计算系统生物学

王 勇

中国科学院数学与系统科学研究院

http://zhangroup.aporc.org

Chinese Academy of Sciences

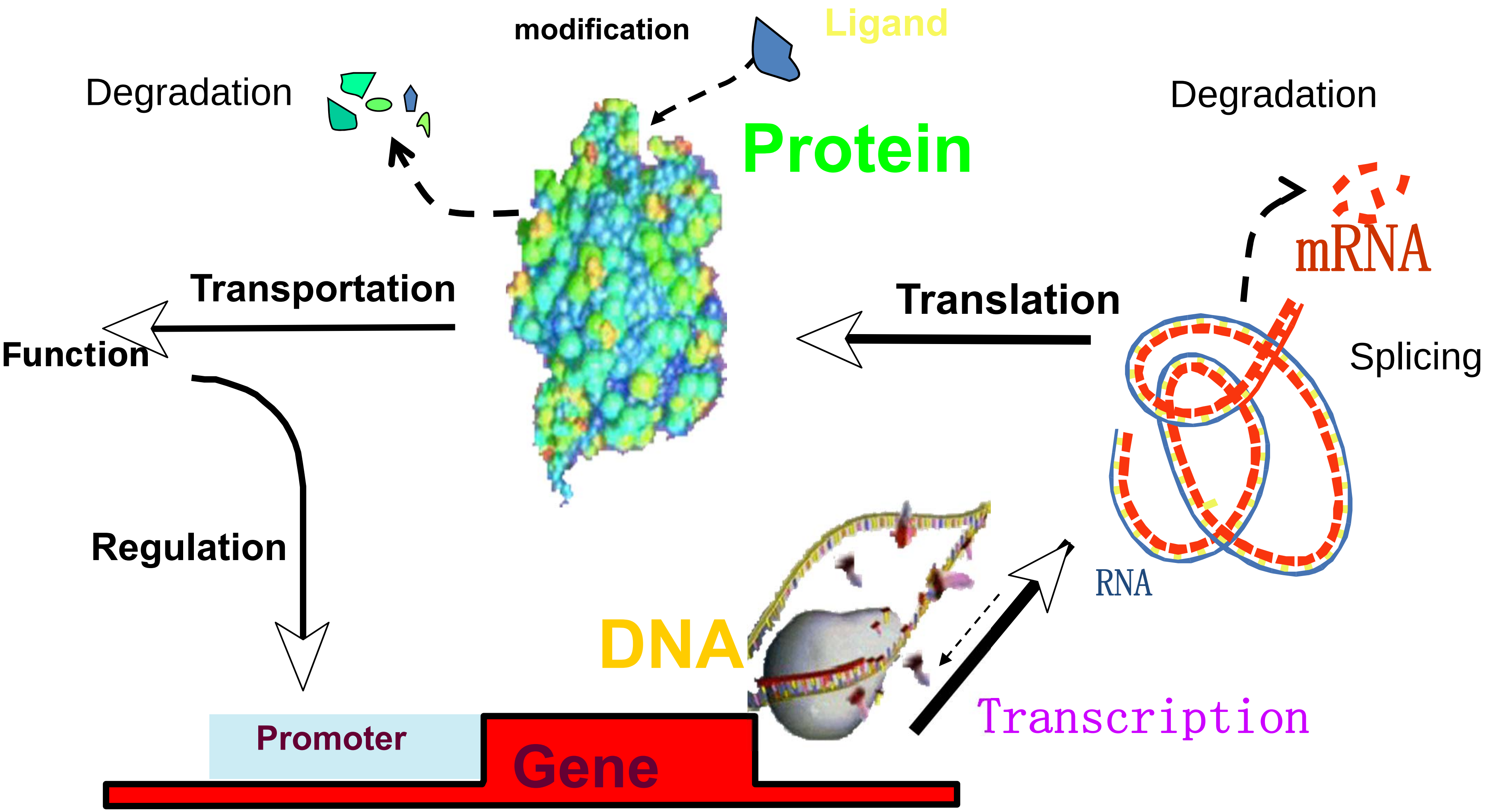
Predicting cooperativity of transcription factors

by integrating heterogenous data sources

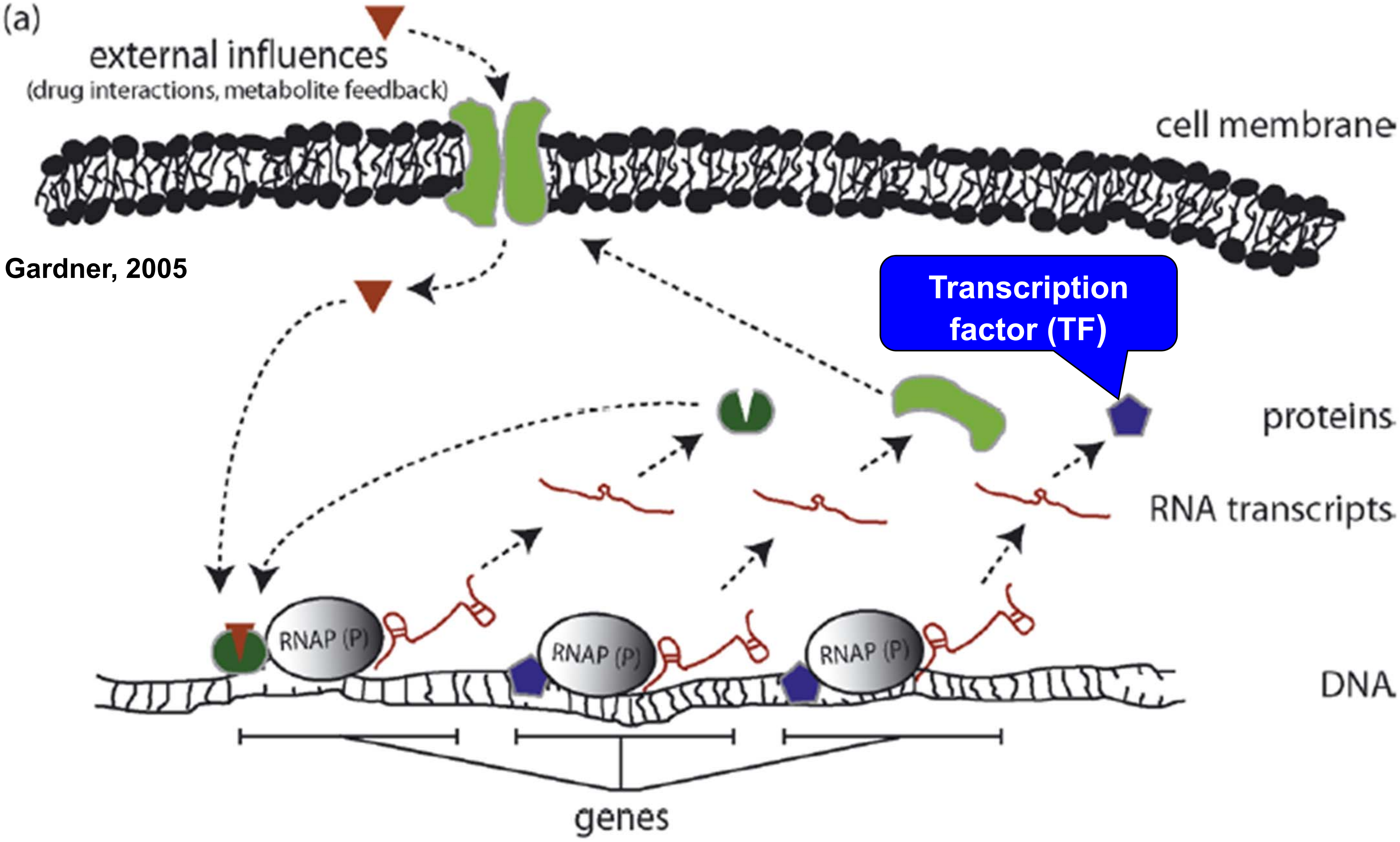
Yong Wang

Academy of Mathematics & Systems Science

Central dogma of molecular biology



Gene regulation

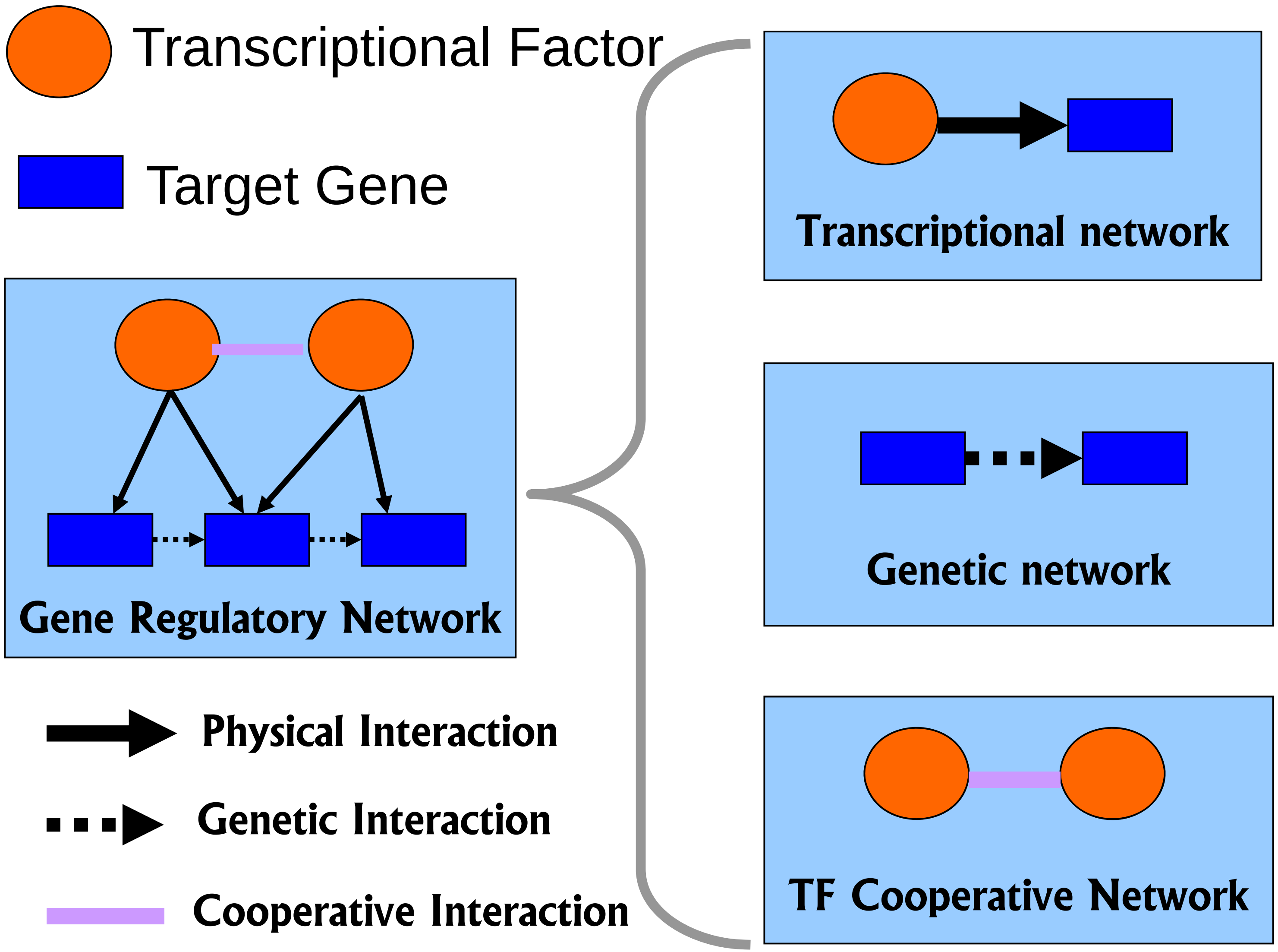


Gardner, 2005

Transcription factor (TF)

Transcription factors (TFs) are proteins that **dynamically** read and interpret the **static** genetic instructions in the DNA

Basic building blocks for gene regulatory network



Transcriptional Factor

Target Gene

Gene Regulatory Network

Physical Interaction

Genetic Interaction

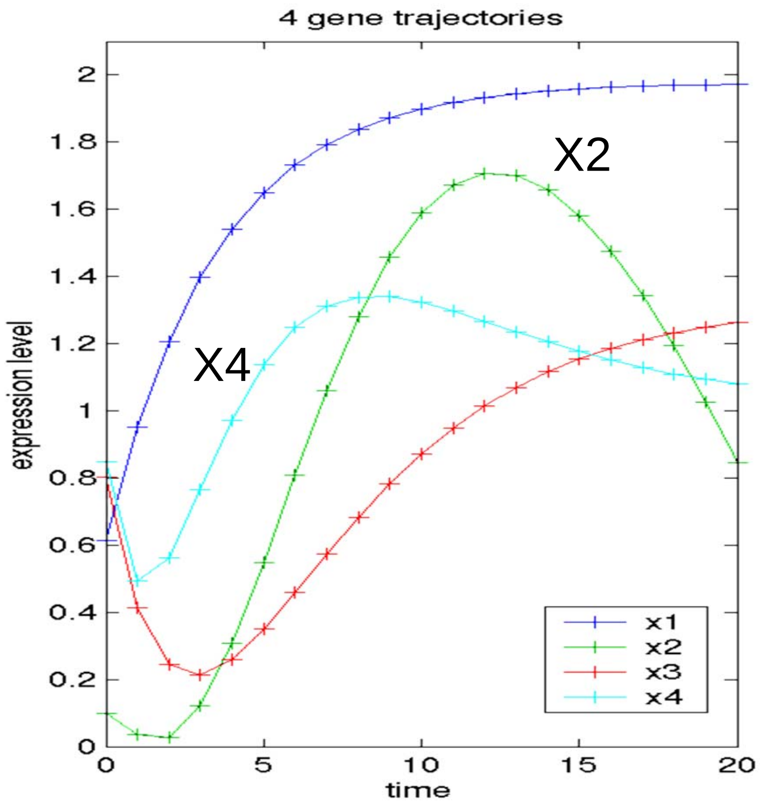
Cooperative Interaction

Transcriptional network

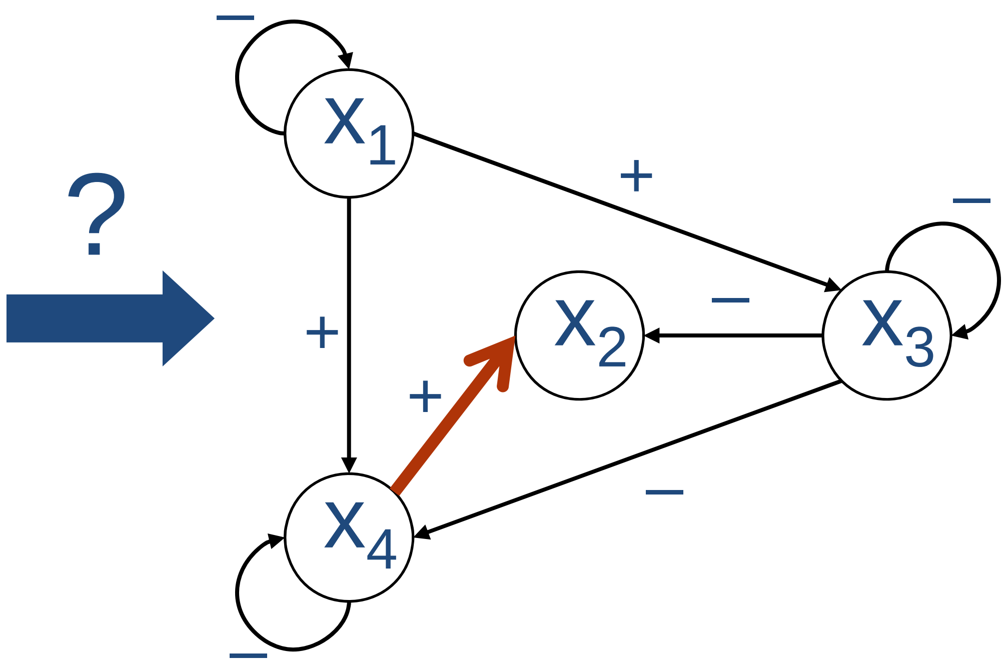
Genetic network

TF Cooperative Network

Gene regulatory network inference

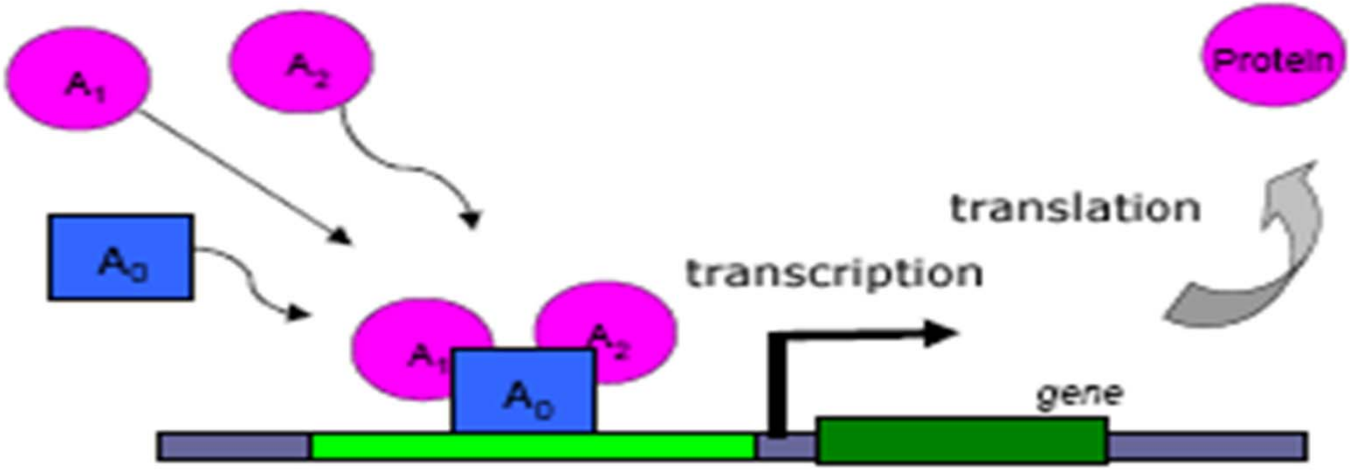


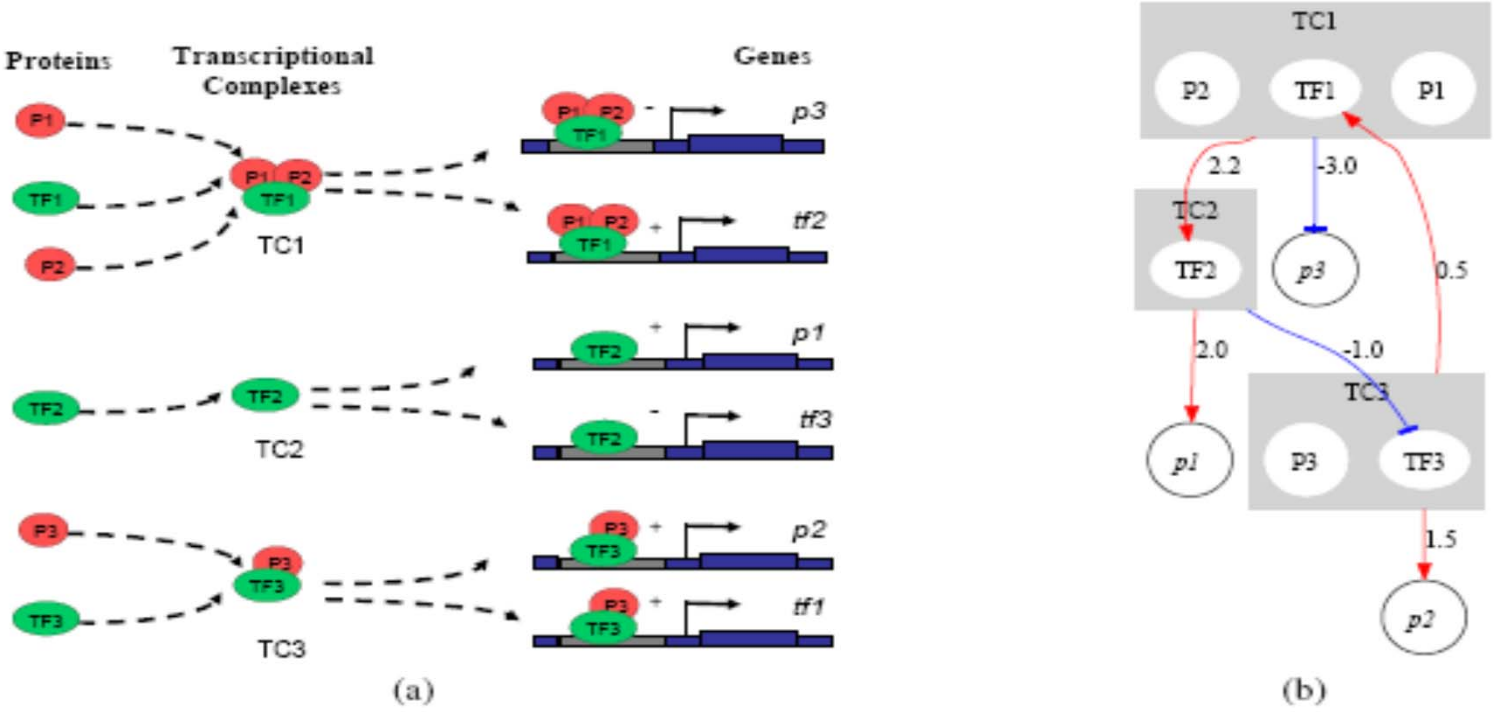
Time series data of gene expression



Indirect influence among genes

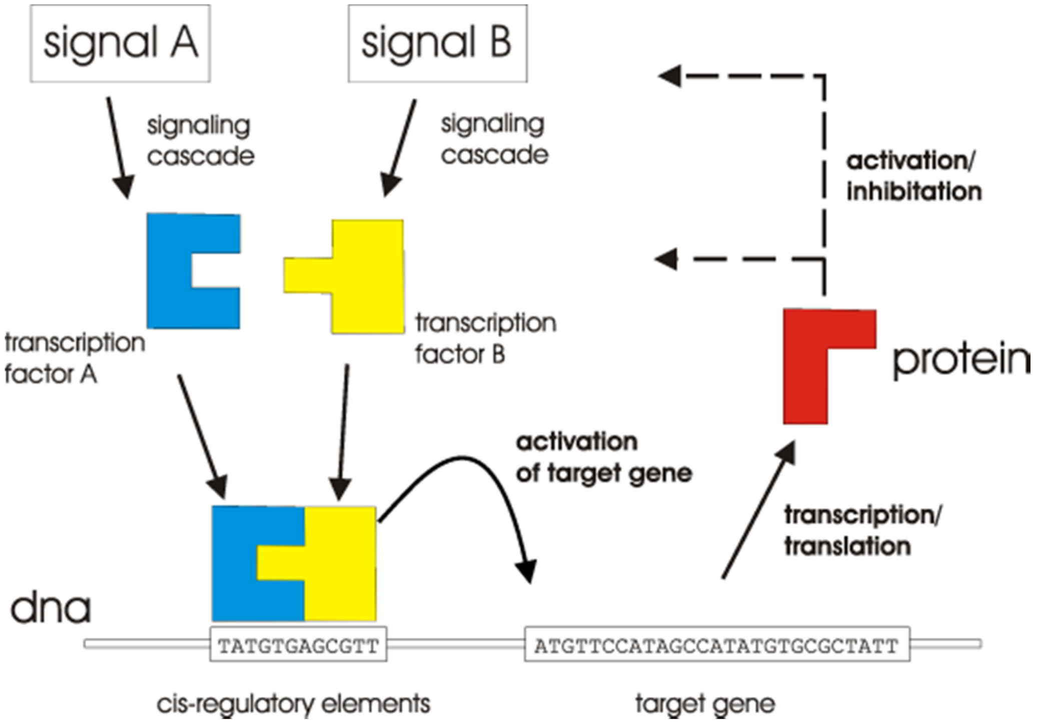
Transcriptional regulation inference



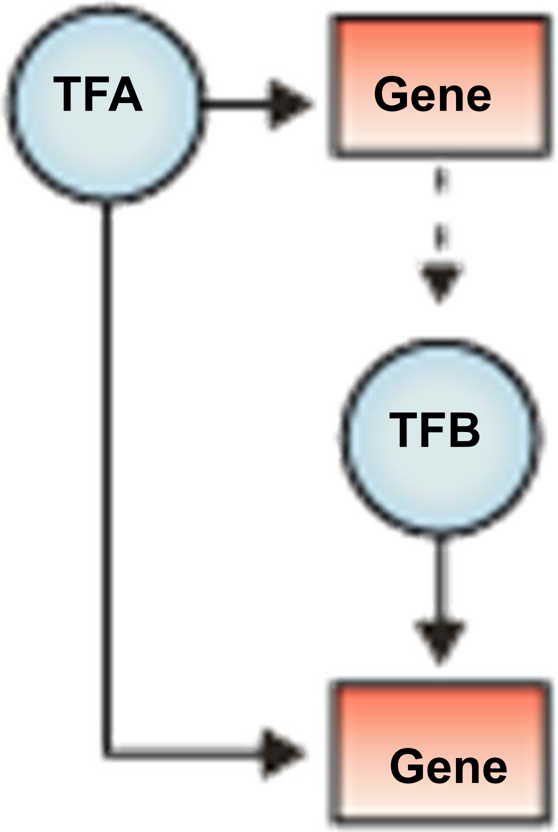


Physical interactions between TFs and target genes

What is TF cooperativity?



Direct interaction

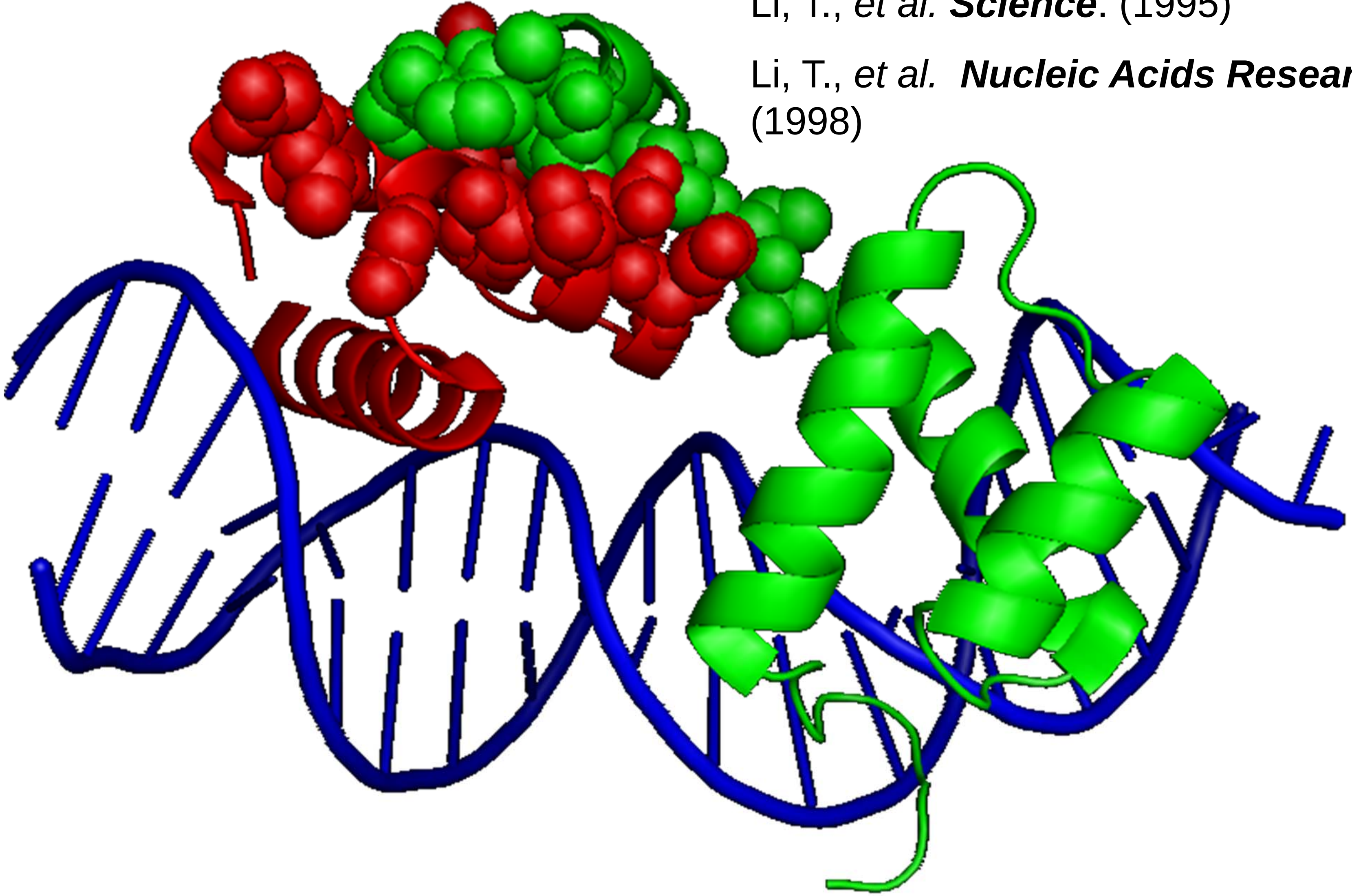


Indirect interaction

Example 1

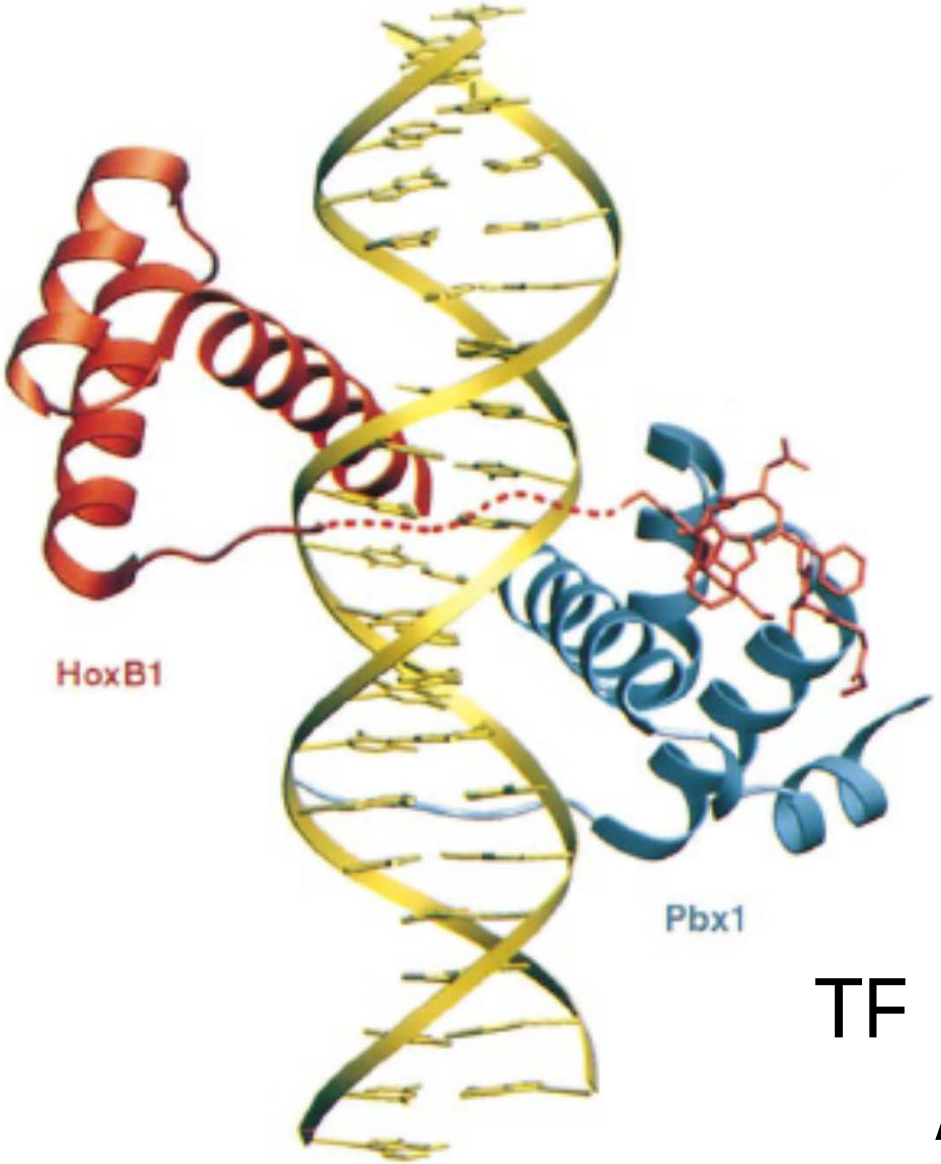
Li, T., *et al.* ***Science***. (1995)

Li, T., *et al.* ***Nucleic Acids Research***. (1998)



Structure of the MAT a1/α2 TF complex bound to DNA fragment in yeast

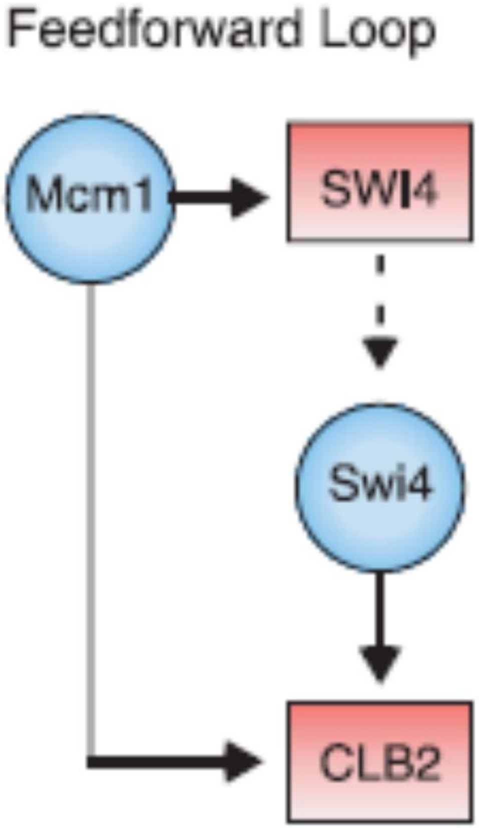
Example 2

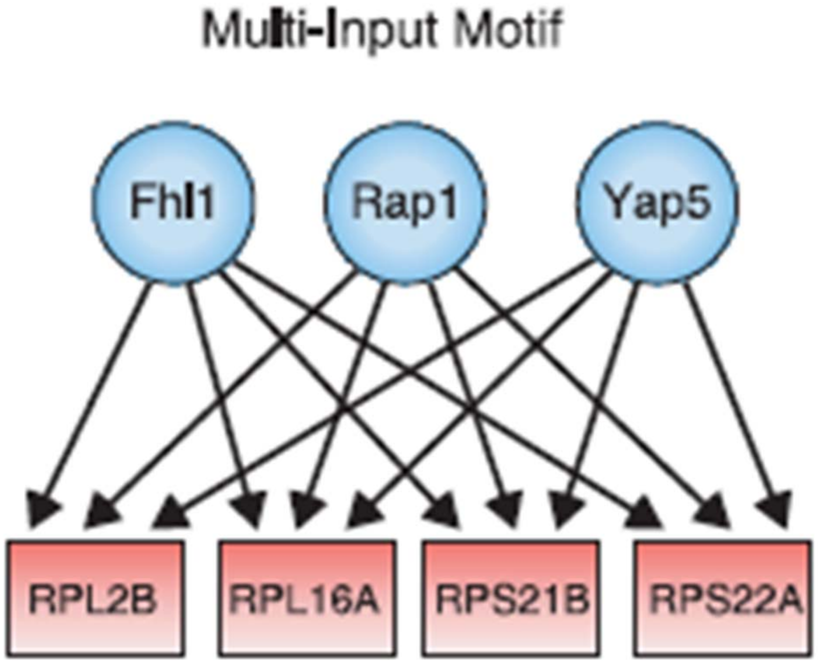


Wolberger, *et al.* ***Cell,*** *(*1999*)*

TF Interactions ***In Drosophila***

Example 3





Here, a master regulator binds to the promoter of a second regulator, then both regulators bind a common target gene.

Lee, T. *et al.* ***Science*** (2002)

TF-TF cooperativity is important

- Transcription factors usually cooperate with other TFs to facilitate (as an activator) or inhibit (as a repressor) the recruitment of RNA polymerase

- TFs use complex logic rules building upon simple ones (AND, OR, and NOT) to control the precise condition-dependent expression of target genes

- The idea of combinatorial regulation as a primary mechanism for achieving fine-tuned transcriptional control

- One reason for the complexity of gene regulatory network.

Can we predict TF cooperativity?

- Experimental methods for detecting TF interaction include co-immunoprecipitation and super-gel shift.

- These methods are generally time-consuming and expensive.

- It is difficult to apply them to mapping the whole-genome TF cooperativity network in the living cell.

Existing studies

- Case studies
- Using information from a single data source such as TF binding motif, target gene, and TF activity
- Unsupervised framework

Our work

- **Main idea:** first supervised framework to integrate all the available data together
- **Challenges:**

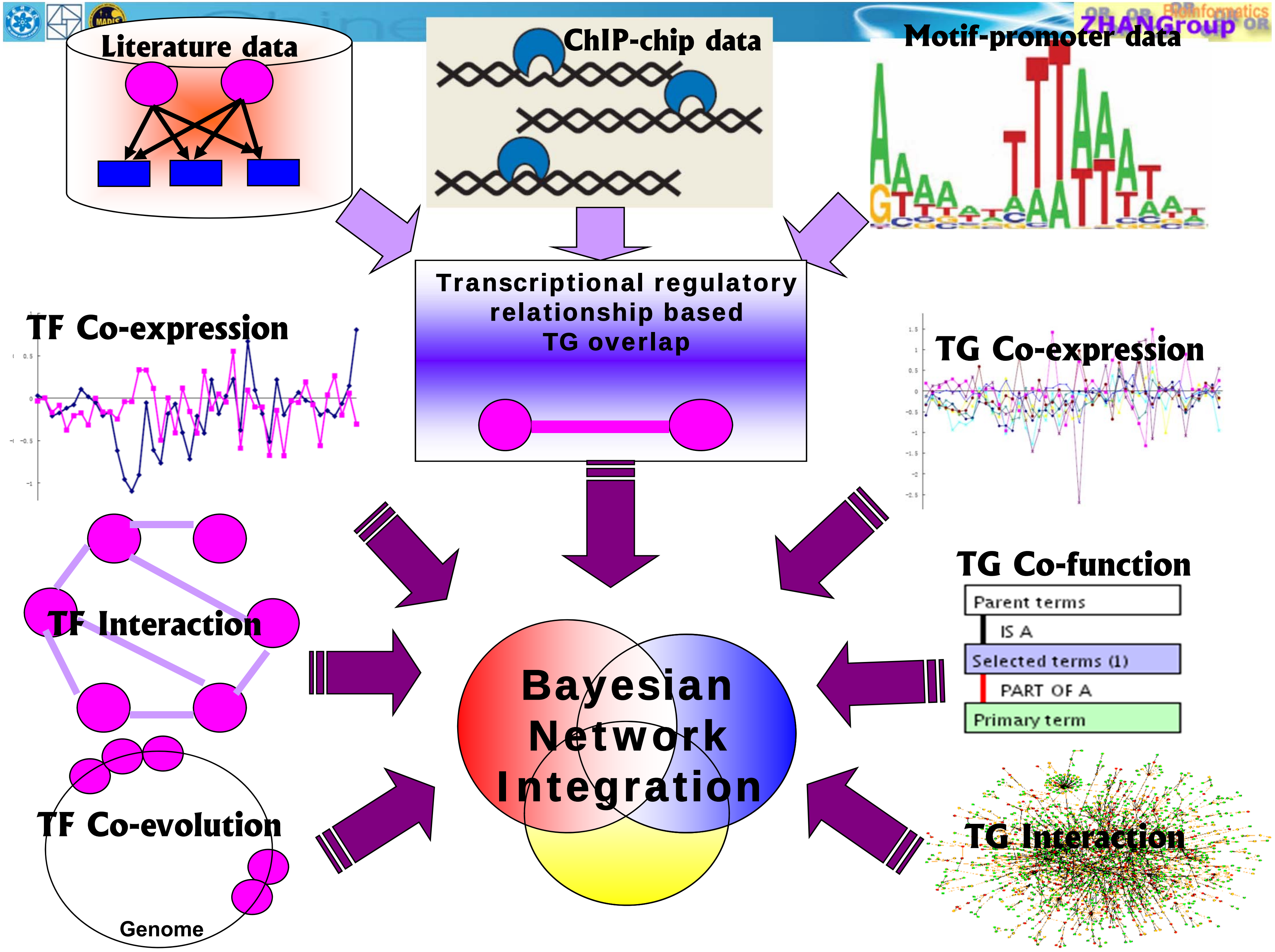
1. Scarcity of gold-standard data
2. Collection and assessment of genomic predictors for TF cooperativity
3. Optimally integrate these predictors

Part I: Collecting features

Model eukaryote: Yeast

- We choose *Saccharomyces cerevisiae* as our model eukaryote since many different types of genome-wide data sources are available in yeast.

- 174 TFs and 6000 target genes
- ChIP-chip data: 143 TFs, 4,774 TGs, and 16,656 transcriptional regulations
- Literature data: 162 TFs, 4,716 TGs, and 17,616 transcriptional regulations
- 108 TFs have 281 specific DNA binding motifs
- Sequence for the 6000 target genes
- 5,193 proteins and 111,883 protein interactions
- Many gene expression data under different conditions
- 41,984 co-evolutionary linkages among 3,047 proteins



Literature data

ChIP-chip data

Motif-promoter data

Transcriptional regulatory relationship based TG overlap

TF Co-expression

TG Co-expression

TF Interaction

TG Co-function

Parent terms

IS A

Selected terms (1)

PART OF A

Primary term

TF Co-evolution

Genome

TG Interaction

Bayesian Network Integration

List of Features

- 15 TF pair features that potentially correlate with TF cooperativity relationships.
- TF physical/genetic interaction
- TF co-expression,
- TF co-evolution relationships
- The degree of overlap among the corresponding target genes (TGs) based on literature, ChIP-chip, and motif occurrence evidence (3 features)
- The degree of coherence among the corresponding target genes (TGs) in terms of co-expression, co-function, and interaction, based on literature, ChIP-chip, and motif occurrence evidence (9 features)

TG overlap

- Cooperative TFs tend to share larger number of target genes than expected by chance
- We developed two scores, a p-value score and an enrichment score, to assess the significance of TG overlap for a pair of TFs.
- For a given TF pair, to determine whether the observed TG overlap m is statistically significant, we fix the total number TGs in the yeast genome (*N*), the number of TGs regulated by the first TF (*N*1), the number of TGs regulated by the second TF (*N*2).

- An enrichment score is defined as the ratio of the observed TG overlap versus the expected TG overlap by chance, as follows:

$$F = \frac{Nm}{N_1 N_2}$$

A score larger than 1 indicates that there is more TG overlap than expected by chance.

- Let the number of TGs regulated by both TFs be a random variable *X*. *X* follows a hypergeometric distribution:

$$P(X = i) = \frac{\binom{N_1}{i}\binom{N - N_1}{N_2 - i}}{\binom{N}{N_2}}$$

- A p-value score, which is defined as the probability that the TG overlap would assume a value greater than or equal to the observed value, *m*, by chance:

$$P(X \geq m) = 1 - \sum_{i=0}^{m-1} P(X = i)$$

TG coherence

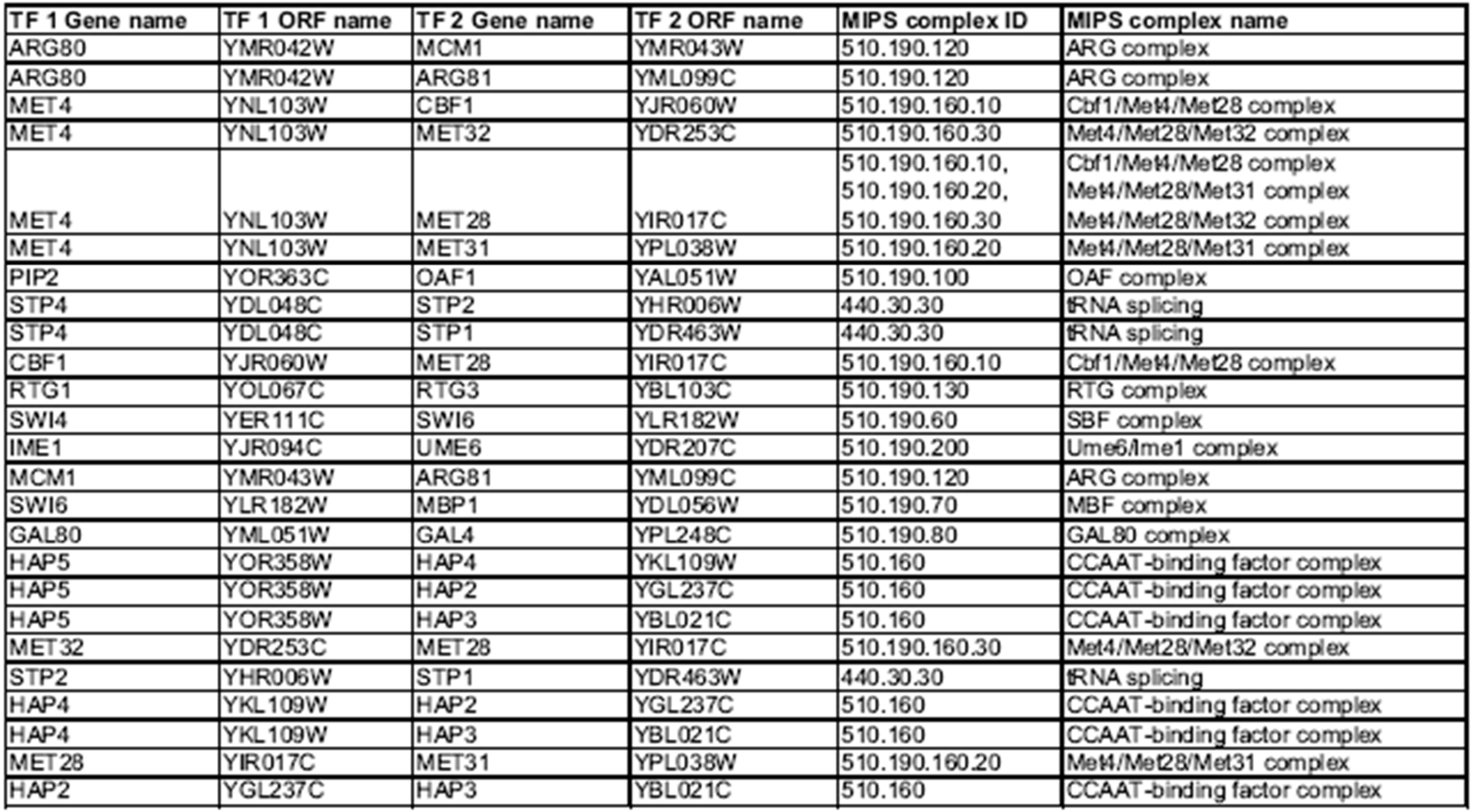
- The target genes of cooperative TFs tend to share significant co-expression, co-function, and interaction relationships.
- First step: computing the pairwise co-expression, co-function and interaction relationships for co-regulated TG pairs
- Second step: calculating the fraction of TG pairs with a specific relationship (co-expression, co-function, or interaction) for the co-regulated TG set.
- Third step: calculating the fraction of TG pairs with the same specific relationship for the entire set of TGs regulated by any TF.
- Fourth step: An coherence enrichment score is defined as the ratio of these two fractions.

Assessing features

Gold standard data

- We compiled 25 TF pairs each belonging to the same biochemically well-defined complex according to the MIPS complex catalogue as our gold-standard positives (GSP).
- We constructed an approximate gold-standard negative (GSN) set for TF cooperativity by identifying all TF pairs that do not belong to any known MIPS complex.
- The GSP set is the only high-quality dataset of TF cooperativity currently available, and is more restrictive.
- The GSN set is expected to contain a small fraction of false negatives.

- Nevertheless, our results suggest that the quality of the GSP and GSN sets are good enough to approximately assess the contribution of each piece of evidence before integrating them in an optimal way.



TF 1 Gene name	TF 1 ORF name	TF 2 Gene name	TF 2 ORF name	MIPS complex ID	MIPS complex name
ARG80	YMR042W	MCM1	YMR043W	510.190.120	ARG complex
ARG80	YMR042W	ARG81	YML099C	510.190.120	ARG complex
MET4	YNL103W	CBF1	YJR060W	510.190.160.10	Cbf1/Met4/Met28 complex
MET4	YNL103W	MET32	YDR253C	510.190.160.30	Met4/Met28/Met32 complex
MET4	YNL103W	MET28	YIR017C	510.190.160.10, 510.190.160.20, 510.190.160.30	Cbf1/Met4/Met28 complex Met4/Met28/Met31 complex Met4/Met28/Met32 complex
MET4	YNL103W	MET31	YPL038W	510.190.160.20	Met4/Met28/Met31 complex
PIP2	YOR363C	OAF1	YAL051W	510.190.100	OAF complex
STP4	YDL048C	STP2	YHR006W	440.30.30	tRNA splicing
STP4	YDL048C	STP1	YDR463W	440.30.30	tRNA splicing
CBF1	YJR060W	MET28	YIR017C	510.190.160.10	Cbf1/Met4/Met28 complex
RTG1	YOL067C	RTG3	YBL103C	510.190.130	RTG complex
SWI4	YER111C	SWI6	YLR182W	510.190.60	SBF complex
IME1	YJR094C	UME6	YDR207C	510.190.200	Ume6/Ime1 complex
MCM1	YMR043W	ARG81	YML099C	510.190.120	ARG complex
SWI6	YLR182W	MBP1	YDL056W	510.190.70	MBF complex
GAL80	YML051W	GAL4	YPL248C	510.190.80	GAL80 complex
HAP5	YOR358W	HAP4	YKL109W	510.160	CCAAT-binding factor complex
HAP5	YOR358W	HAP2	YGL237C	510.160	CCAAT-binding factor complex
HAP5	YOR358W	HAP3	YBL021C	510.160	CCAAT-binding factor complex
MET32	YDR253C	MET28	YIR017C	510.190.160.30	Met4/Met28/Met32 complex
STP2	YHR006W	STP1	YDR463W	440.30.30	tRNA splicing
HAP4	YKL109W	HAP2	YGL237C	510.160	CCAAT-binding factor complex
HAP4	YKL109W	HAP3	YBL021C	510.160	CCAAT-binding factor complex
MET28	YIR017C	MET31	YPL038W	510.190.160.20	Met4/Met28/Met31 complex
HAP2	YGL237C	HAP3	YBL021C	510.160	CCAAT-binding factor complex

Likelihood ratio score

- Given the GSP and GSN datasets. The likelihood ratio for a binary feature f taking on a particular value (1 or 0; presence or absence) is defined as the fraction of GSP where the feature takes on the given value, divided by the fraction of GSN where the feature takes on the given value

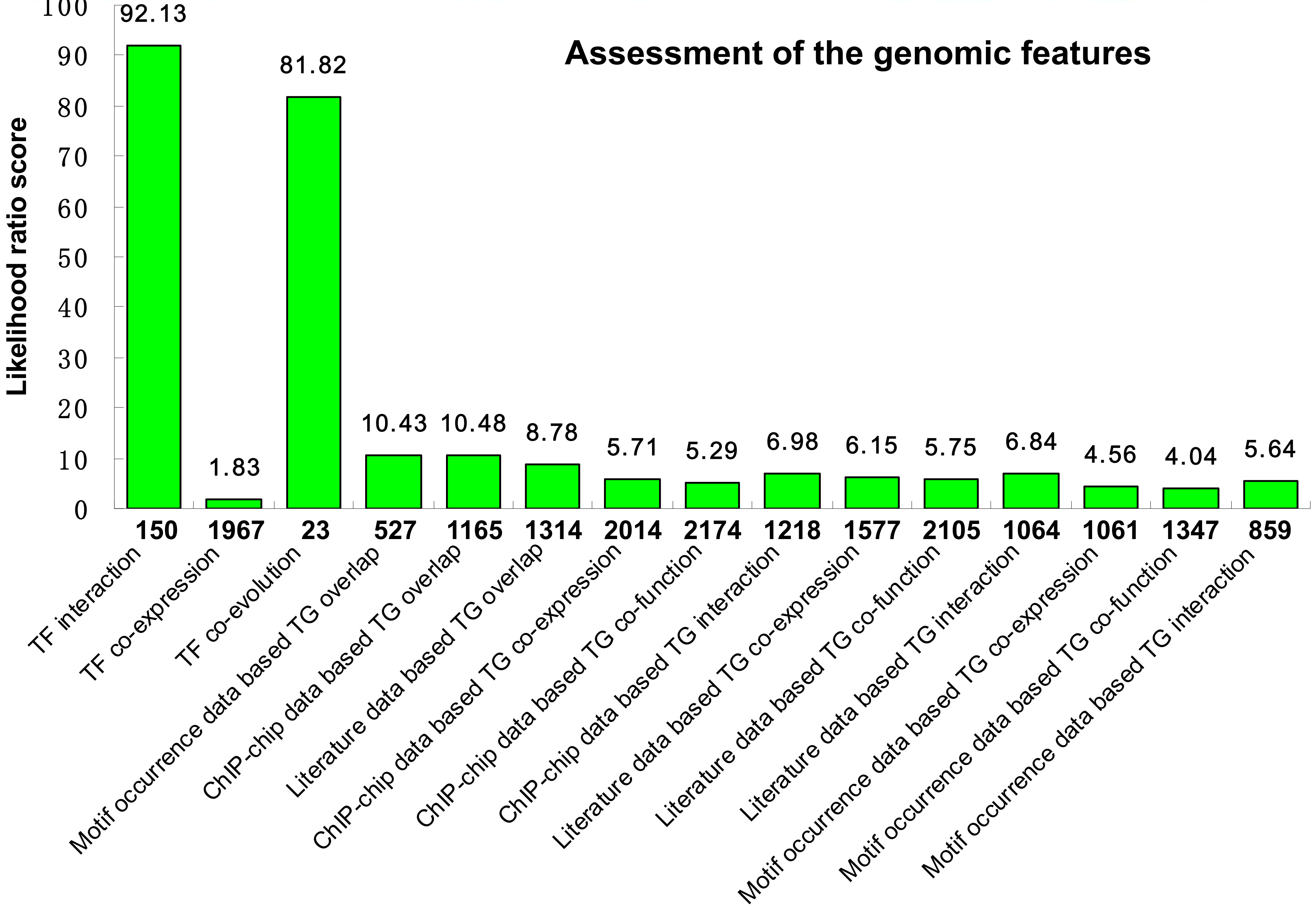
- LR = Prob(f | GSP) / Prob(f | GSN)

- A likelihood ratio score much larger than 1 indicates that the feature is a good predictor for TF cooperativity.

- Likelihood ratio scores can guide the choice of proper cutoffs to convert numerical features into categorical ones.

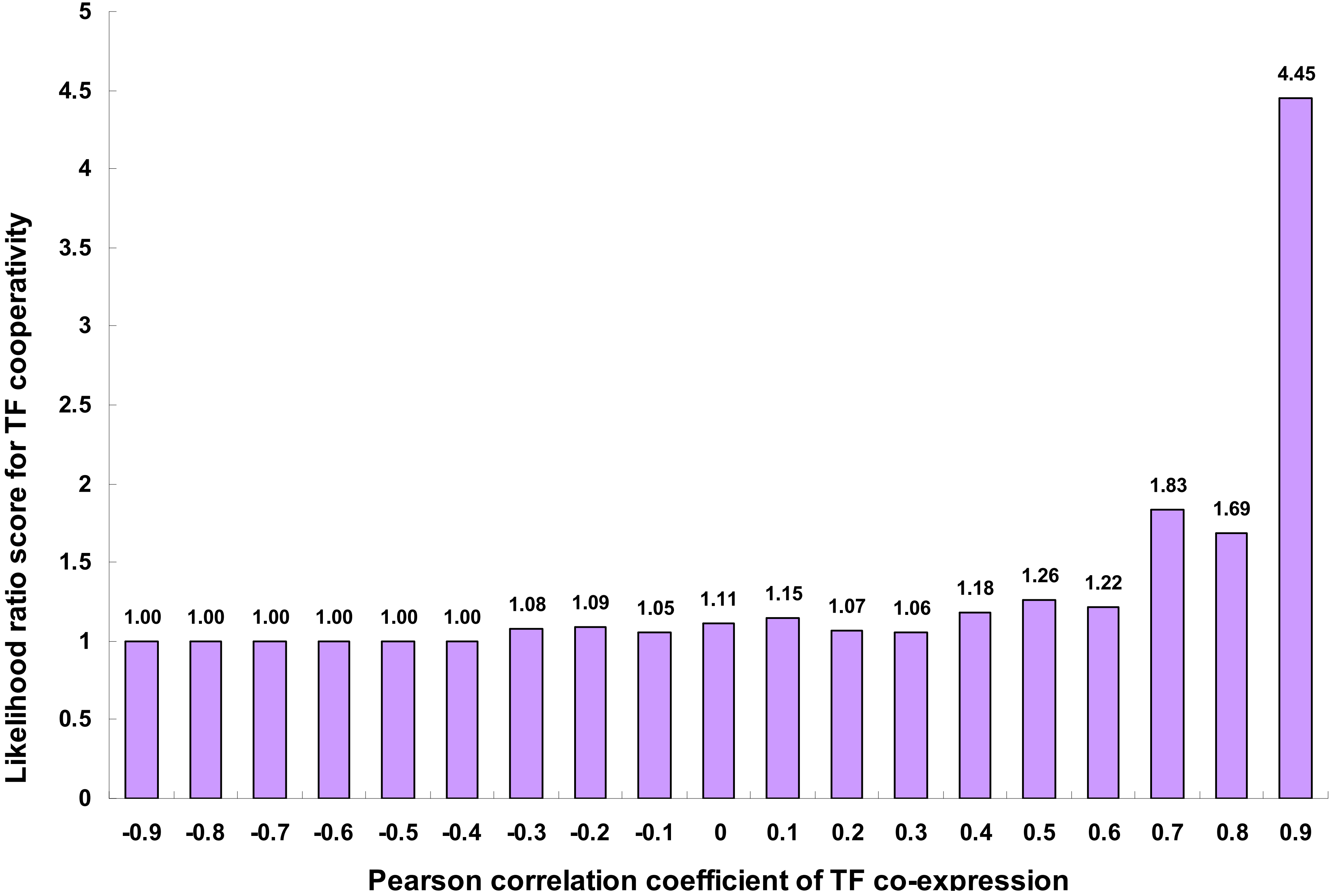
- We can quantitatively assess the predictive power of each feature.

- Because the same gold-standard data is used throughput the integration process, each feature can be assessed on a common benchmark. Different features are directly comparable by the likelihood ratio scores even though the data sources are highly heterogeneous.

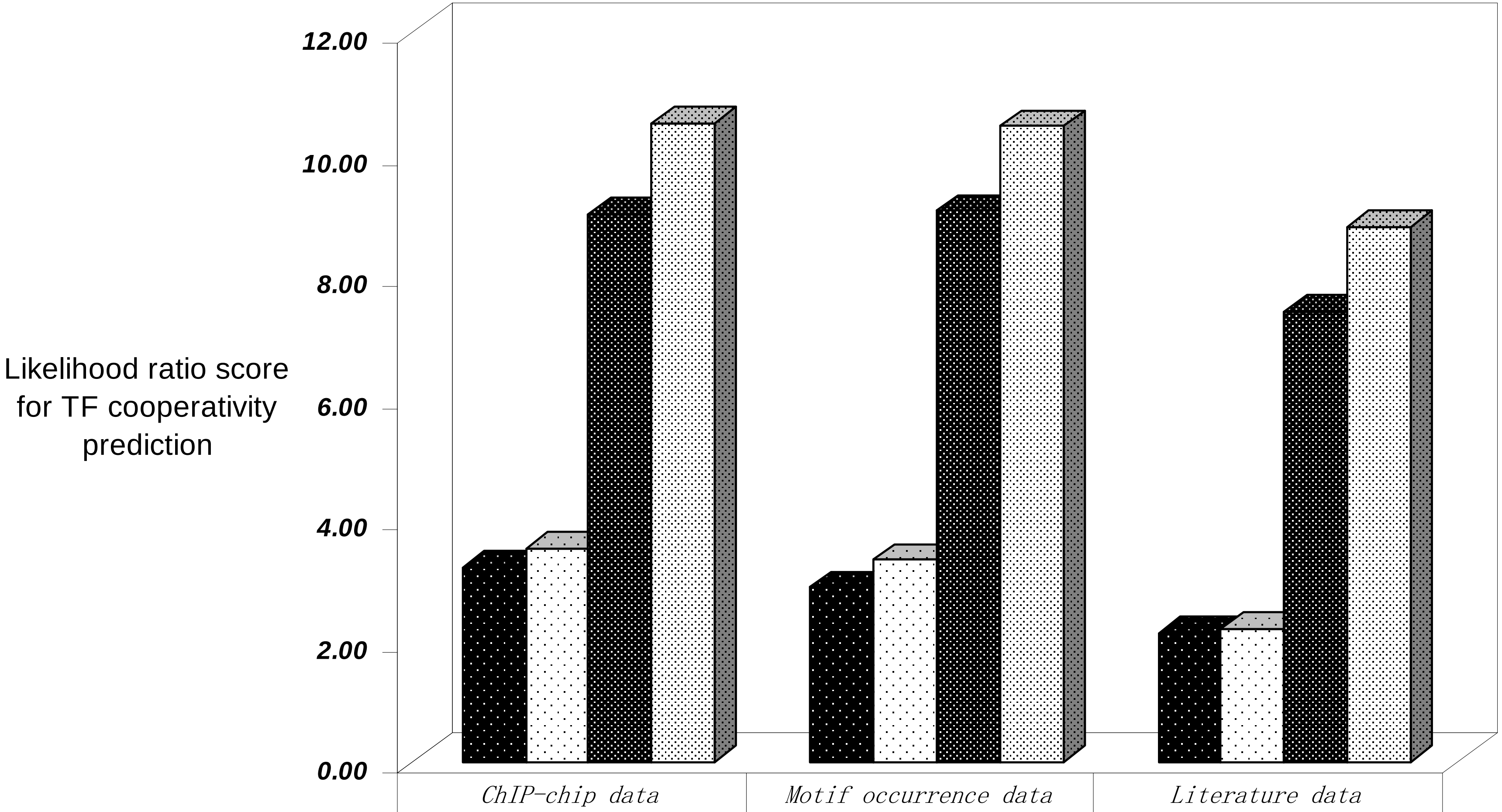


Assessment of the genomic features

Looking back: Guide the choice of parameters

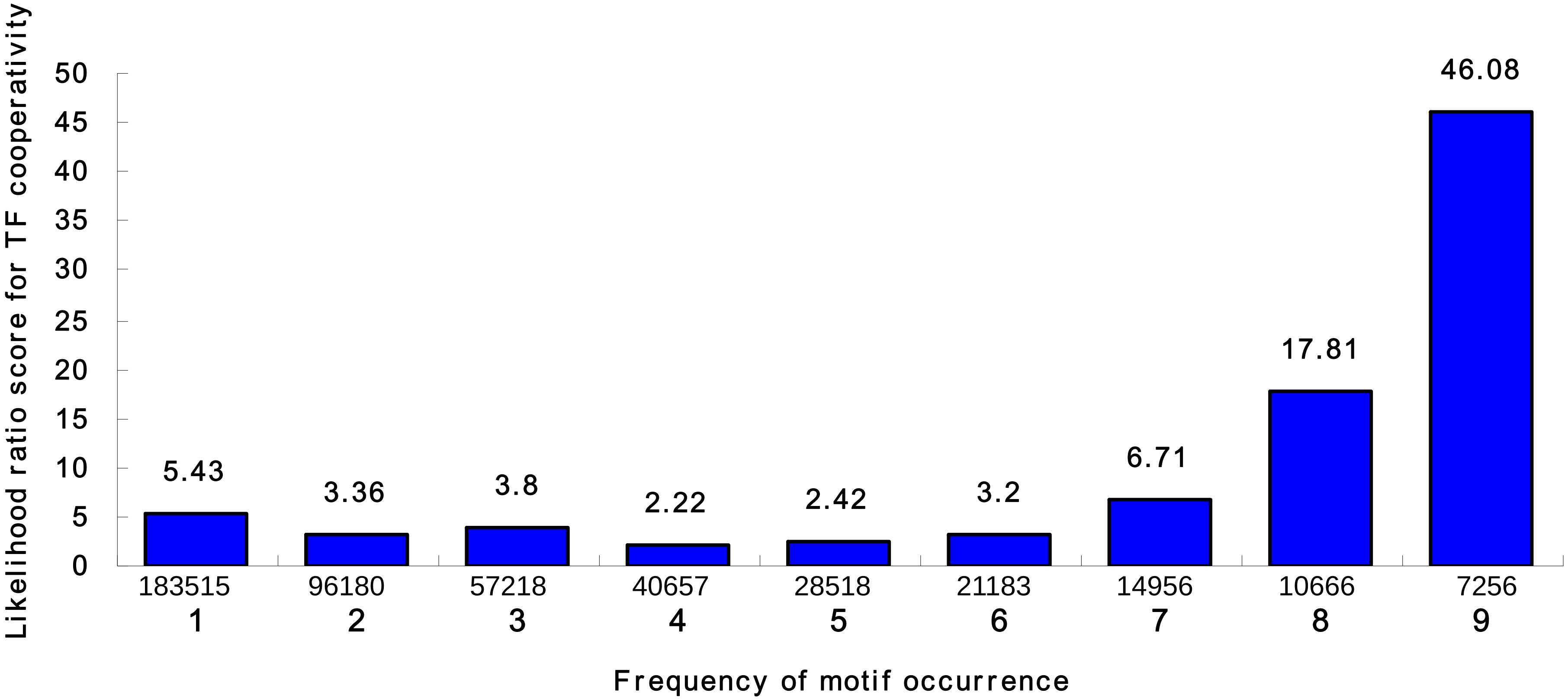


Looking back: combination of p-value and enrichment score

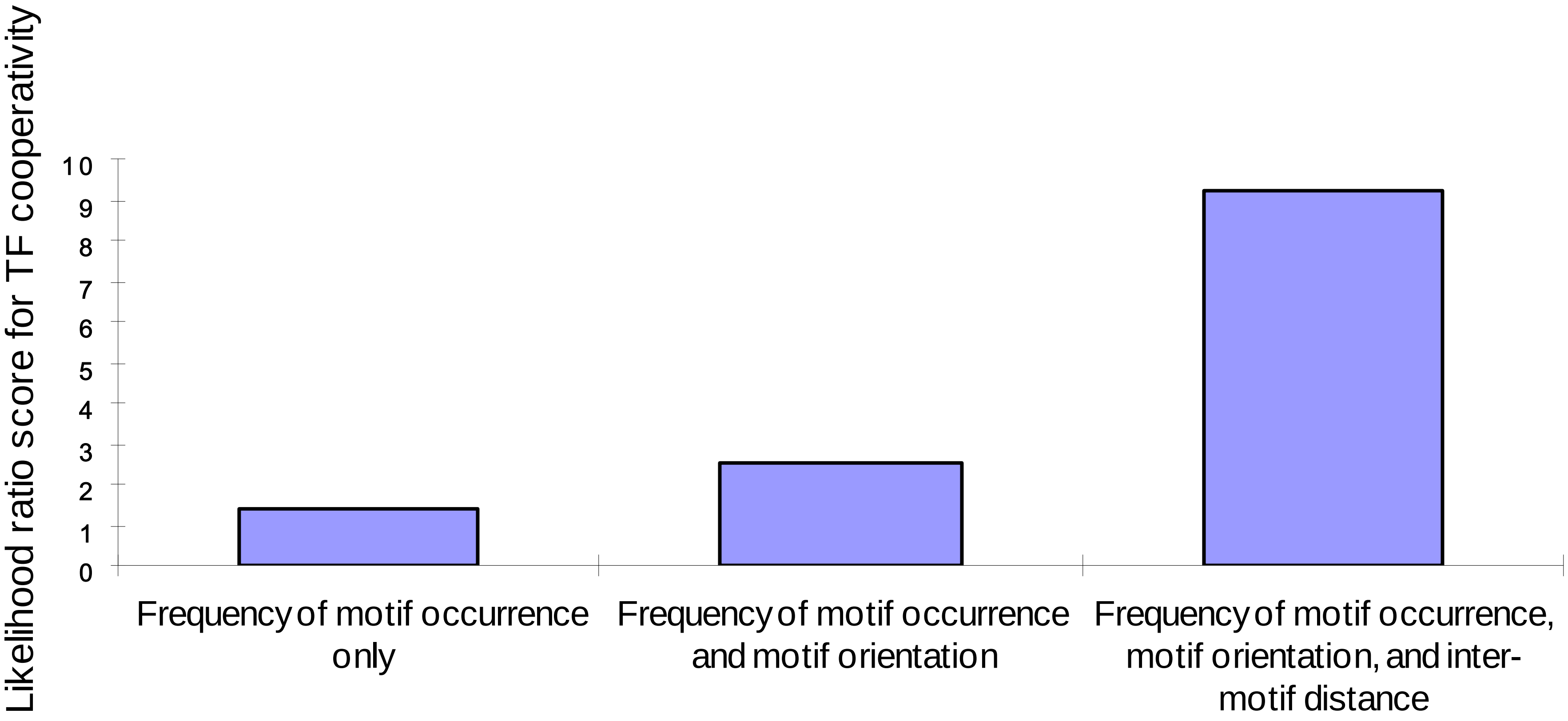


	ChIP-chip data	Motif occurrence data	Literature data
■ *Control score*	*3.20*	*2.86*	*2.10*
□ *Enrichment score only*	*3.49*	*3.31*	*2.18*
▩ *P-value score only*	*8.99*	*9.04*	*7.39*
▦ *Combination of p-value and enrichment score*	*10.48*	*10.43*	*8.78*

Looking back: construction of the motif-occurrence transcriptional regulatory network



Looking back: construction of the motif-occurrence transcriptional regulatory network



Integrating features

Bayesian network method

- For each TF pair, the prediction of cooperativity is based on the calculation of the posterior odds of cooperativity given the presence of genomic features.
- The posterior odds for predicting the class label *y* (1 if exists cooperativity, and 0 otherwise) by integrating genomic features *f*1, *f*2,…, *f*n can be written as follows using the Bayes rule:

$$\log\frac{P(y=1\mid f_1,f_2,\ldots f_n)}{P(y=0\mid f_1,f_2,\ldots f_n)}=\log\frac{P(y=1)}{P(y=0)}+\log\frac{P(f_1,f_2,\ldots f_n\mid y=1)}{P(f_1,f_2,\ldots f_n\mid y=0)}$$

- where *y*=1 represents TF cooperativity and *y*=0 represents non-cooperativity. *f*1 through *f*n are different genomic features that are predictive for TF cooperativity.

- $P(y=1|f1, f2,\ldots, fn)$ is the probability that the TF pair is cooperative given these features.
- $P(y=1)/P(y=0)$ is the prior odds.
- $P(f1, f2,\ldots,fn|y=1)/P(f1, f2,\ldots,fn|y=0)$ is the likelihood ratio for the combined features.
- A TF pair is predicted to be cooperative if the calculated posterior odds of cooperativity is greater than a predetermined threshold.

The simplest case: naïve Bayes

- Genomic features are assumed to be conditionally independent given TF cooperativity.
- In this case, the likelihood ratio of the combined features is equal to the product of the likelihood ratios for individual features.
- $P(f1, f2,\ldots,f\mathrm{n}|y=1)/P(f1, f2,\ldots,f\mathrm{n}|y=0)$

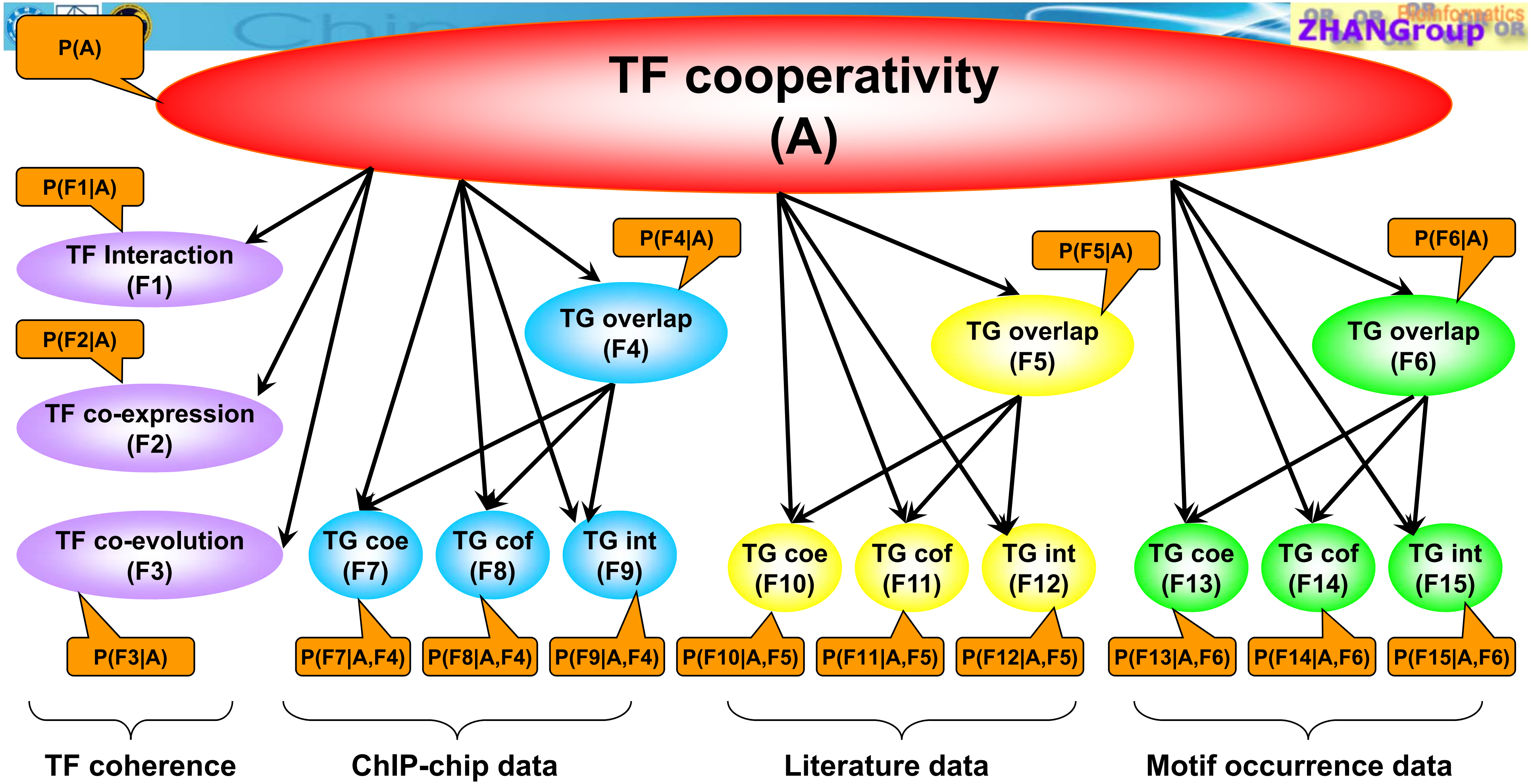
$=P(f1|y=1)/P(f1|y=0)$

$*P(f2|y=1)/P(f2|y=0)$

$*\ldots$

$*P(fn|y=1)/P(fn|y=0)$

- However in our TF cooperativity prediction task, the TG overlap feature and the TG coherence features are not conditionally independent, as they all rely on the target gene information and are thus are partially redundant.
- It is possible to learn the optimal Bayesian network structure from training data, but this problem is hard in terms of computational complexity, and requires a large training data.
- We rely on prior knowledge to determine the Bayesian network structure.



P(A,F1,F2,F3,F5,F5,F6,F7,F8,F9,F10,F11,F12,F13,F14,F15)=

P(A)*P(F1|A)*P(F2|A)*P(F3|A)*

P(F4|A)*P(F7|A,F4)*P(F8|A,F4)*P(F9|A,F4)*

P(F5|A)*P(F10|A,F5)*P(F11|A,F5)*P(F12|A,F5)*

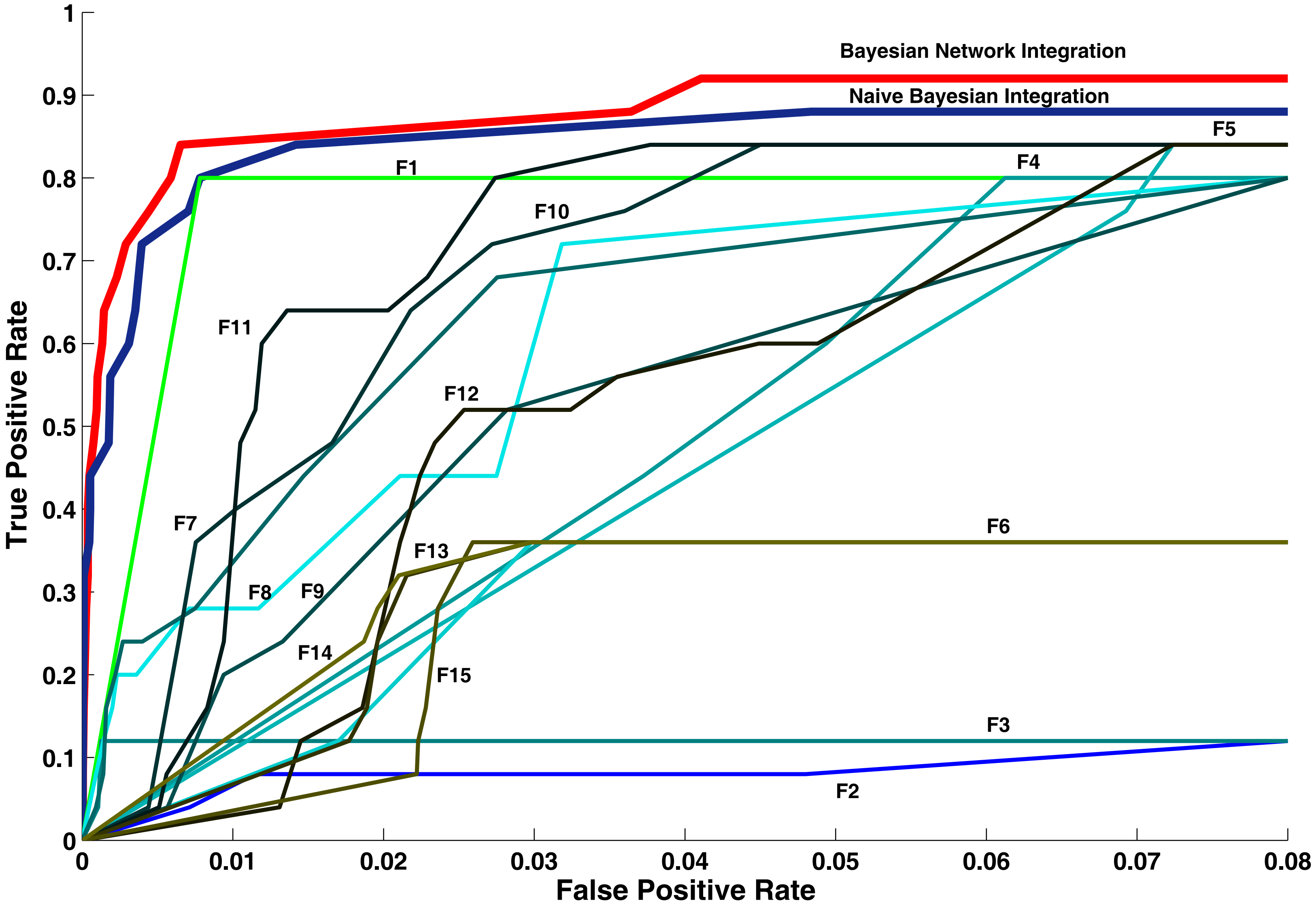
P(F6|A)*P(F13|A,F6)*P(F14|A,F6)*P(F15|A,F6)

Bayesian network structure

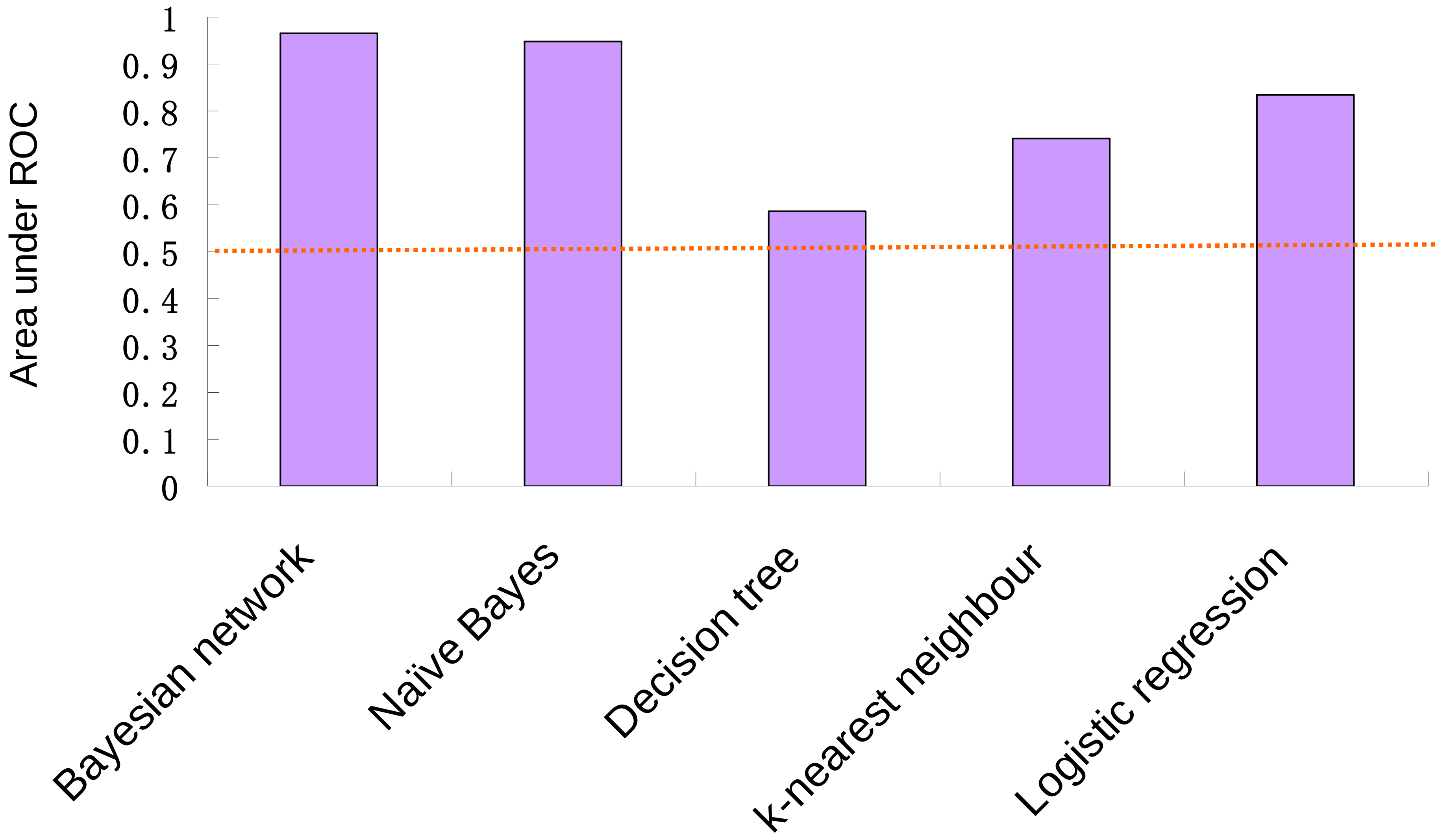
- Guiding rule: the structure should be as simple as possible, i.e., maximize the number of conditional independencies among features, while at the same time still be able to capture the dominant dependencies within data.
- Given the Bayesian network structure, we can determine the posterior odds for every TF pair:

$$\log\frac{P(y=1\mid f_1,f_2,\ldots f_n)}{P(y=0\mid f_1,f_2,\ldots f_n)}=\log\frac{P(y=1)}{P(y=0)}+\sum_{i=1}^{n}\log\frac{P(f_i\mid y=1,S_i)}{P(f_i\mid y=0,S_i)}$$

- Where *Si* is the set of parent features that *fi* conditionally depends upon.



ROC plot comparison of our Bayesian network classifier, naïve Bayes classifier, and 15 individual feature based classifiers.

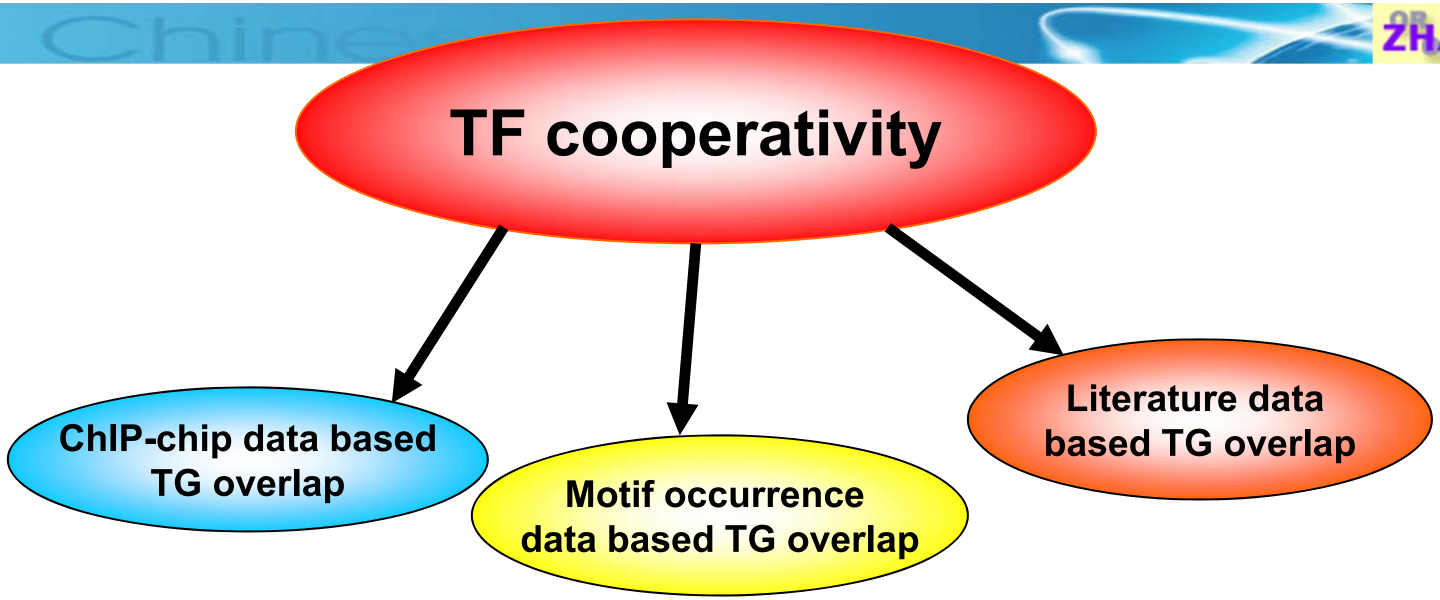


Comparing with other machine learning methods

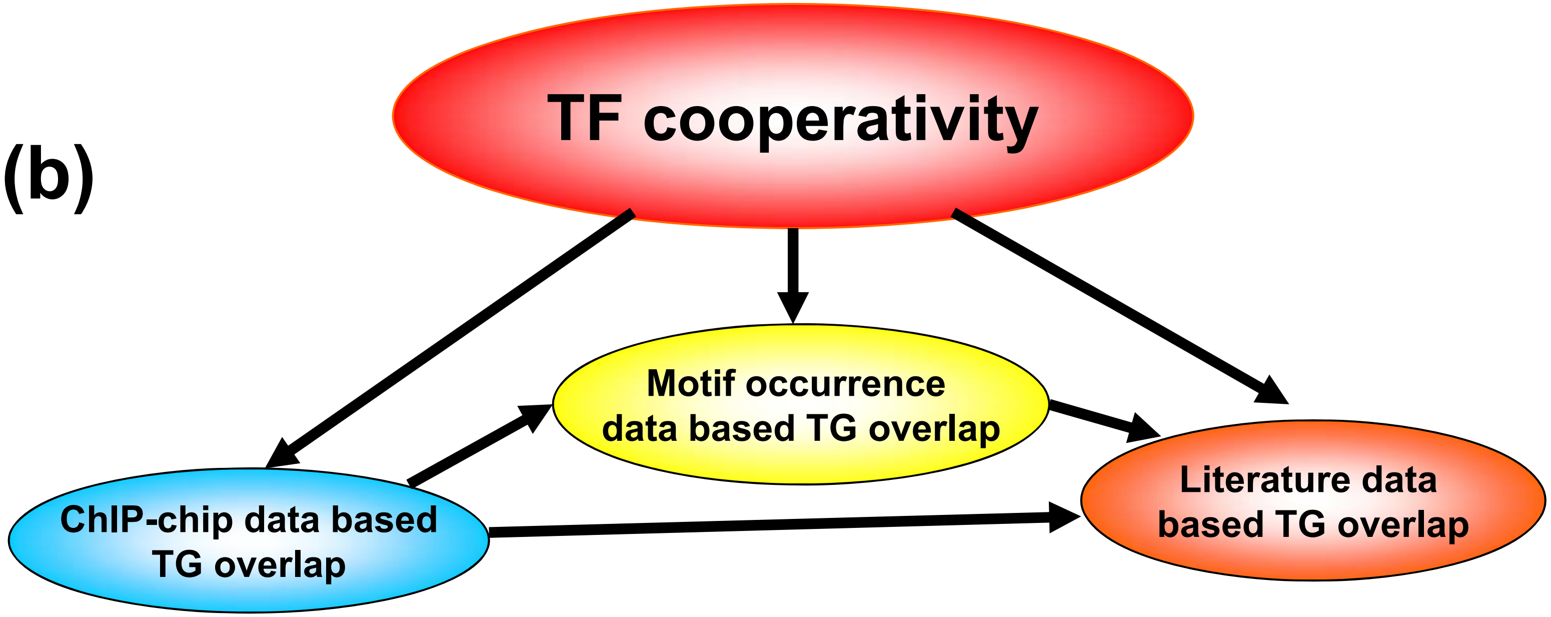
Question: Is it the best structure?

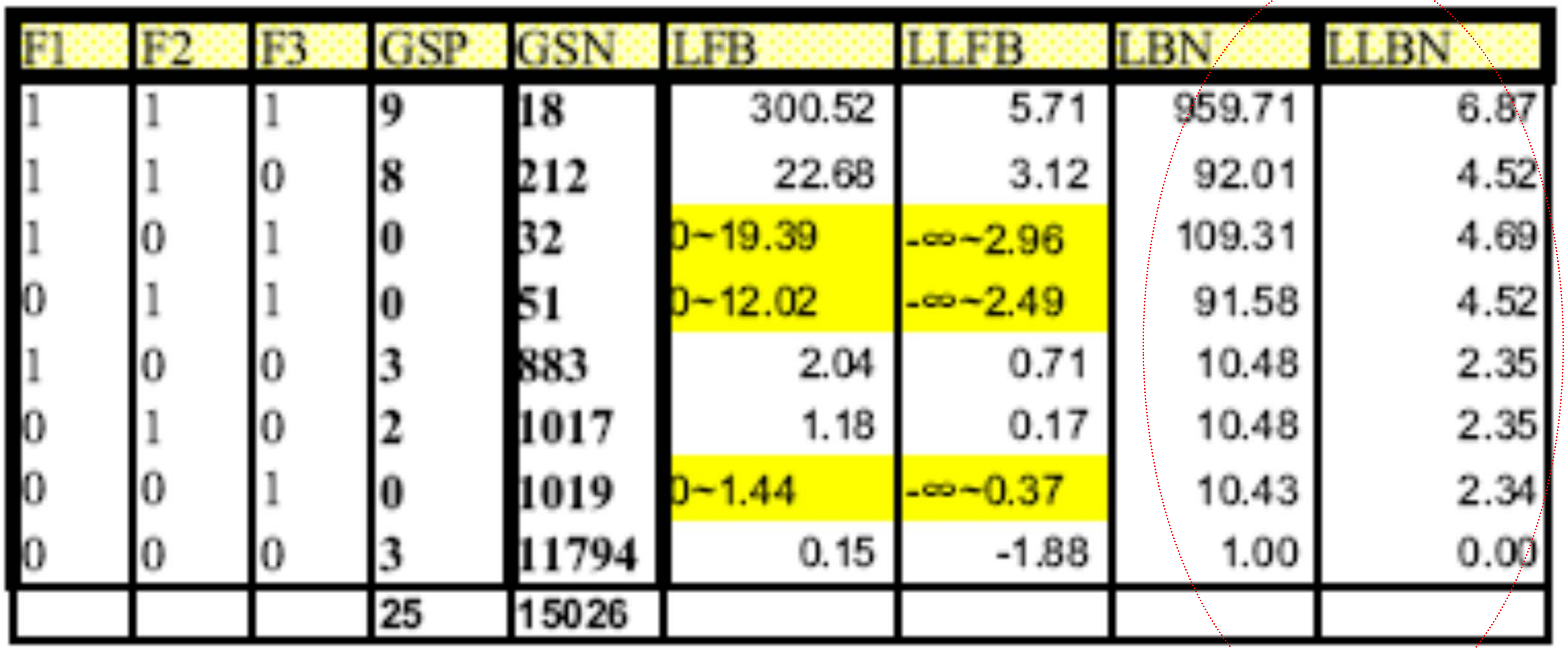
- The Bayesian network architecture we proposed is the simplest that still captures the important relationships between different features for integrated prediction.
- We used the contingency tables to carefully compare the substructures of our Bayesian network with other possible substructures

(a)


TF cooperativity
ChIP-chip data based TG overlap
Motif occurrence data based TG overlap
Literature data based TG overlap

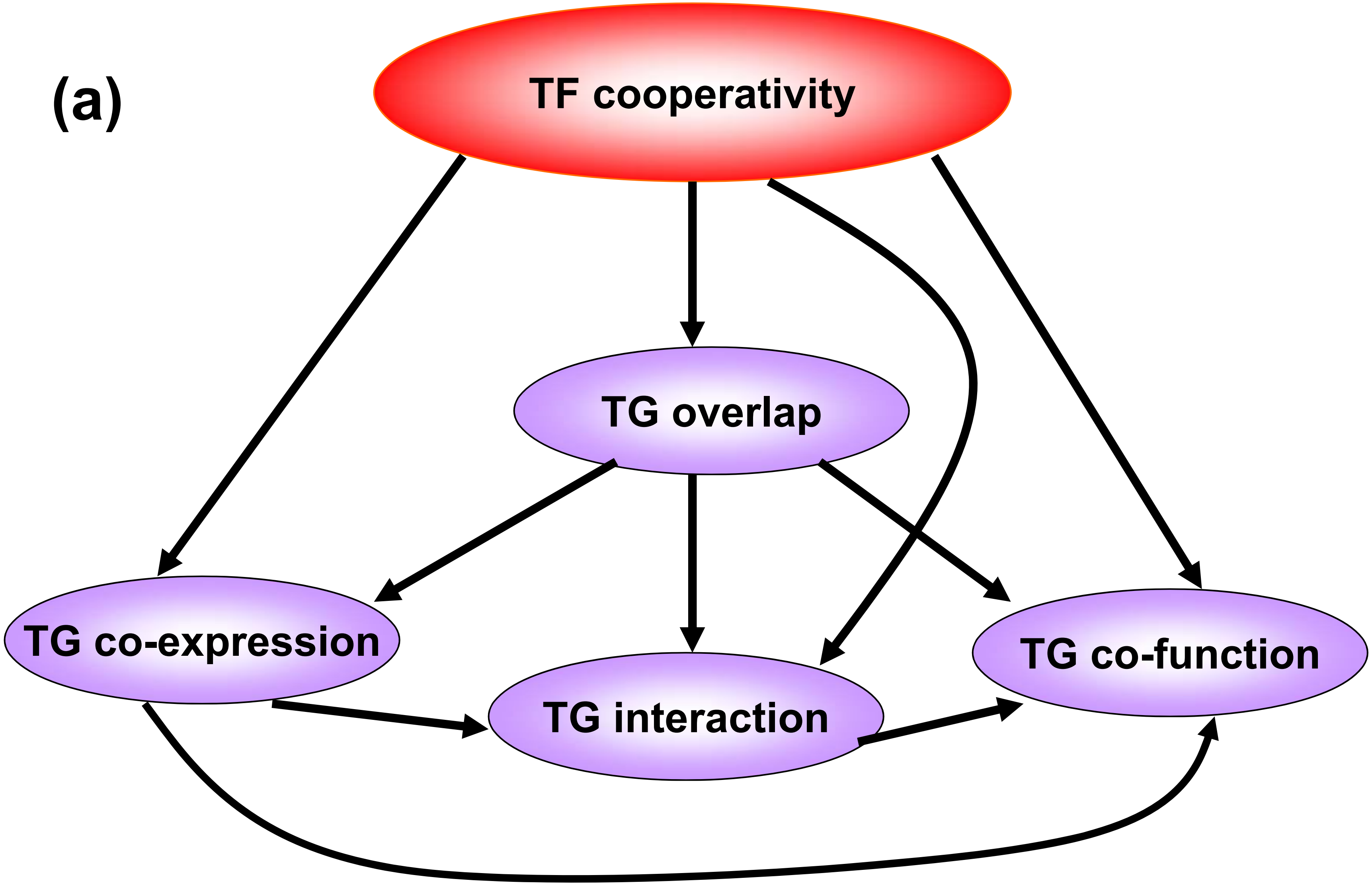
(b)

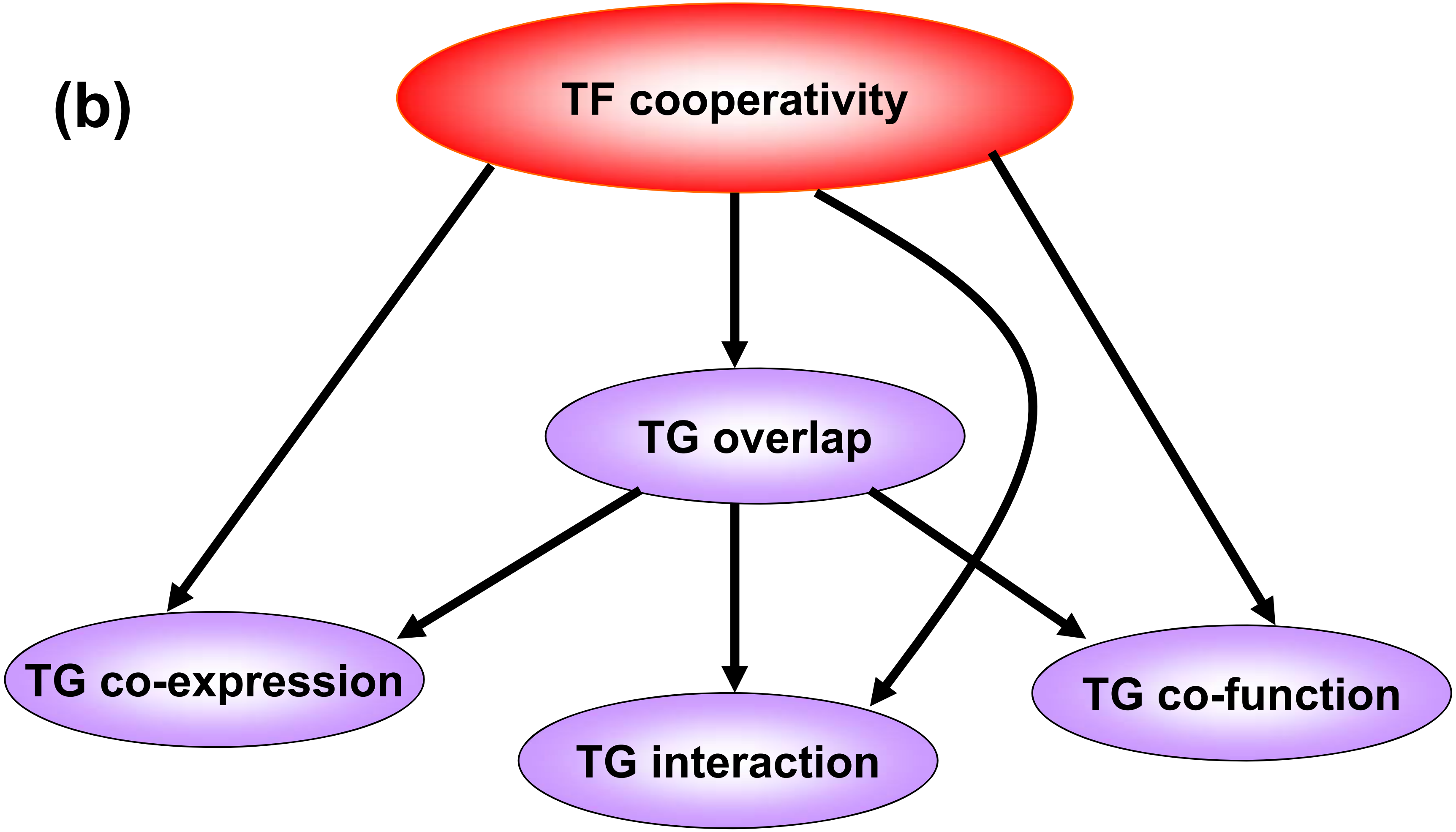

TF cooperativity
Motif occurrence data based TG overlap
ChIP-chip data based TG overlap
Literature data based TG overlap



F1	F2	F3	GSP	GSN	LFB	LLFB	LBN	LLBN
1	1	1	9	18	300.52	5.71	959.71	6.87
1	1	0	8	212	22.68	3.12	92.01	4.52
1	0	1	0	32	0~19.39	-∞~2.96	109.31	4.69
0	1	1	0	51	0~12.02	-∞~2.49	91.58	4.52
1	0	0	3	883	2.04	0.71	10.48	2.35
0	1	0	2	1017	1.18	0.17	10.48	2.35
0	0	1	0	1019	0~1.44	-∞~0.37	10.43	2.34
0	0	0	3	11794	0.15	-1.88	1.00	0.00
			25	15026				

The highlighted yellow region means this parameter cannot be accurately estimated by full Bayesian network due to scarcity of gold-standard positive data. Instead we give an estimated interval.


(a)
TF cooperativity
TG overlap
TG co-expression
TG interaction
TG co-function



(b)

TF cooperativity

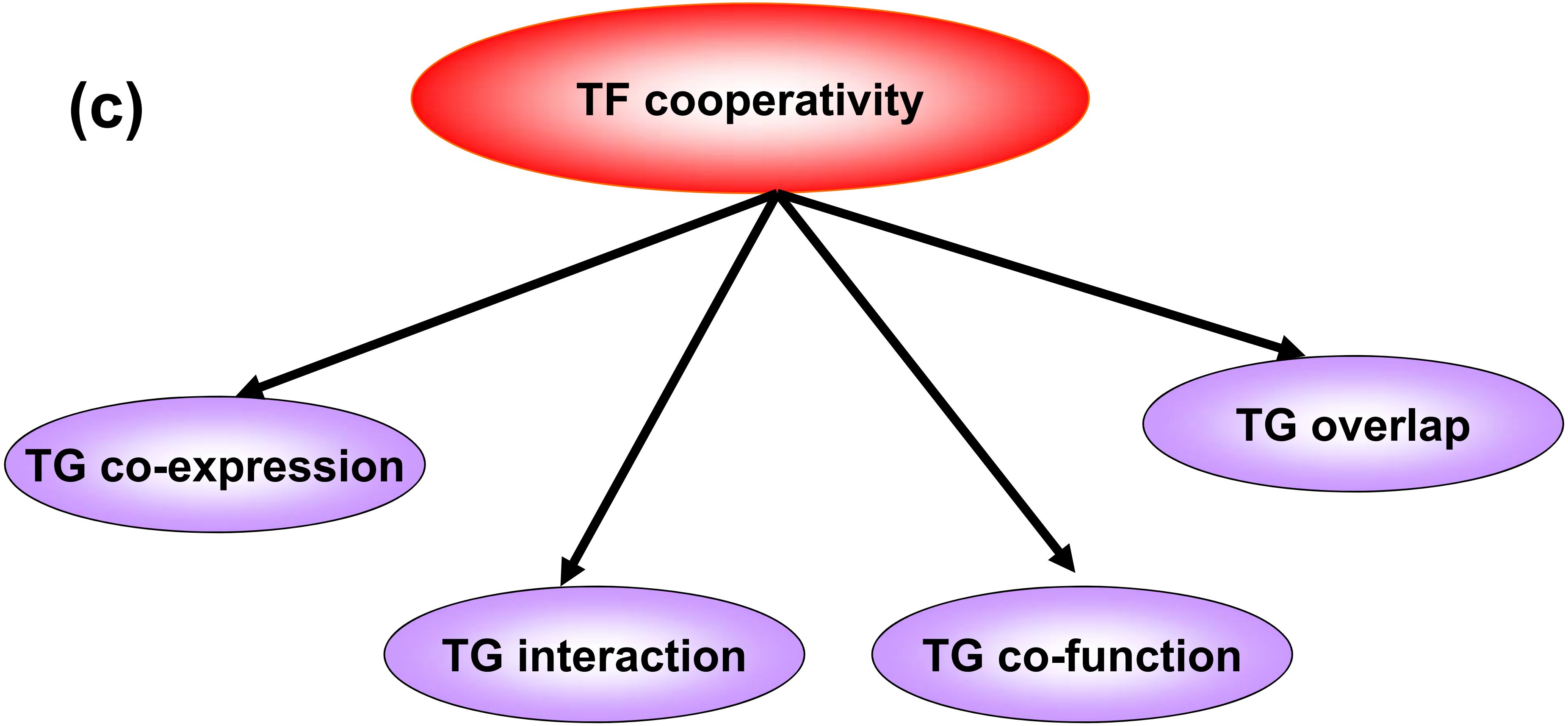
TG overlap

TG co-expression

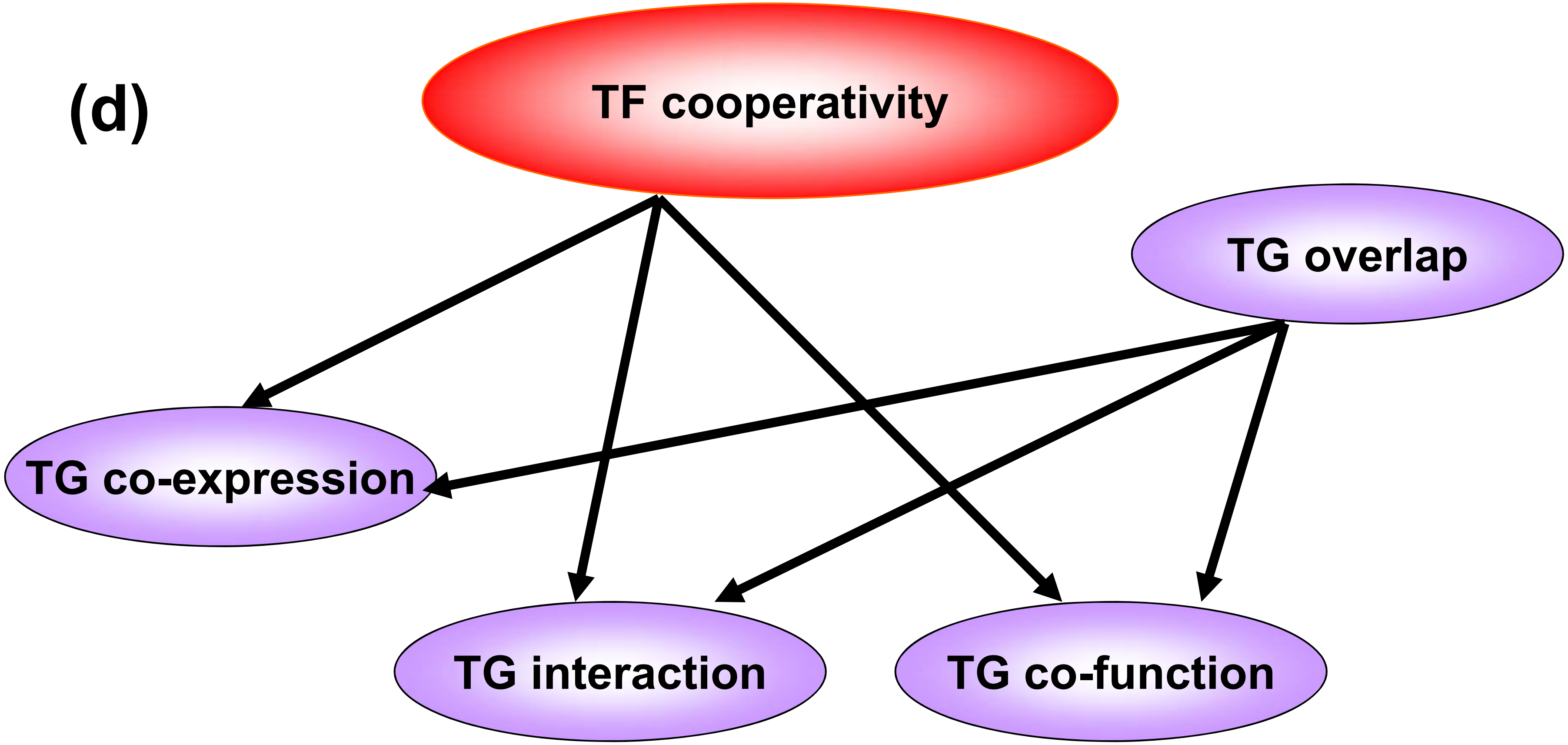
TG interaction

TG co-function

(c)


TF cooperativity
TG co-expression
TG interaction
TG co-function
TG overlap

(d)



TF cooperativity

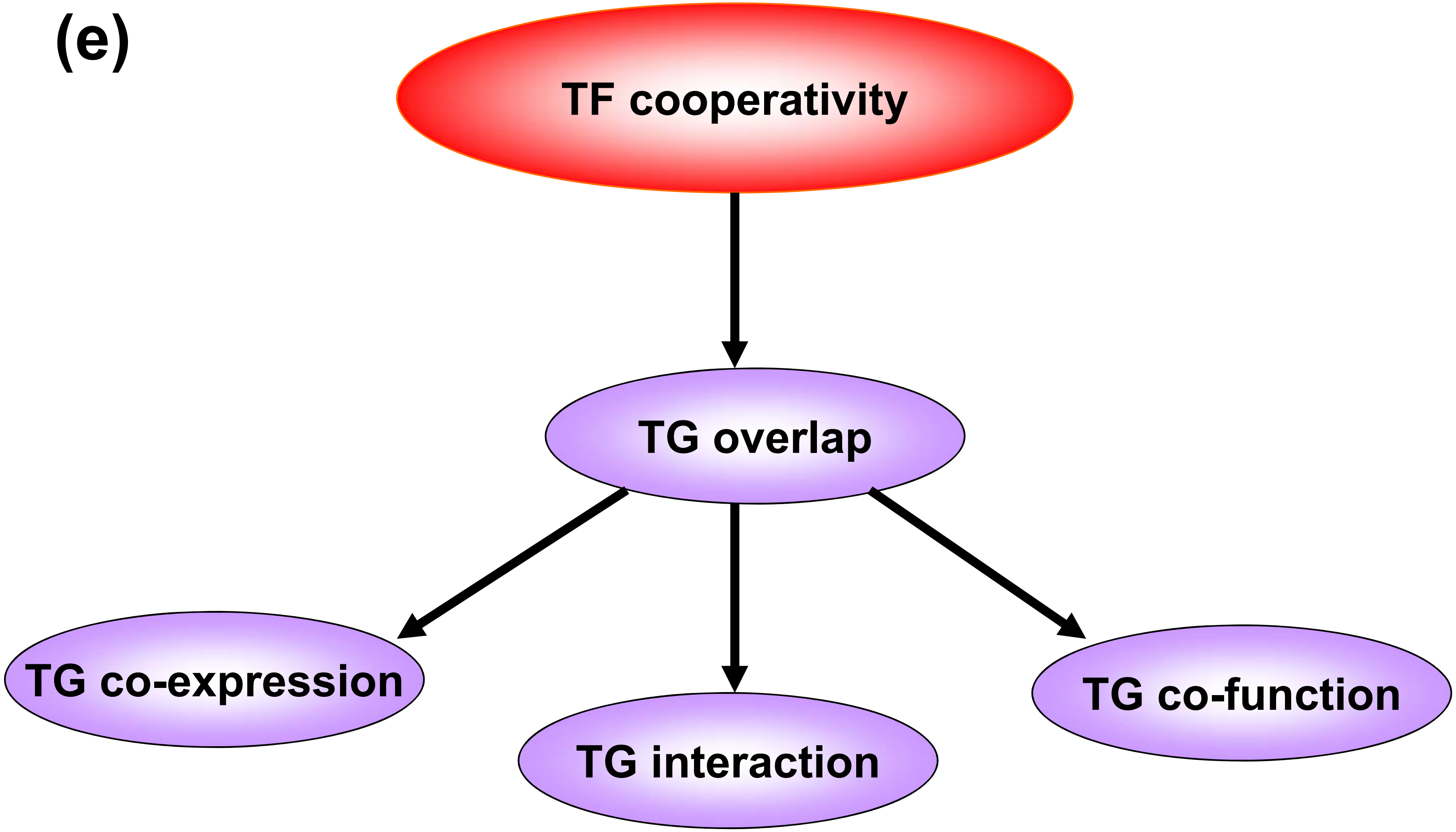
TG overlap

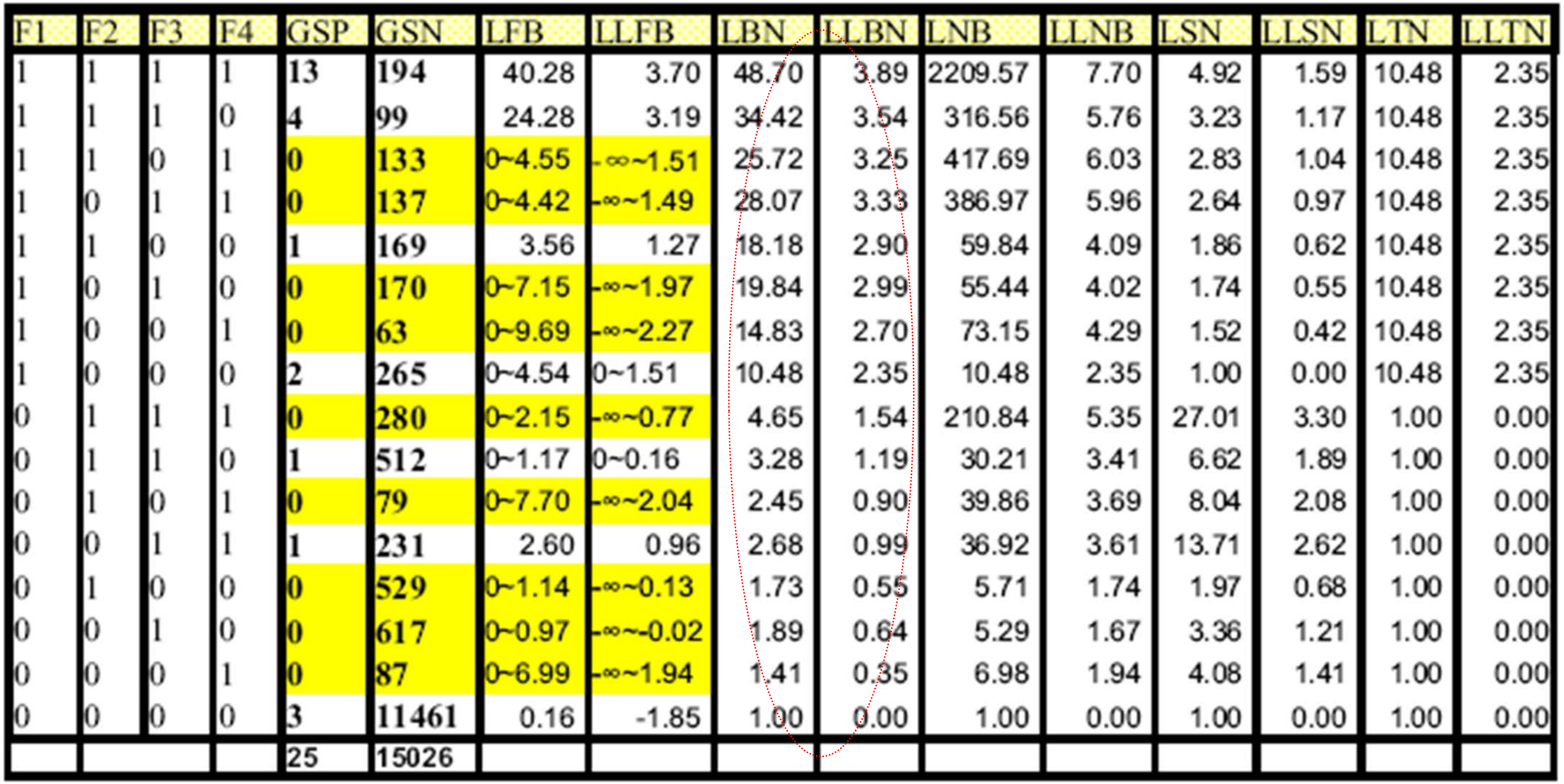
TG co-expression

TG interaction

TG co-function

(e)


TF cooperativity
TG overlap
TG co-expression
TG interaction
TG co-function



F1	F2	F3	F4	GSP	GSN	LFB	LLFB	LBN	LLBN	LNB	LLNB	LSN	LLSN	LTN	LLTN
1	1	1	1	13	194	40.28	3.70	48.70	3.89	2209.57	7.70	4.92	1.59	10.48	2.35
1	1	1	0	4	99	24.28	3.19	34.42	3.54	316.56	5.76	3.23	1.17	10.48	2.35
1	1	0	1	0	133	0~4.55	-∞~1.51	25.72	3.25	417.69	6.03	2.83	1.04	10.48	2.35
1	0	1	1	0	137	0~4.42	-∞~1.49	28.07	3.33	386.97	5.96	2.64	0.97	10.48	2.35
1	1	0	0	1	169	3.56	1.27	18.18	2.90	59.84	4.09	1.86	0.62	10.48	2.35
1	0	1	0	0	170	0~7.15	-∞~1.97	19.84	2.99	55.44	4.02	1.74	0.55	10.48	2.35
1	0	0	1	0	63	0~9.69	-∞~2.27	14.83	2.70	73.15	4.29	1.52	0.42	10.48	2.35
1	0	0	0	2	265	0~4.54	0~1.51	10.48	2.35	10.48	2.35	1.00	0.00	10.48	2.35
0	1	1	1	0	280	0~2.15	-∞~0.77	4.65	1.54	210.84	5.35	27.01	3.30	1.00	0.00
0	1	1	0	1	512	0~1.17	0~0.16	3.28	1.19	30.21	3.41	6.62	1.89	1.00	0.00
0	1	0	1	0	79	0~7.70	-∞~2.04	2.45	0.90	39.86	3.69	8.04	2.08	1.00	0.00
0	0	1	1	1	231	2.60	0.96	2.68	0.99	36.92	3.61	13.71	2.62	1.00	0.00
0	1	0	0	0	529	0~1.14	-∞~0.13	1.73	0.55	5.71	1.74	1.97	0.68	1.00	0.00
0	0	1	0	0	617	0~0.97	-∞~-0.02	1.89	0.64	5.29	1.67	3.36	1.21	1.00	0.00
0	0	0	1	0	87	0~6.99	-∞~1.94	1.41	0.35	6.98	1.94	4.08	1.41	1.00	0.00
0	0	0	0	3	11461	0.16	-1.85	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
				25	15026										

Comparison with other methods

- Five existing methods.

(1) Jansen et al. used Bayesian networks to predict yeast protein-protein interactions in general

(2) Datta et al. used log-linear models to predict cooperative binding among cell cycle specific TFs.

(3) Banerjee et al. integrated genome-wide location data from ChIP-chip and gene expression data to infer cooperative TF pairs

(4) Tsai et al. used statistical methods (ANOVA) to identify synergistic pairs of yeast cell cycle TFs

(5) Balaji et al. used a specific network transformation procedure to obtain a co-regulatory network

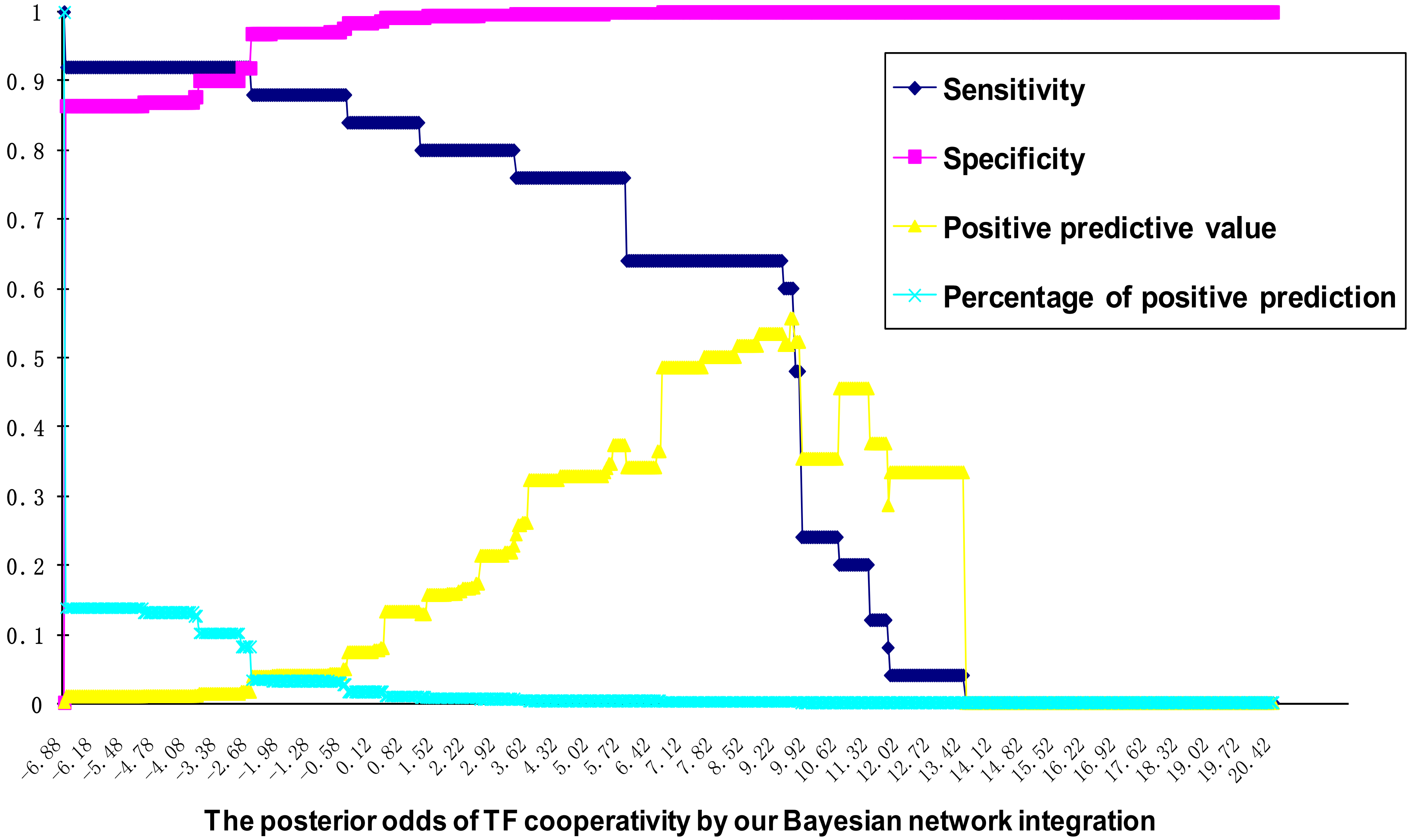
Two independent benchmark datasets

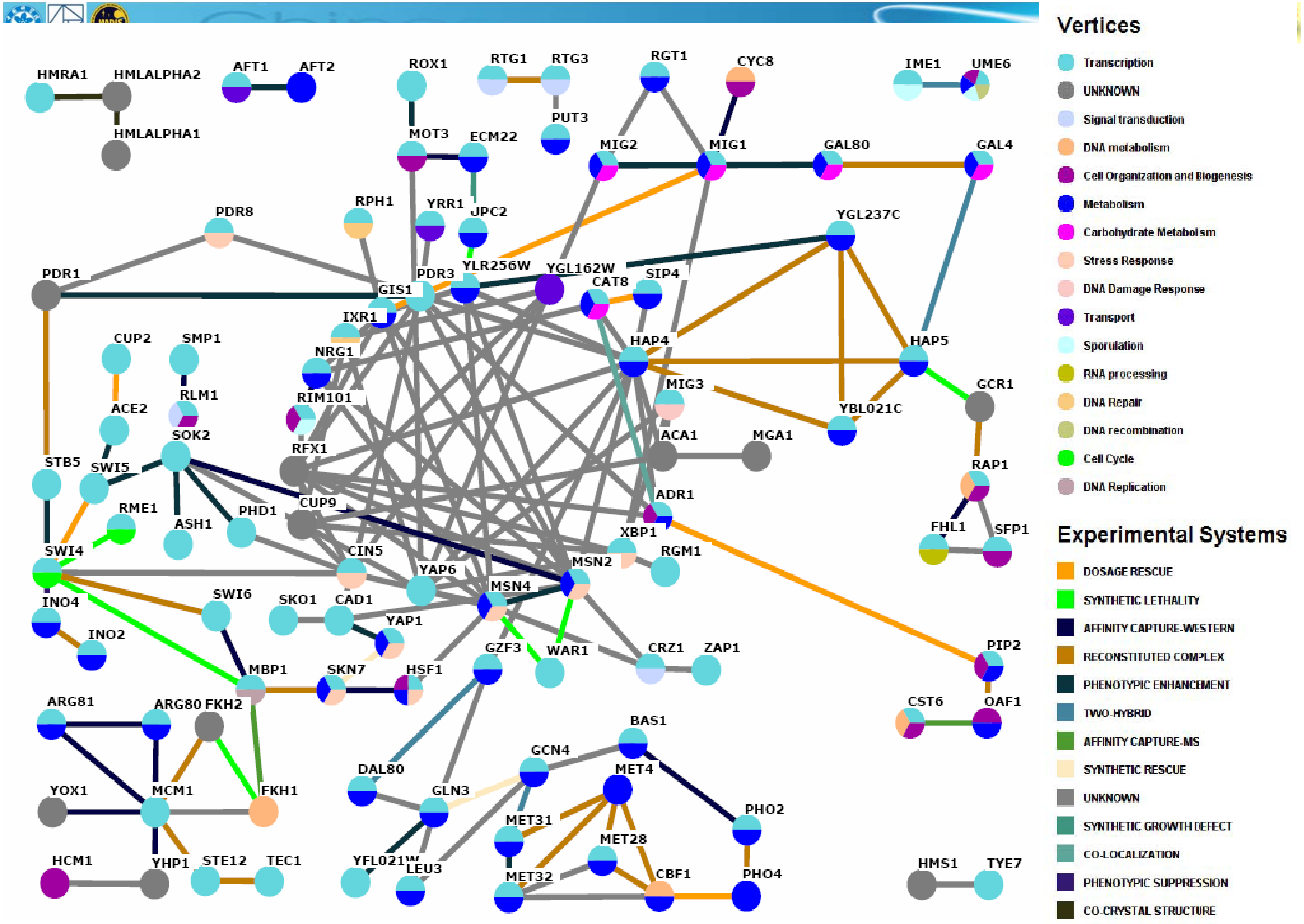
- The first benchmark dataset is based on the KEGG pathway database, and contains 48 TF pairs among 13 TFs that co-occur in at least one KEGG pathway.
- The second benchmark dataset is based on the recently published high-quality experimental binary protein-protein interaction map in yeast (CCSB-YI1) by Yu et al. , and contains 17 interacting TF pairs among 24 TFs.

Benchmark Dataset		Our method	Datta et al. [19]	Banerjee et al. [20]	Tsai et al. [22] (Doubtful)	Tsai et al. [22] (Plausible)	Tsai et al. [22] (Confident)	Balaji et al. [9] (All)	Balaji et al. [9] (Core)	Jansen et al. [26]
KEGG pathway database [63] (13 TFs, 48 TF pairs)	# of overlapping TFs	13	4	8	7	8	6	13	13	7
	# of possible interactions among overlapping TFs	78	6	28	21	28	15	78	78	21
	# of KEGG interactions among overlapping TFs	48	6	20	15	18	9	48	48	13
	# of predicted interactions among overlapping TFs	8	3	8	3	1	2	69	48	2
	# of KEGG interactions that are correctly predicted	8	3	5	3	1	1	45	33	2
	Fisher's exact test p-value	**0.016**	**1.0**	**0.87**	**0.34**	**0.64**	**0.86**	**0.071**	**0.079**	**0.37**
CCSB-YI1 dataset [64] (24 TFs, 17 TF pairs)	# of overlapping TFs	20	2	11	2	3	1	18	18	8
	# of possible interactions among overlapping TFs	190	1	55	1	3	0	153	153	28
	# of CCSB-YI1 interactions among overlapping TFs	13	0	2	0	0	0	12	12	3
	# of predicted interactions among overlapping TFs	5	1	2	0	1	0	91	50	2
	# of CCSB-YI1 interactions that are correctly predicted	5	0	0	0	0	0	3	3	1
	Fisher's exact test p-value	**6.6×10^{-7}**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**0.82**	**0.21**

Biological results

Choosing cutoff for final prediction

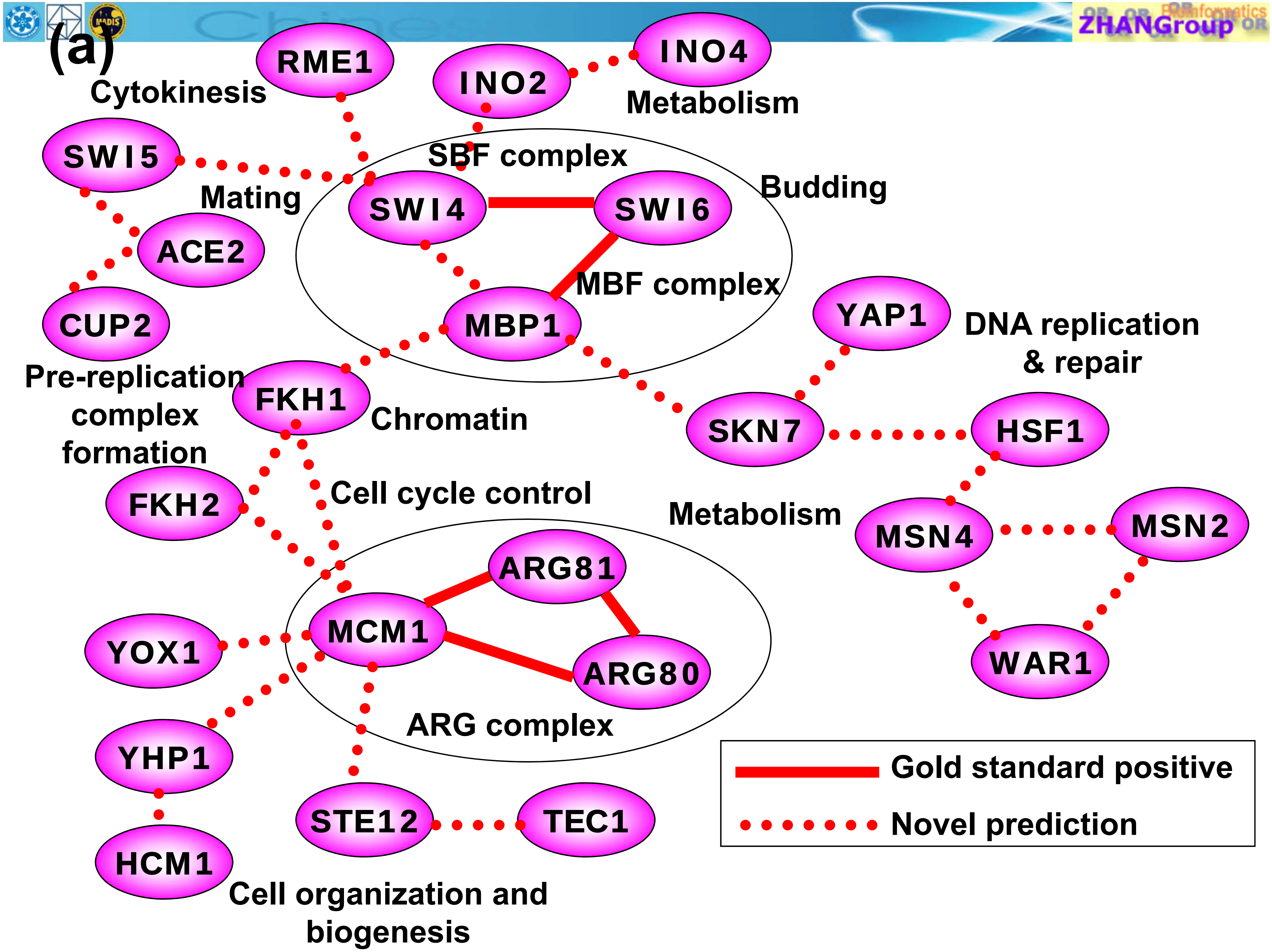




The genome wide TF cooperation network(174 nodes and 159 edges)

Validation of novel TF cooperativity predictions

- Structural evidence in PDB
- We then manually curated TF cooperativity information from PubMed abstracts and found that most of the 159 predicted TF cooperativity relationships are supported by one or more published literatures (143 out of 159 are supported by literature evidence including 21 gold-standard positives).
- The extensive literature validation demonstrates the overall high quality of the prediction results



(a)

RME1

INO2

INO4

Cytokinesis

Metabolism

SWI5

SBF complex

Mating

SWI4

SWI6

Budding

ACE2

MBF complex

CUP2

MBP1

YAP1

DNA replication & repair

Pre-replication complex formation

FKH1

Chromatin

SKN7

HSF1

FKH2

Cell cycle control

Metabolism

MSN4

MSN2

ARG81

MCM1

YOX1

ARG80

WAR1

ARG complex

YHP1

STE12

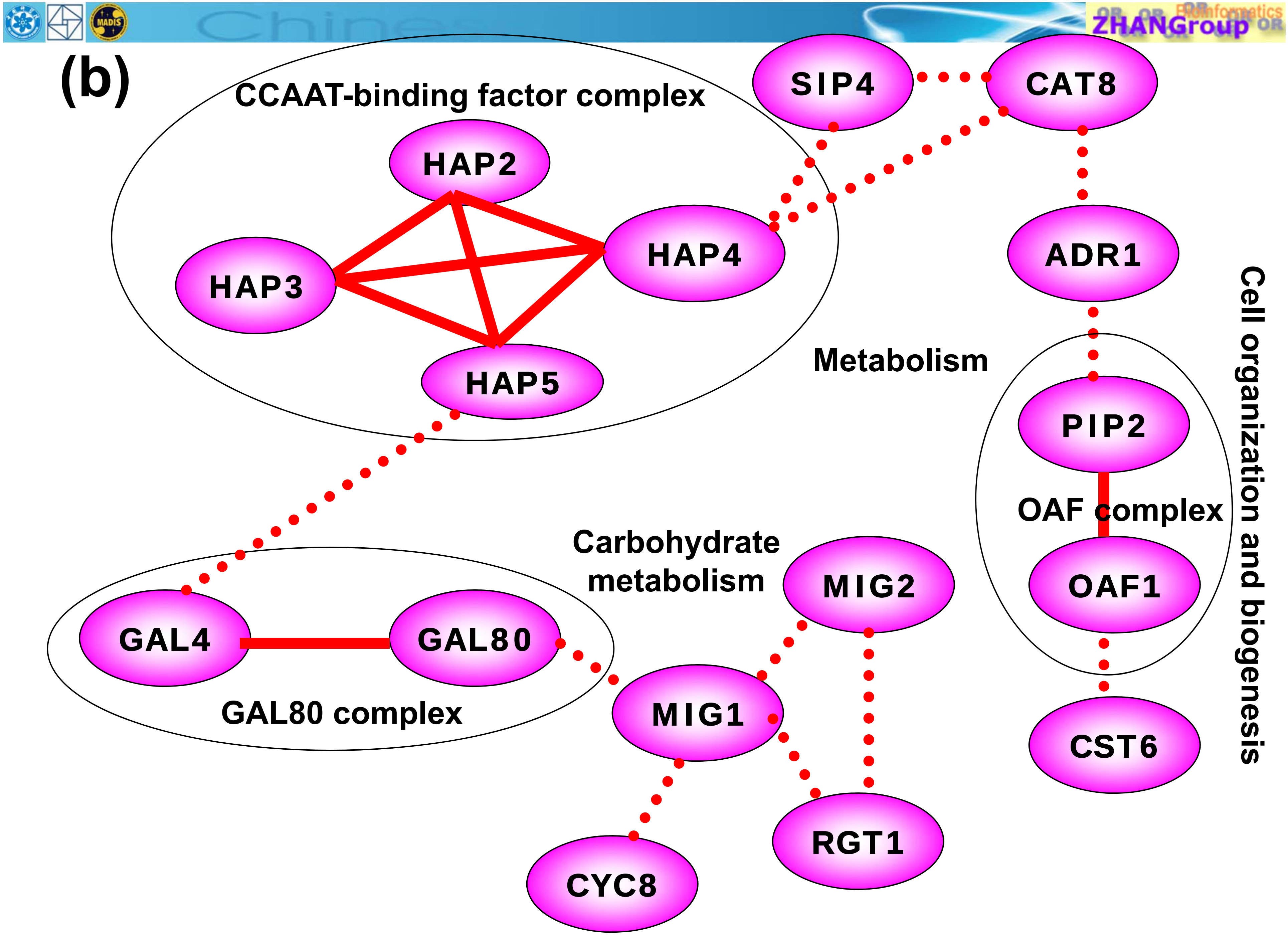
TEC1

HCM1

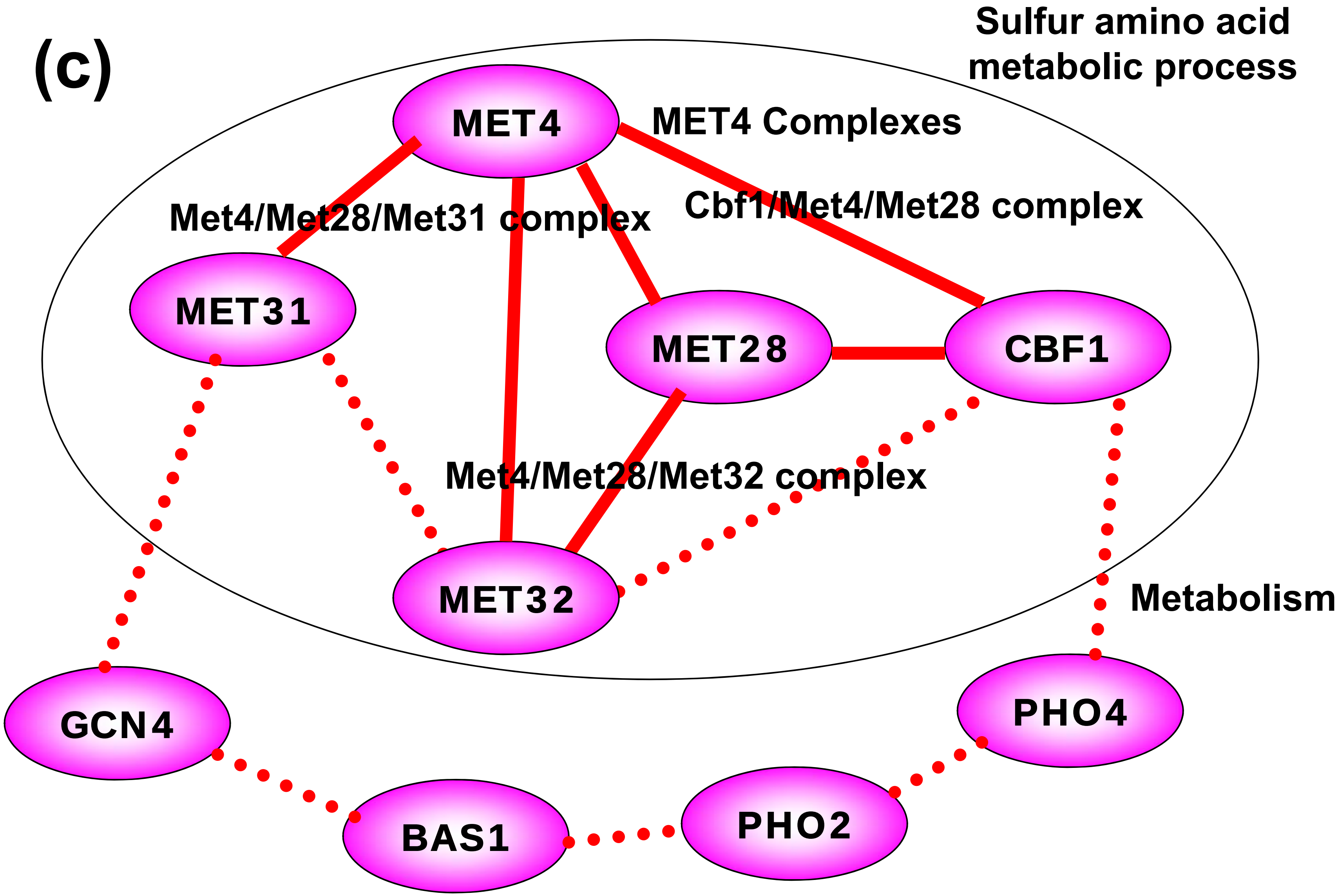
Cell organization and biogenesis

▬▬ Gold standard positive

• • • • • • • Novel prediction


(b)
CCAAT-binding factor complex
SIP4
CAT8
HAP2
HAP4
HAP3
HAP5
ADR1
Metabolism
PIP2
OAF complex
OAF1
CST6
Cell organization and biogenesis
Carbohydrate metabolism
MIG2
GAL4
GAL80
GAL80 complex
MIG1
RGT1
CYC8

(c)



Sulfur amino acid metabolic process

MET4 Complexes

Met4/Met28/Met31 complex

Cbf1/Met4/Met28 complex

Met4/Met28/Met32 complex

Metabolism

MET4

MET31

MET28

CBF1

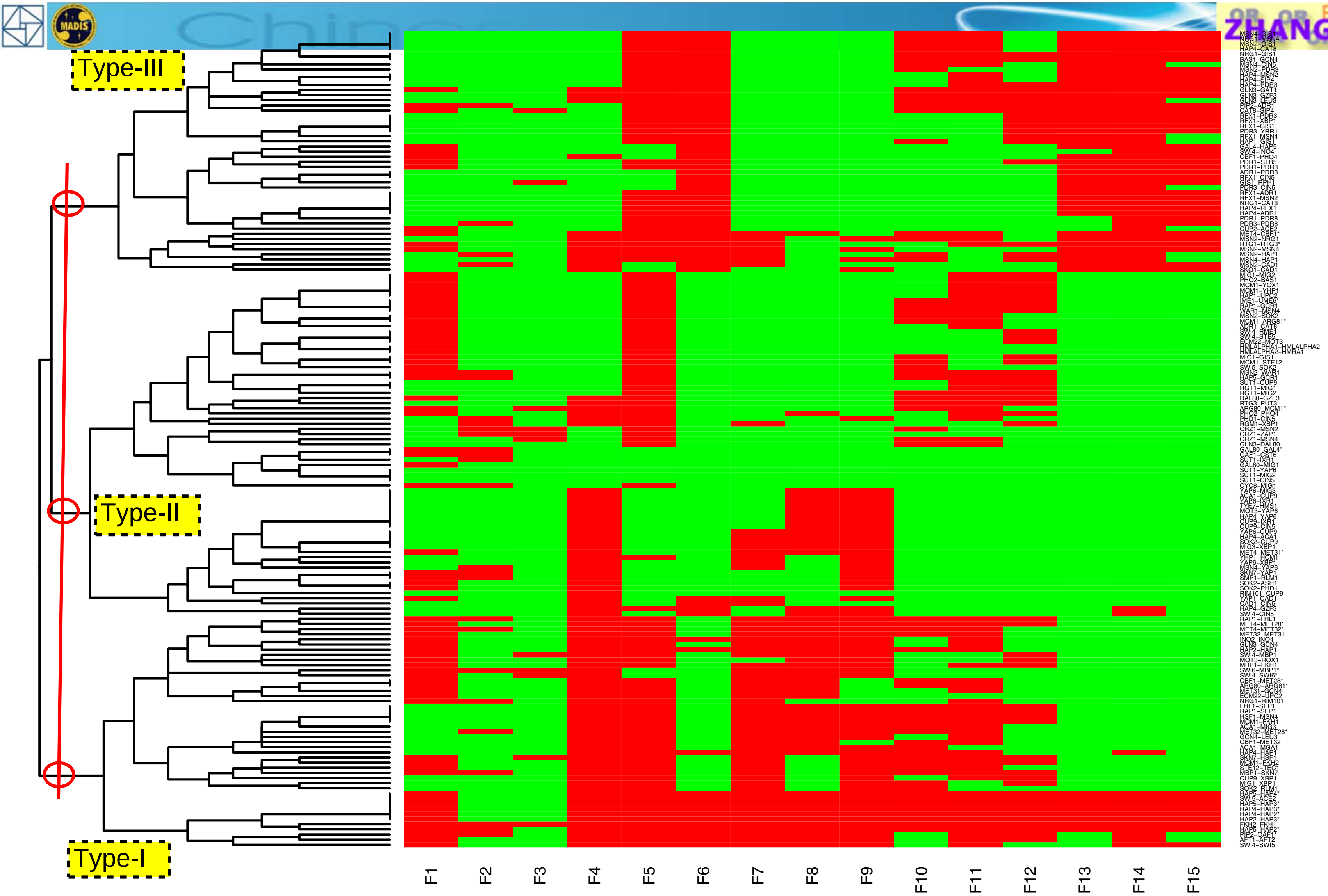
MET32

GCN4

PHO4

BAS1

PHO2



The heatmap for feature profiles of the predicted 159 TF cooperative relationships

Columns represent genomic features

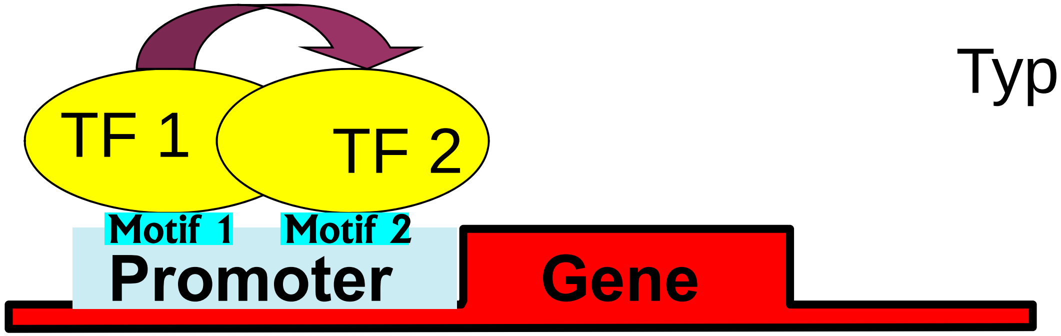
Rows represent predicted cooperative TF pairs.

Red if the feature is present (f = 1), or green if the feature is absent (f = 0)

Physical interaction

(a)

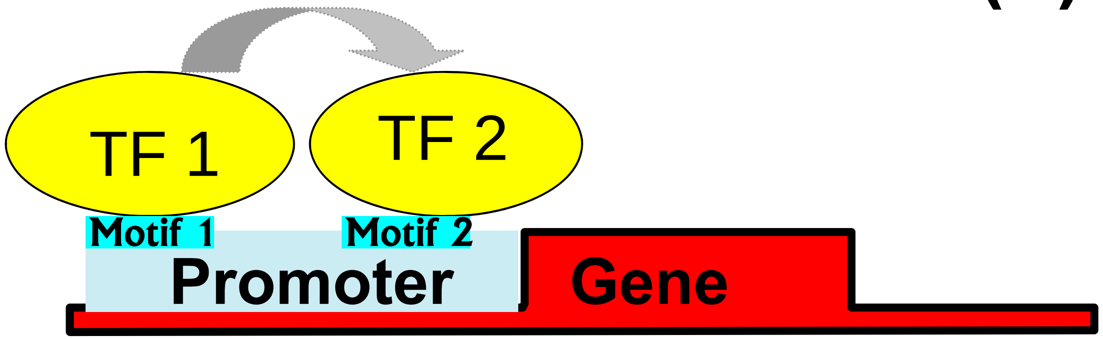
Type-I cooperativity



Regulatory or genetic interaction

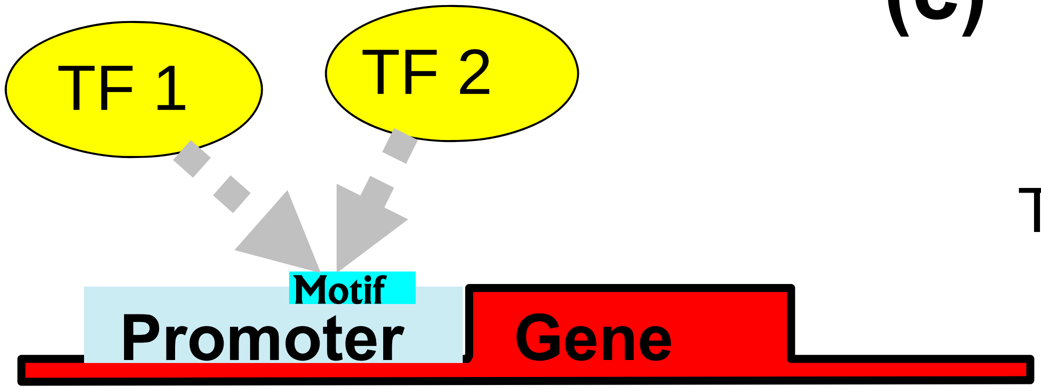
(b)

Type-II cooperativity



Competitive cooperation

(c)



Type-III cooperativity

Conclusion

- Three machine learning ideas for the first time into the prediction of transcriptional cooperativity.
- First, we introduced a small set of well-constructed gold-standard dataset, and emphasized its central role in our data integration framework.
- Second, we used graphical models such as Bayesian networks to capture the casual relationships among genomic features. This framework of transparent data integration is especially important for our case, where the gold-standard data is scarce.
- Third, our Bayesian network structure is pre-chosen by carefully considering the optimal trade-off between predictive bias and variance, and we only need to learn Bayesian network parameters during training.
- In general, our methodology can be applied to other genomic data integration tasks where high-quality gold-standard positive data are scarce.