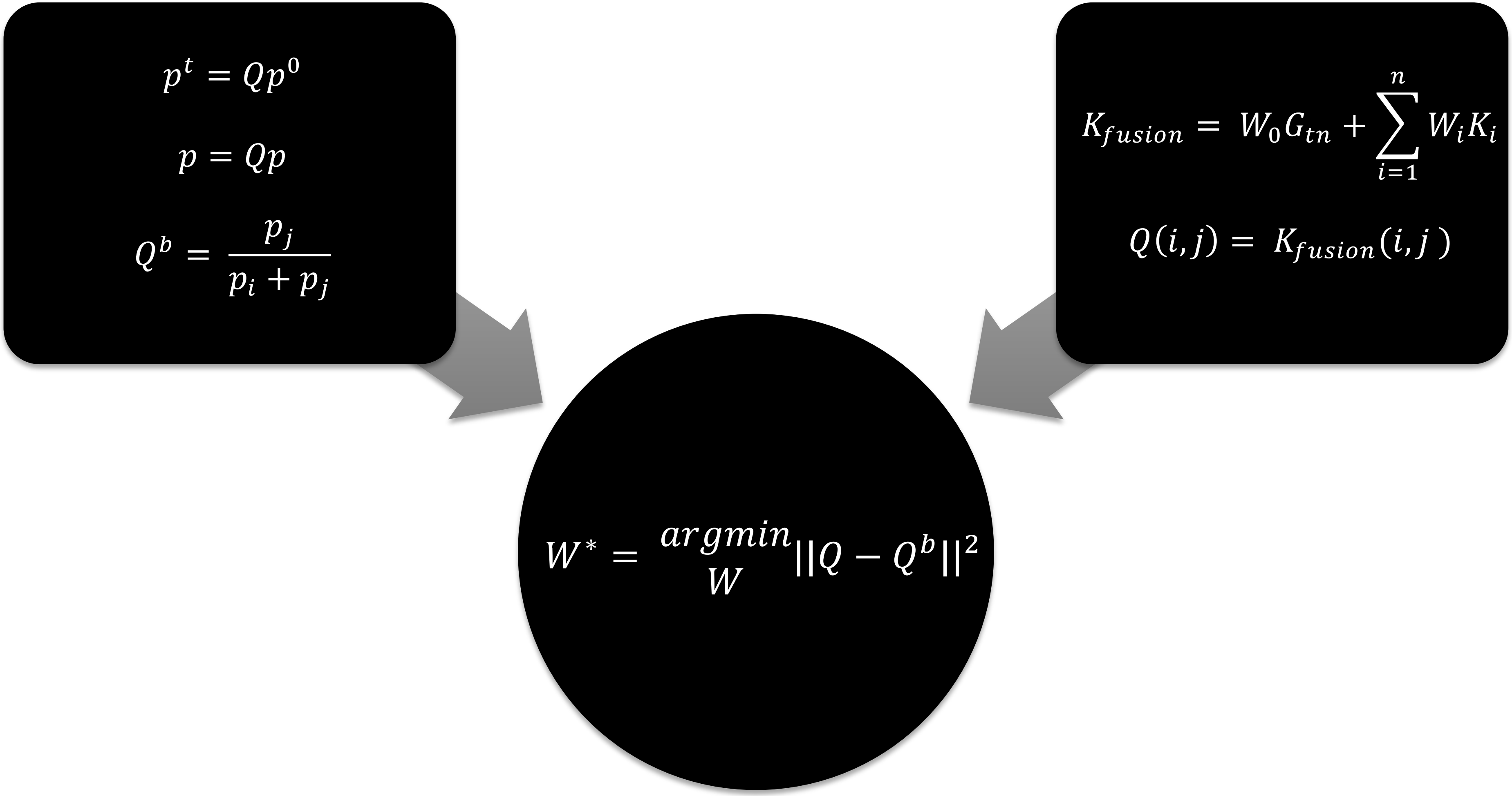
CM-ipHMM vs. ipHMM : p-value, 4.36E-77;
CM-ipHMM vs. CM-Only : p-value, 9.32E-10;
Ground-truth-CM: Replace predicted contact matrix with ground-truth
contact matrix

Basic idea of our method


Training Network
Gtn
Kernels
$W_0 \, G_{tn} + \sum_{i=1}^{n} W_i K_i$
Kernel Fusion
RL
P
$A_{inferred}(i, j) = \begin{cases} 1, & \text{if } P(i, j) > \epsilon \\ 0, & \text{otherwise} \end{cases}$
Inferred Network

$$RL = \sum_{k=0}^{\infty} \alpha^k (-L)^k = (I + \alpha * L)^{-1}$$

Weight Optimization by Linear Programming (WOLP)



$$p^t = Qp^0$$

$$p = Qp$$

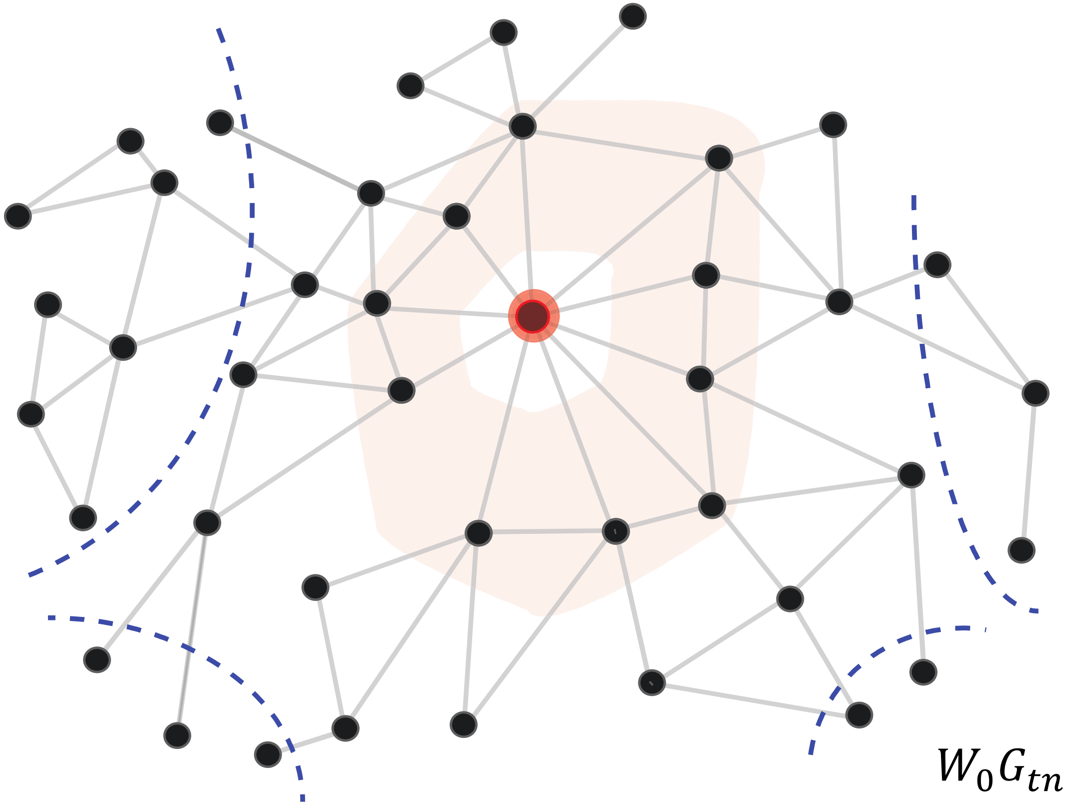
$$Q^b = \frac{p_j}{p_i + p_j}$$

$$K_{fusion} = W_0 G_{tn} + \sum_{i=1}^{n} W_i K_i$$

$$Q(i,j) = K_{fusion}(i,j\ )$$

$$W^* = \underset{W}{argmin} ||Q - Q^b||^2$$

Weight Optimization by Linear Programming (WOLP)



- Toy network G(V, E)
- Start node s: one of hubs
- D node set: $d(s, D_i) < 2$
- L node set: $d(s, L_i) > 2$
- M node set: $M = V \backslash (D \cup L)$

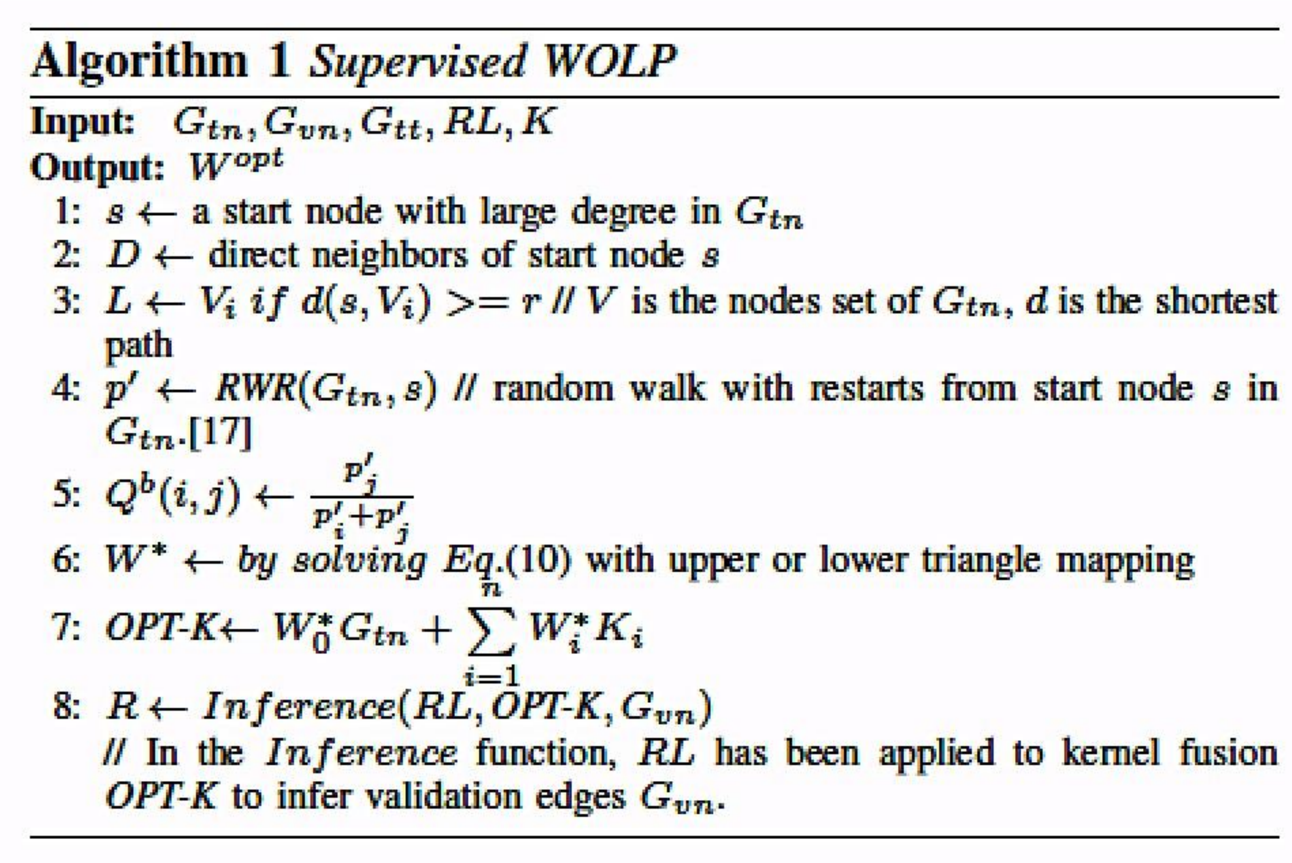
$$W^* = \underset{W}{argmin} ||Q - Q^b||^2$$

$$\Downarrow$$

$$W_0 G_{tn}(u, v) + \sum_{i=1}^{n} W_i K_i(u, v) = Q^b(u, v)$$

$$if\ u, v \in D \cup L\ \wedge K_i(u, v)! = 0\ \wedge (u < v\ \vee u > v)$$

Algorithm



Algorithm 1 *Supervised WOLP*

```
Input: $G_{tn}, G_{vn}, G_{tt}, RL, K$
Output: $W^{opt}$
  $s \leftarrow$ a start node with large degree in $G_{tn}$
  $D \leftarrow$ direct neighbors of start node $s$
  $L \leftarrow V_i\ if\ d(s, V_i) >= r$ // $V$ is the nodes set of $G_{tn}$, $d$ is the shortest path
  $p' \leftarrow RWR(G_{tn}, s)$ // random walk with restarts from start node $s$ in $G_{tn}$.[17]
  $Q^b(i, j) \leftarrow \frac{p'_j}{p'_i + p'_j}$
  $W^* \leftarrow$ by solving Eq.(10) with upper or lower triangle mapping
  OPT-K$\leftarrow W_0^* G_{tn} + \sum_{i=1}^{n} W_i^* K_i$
  $R \leftarrow Inference(RL, OPT\text{-}K, G_{vn})$
  // In the $Inference$ function, $RL$ has been applied to kernel fusion OPT-K to infer validation edges $G_{vn}$.
```

- ➢Golden standard connected network: G(V, E)
- ➢Connected training network: G (V, E)

$G_{tn}(E) \quad G_{vn}(E) \quad G_{tt}(E) = G\ (E)$

$G_{tn}(E) \quad G_{vn}(E) \quad G_{tt}(E) = f$

Experiments on network inference with real data

Data description of DIP yeast PPI networks(Release 20150101)

- Largest connected component: G(V, E) = G(5,030, 22,394)

① Connected training network: G_{tn}(V, E) = (5,030, 5,394)

② Validation edge set: G_{vn}(V, E) = (?, 1,000)

③ Testing edge set: G_{tt}(V, E) = (?, 16,000)

Experiments on network inference with real data)

- Feature kernels [39]

① $K_{Jaccard}$[15]: This kernel measure the similarity of protein pairs i, j in term of neighbors(i)∩neighbors(j)/neighbors(i)∪neighbors(j).

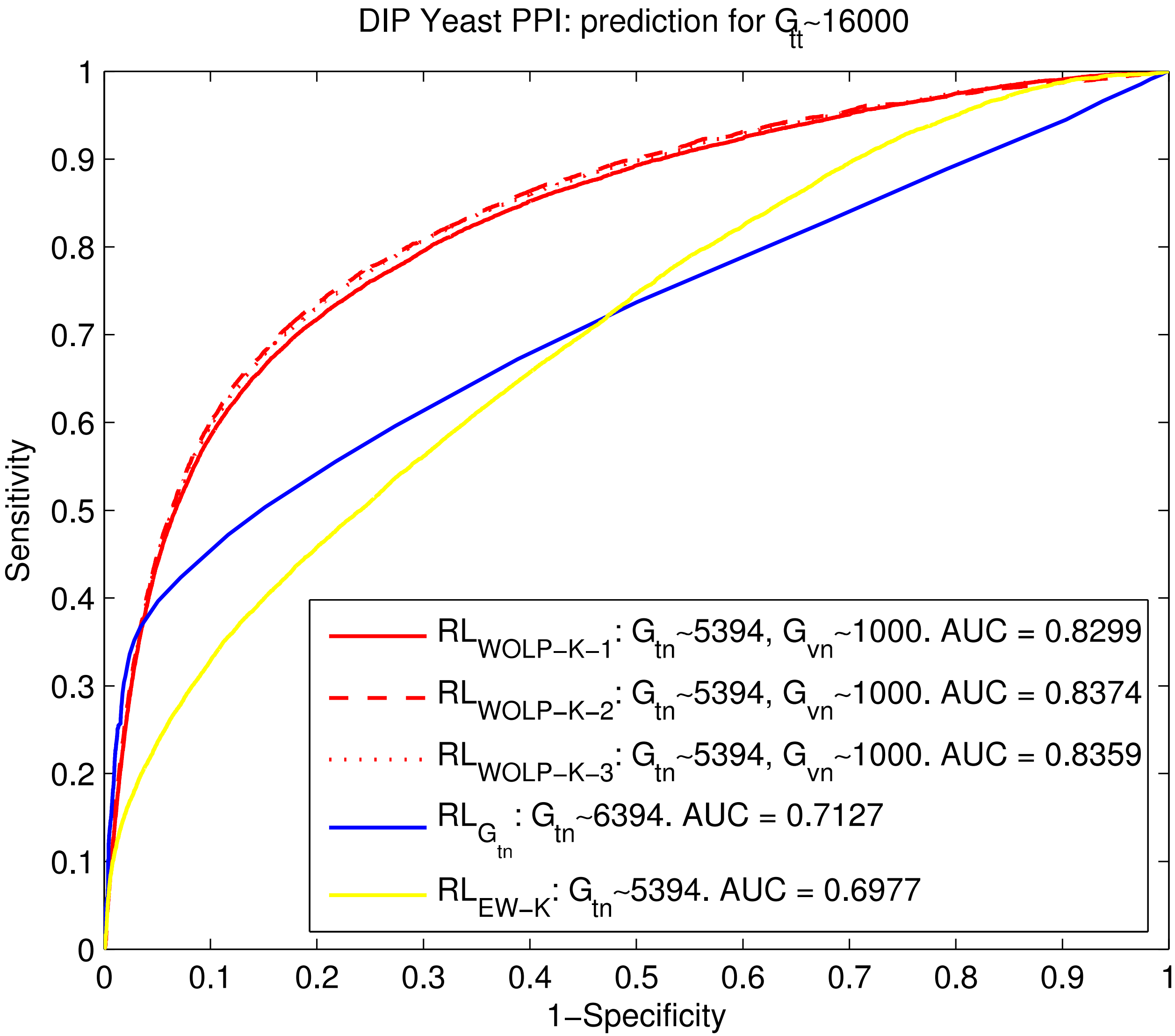
② K_{SN}: It measures the total number of neighbors of protein i and j, KSN = neighbors(i) + neighbors(j).

③ K_B: It is a sequence-based kernel matrix that is generated using the BLAST.

④ K_E: This is a gene co-expression kernel matrix constructed entirely from microarray gene expression measurements.

⑤ K_{Pfam}: This is a generalization of the previous pairwise comparison-based matrices in which the pairwise comparison scores are replaced by expectation values derived from hidden Markov models (HMMs) in the Pfam database.

[39] Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics**, 20(suppl 1):i363–i370, 2004.**


DIP Yeast PPI: prediction for G_{tt}~16000
Sensitivity
1−Specificity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
$RL_{WOLP-K-1}$: G_{tn}~5394, G_{vn}~1000. AUC = 0.8299
$RL_{WOLP-K-2}$: G_{tn}~5394, G_{vn}~1000. AUC = 0.8374
$RL_{WOLP-K-3}$: G_{tn}~5394, G_{vn}~1000. AUC = 0.8359
$RL_{G_{tn}}$: G_{tn}~6394. AUC = 0.7127
RL_{EW-K}: G_{tn}~5394. AUC = 0.6977