Towards Modeling the Gene Regulation Problem

From static to dynamic … from bacterial to human models

SCI2S Group

Departamento de Ciencias de la Computación e Inteligencia Artificial

Escuela Técnica Superior de Ingeniería Informática

Universidad de Granada, Granada, Spain

Computational Biology Group

Departamento de Ciencias de la Computación

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires, Buenos Aires, Argentina

Groisman Lab

Microbiology Department

Howard Hughes Medical Institute

Washington University School of Medicine

St. Louis, Missouri USA

Cobb Lab

Cellular Injury and Adaptation Laboratory

Washington University in St. Louis

St. Louis, Missouri USA

AN EXPERIMENTAL-ORIENTED APPROACH TO BIOINFORMATICS

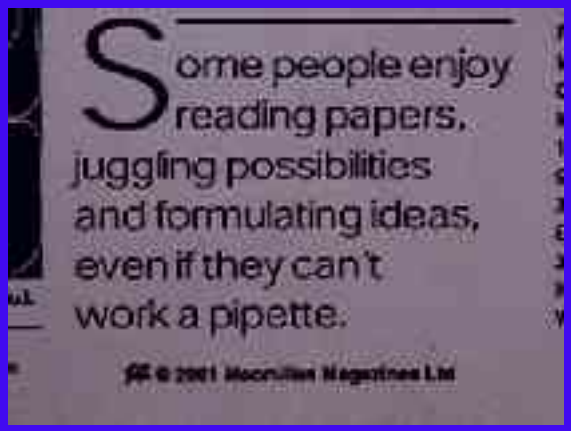

Reasoning for results

Theoretical biology


Así funciona la mente

El ingeniero informático de Silicon Valley Jeff Hawkins propone una teoría revolucionaria capaz de explicar el milagro de la inteligencia humana

Así funciona el córtex cerebral


Some people enjoy reading papers, juggling possibilities and formulating ideas, even if they can't work a pipette.

© 2001 Macmillan Magazines Ltd


EL PAIS

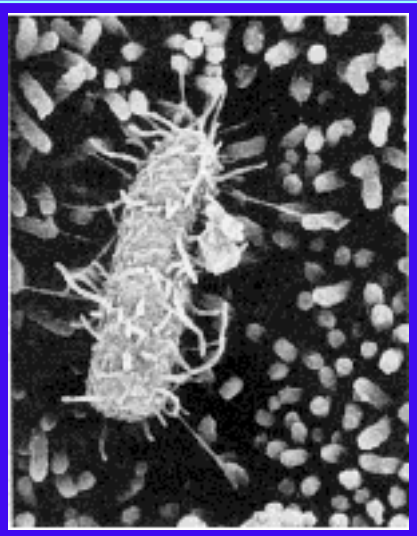
Genotype

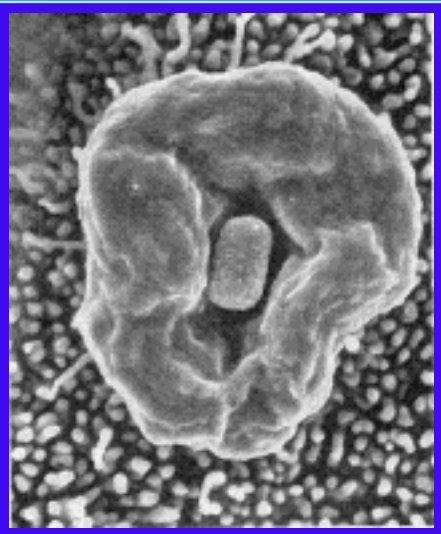
“The analysis of genome sequences gives us a comprehensive protein parts list…But there is one important piece of information that is almost totally missing: the sequence information that specifies when and where and for how long a gene is turned on or off.”

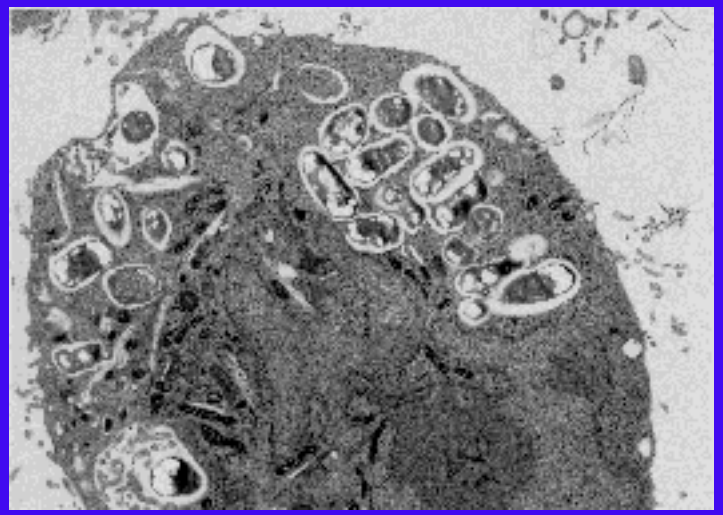
Sydney Brenner
Science 287: 2173

Phenotype

SALMONELLA: A GRAM-NEGATIVE PATHOGEN WITH A VARIED LIFESTYLE

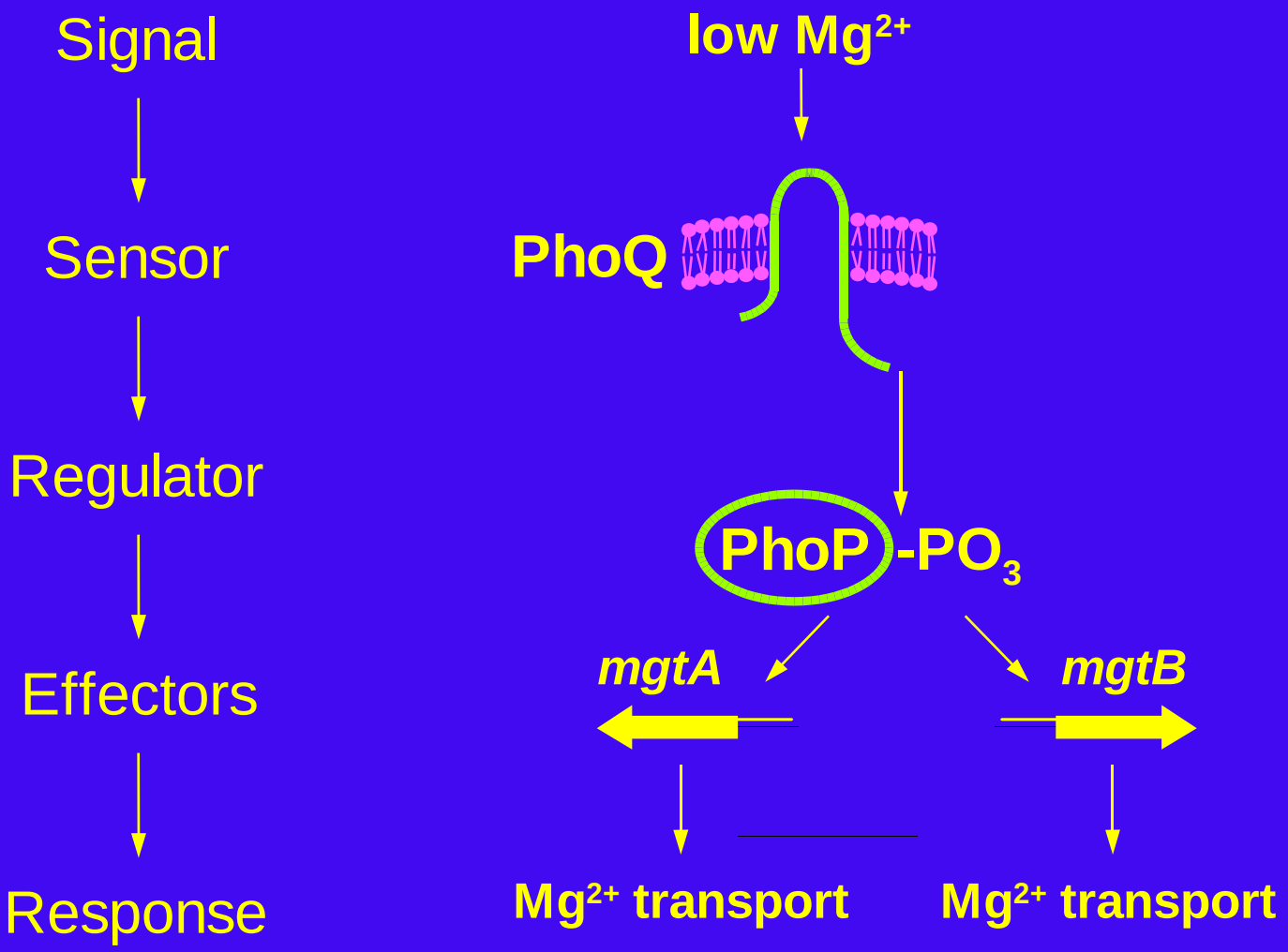




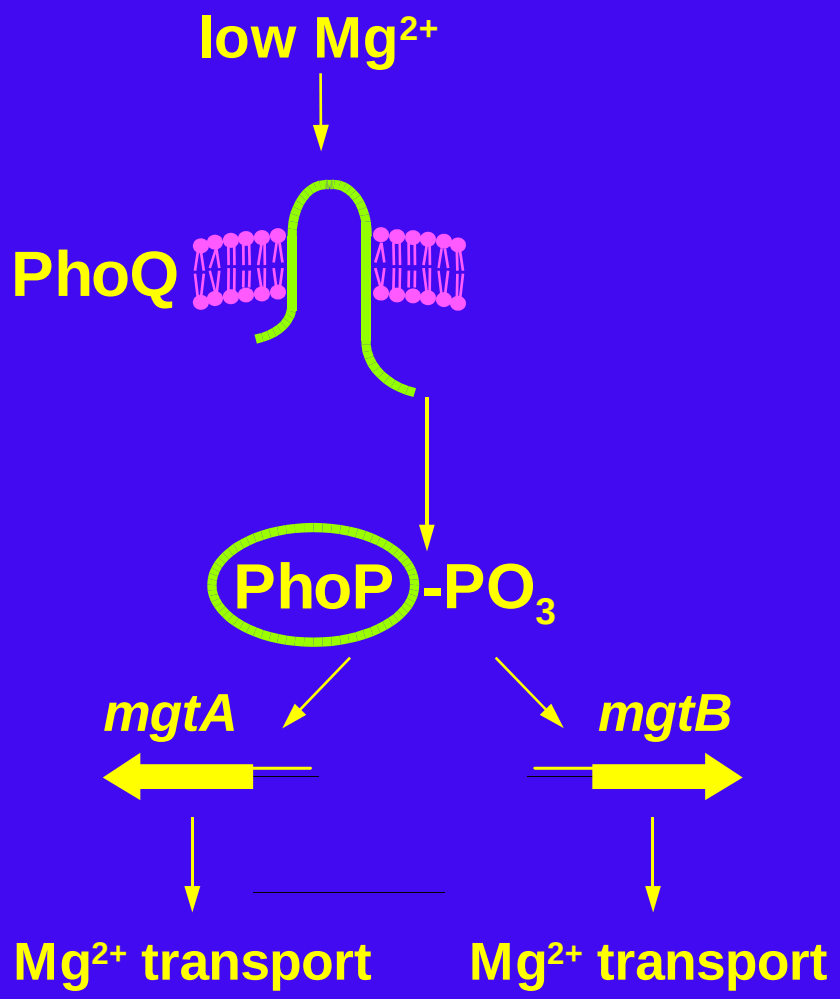


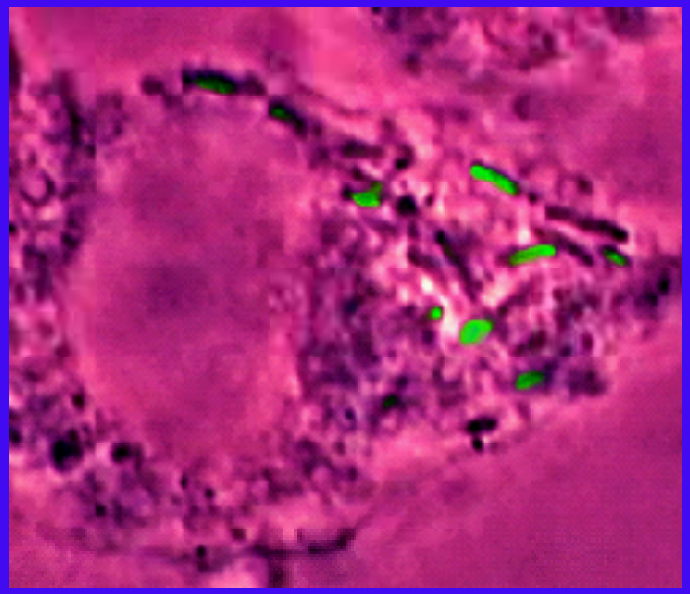
SIGNAL TRANSDUCTION CASCADE

BY TWO-COMPONENT REGULATORY SYSTEMS

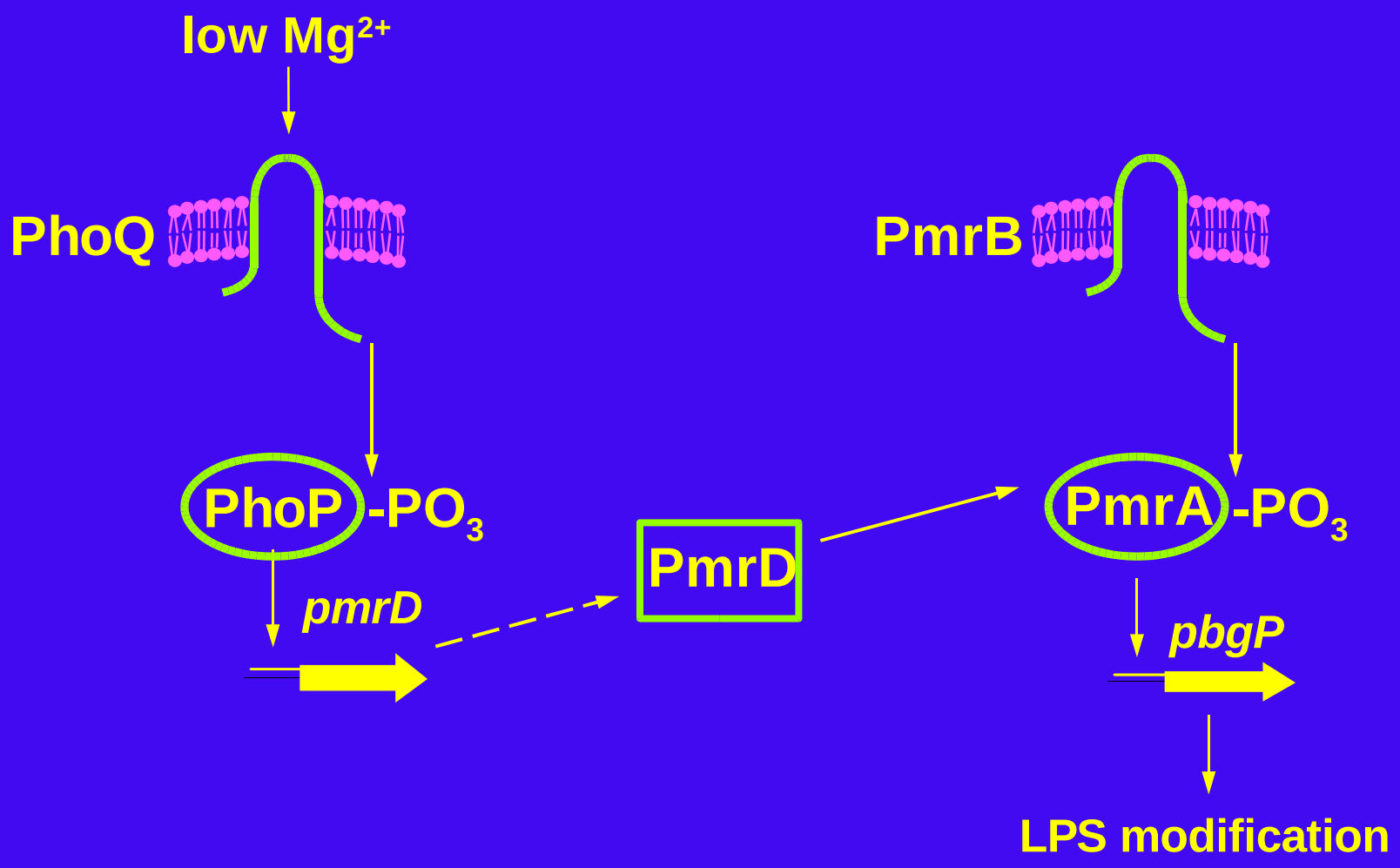


THE PHOP/PHOQ REGULATORY SYSTEM IS ACTIVATED IN LOW Mg^{2+} ENVIRONMENTS

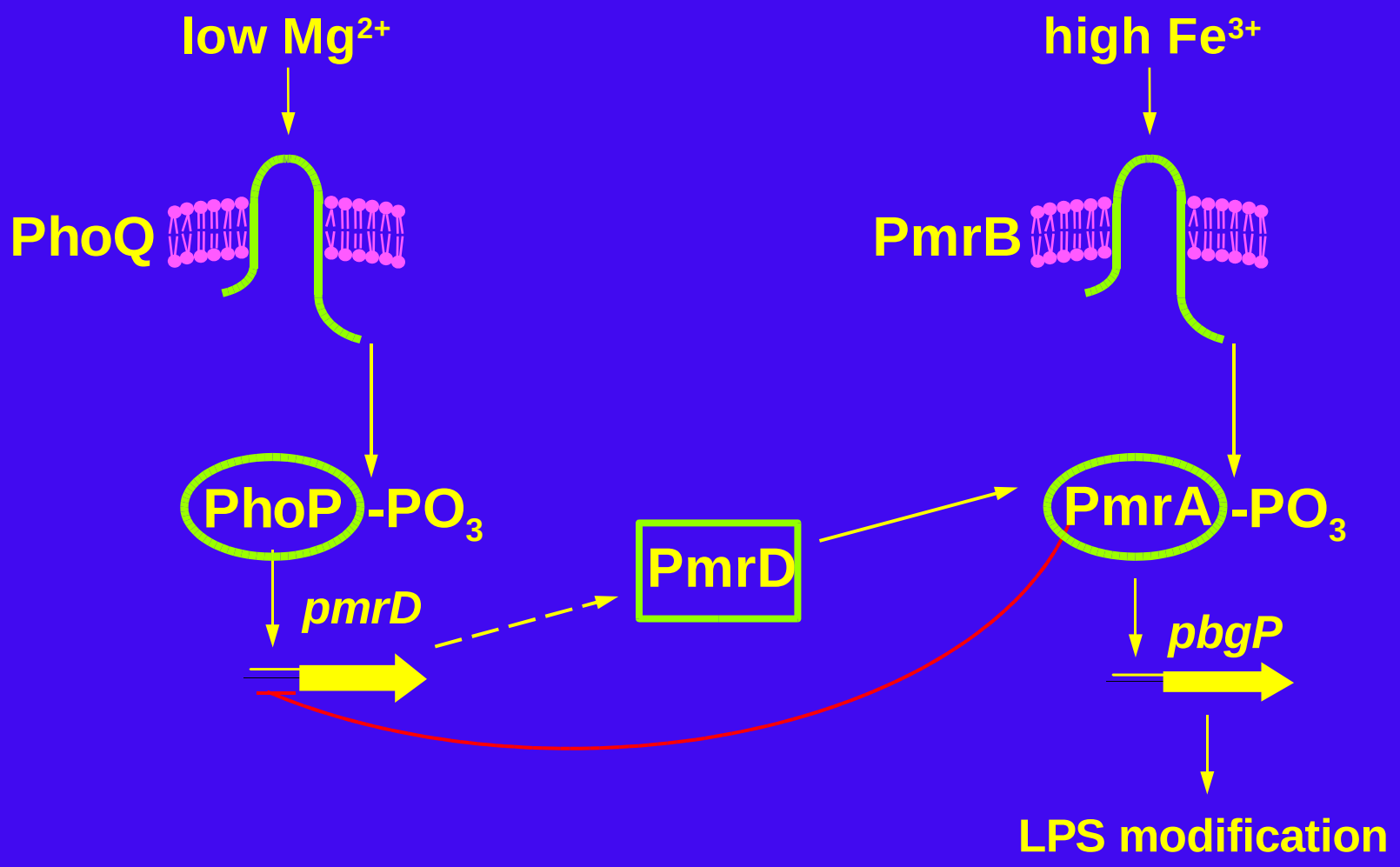




LOW Mg^{2+} INDUCES PMRA-ACTIVATED GENES VIA THE PHOP/PHOQ-ACTIVATED PMRD PROTEIN



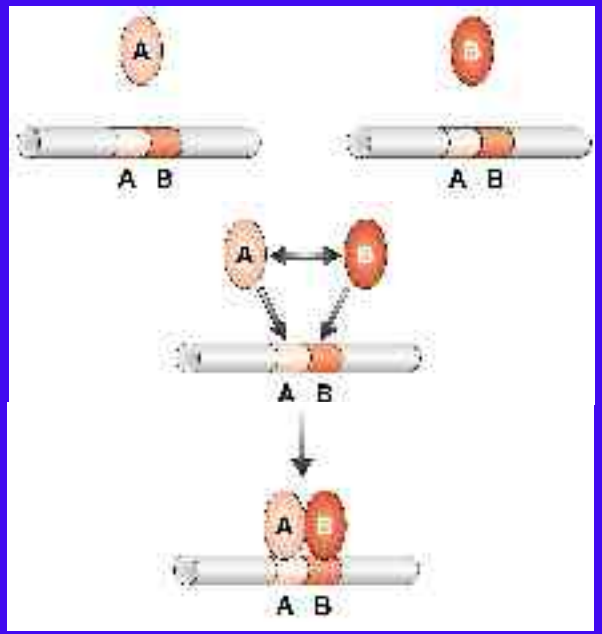
Fe^{3+} REPRESSES *pmrD* EXPRESSION VIA THE PMRA/PMRB SYSTEM



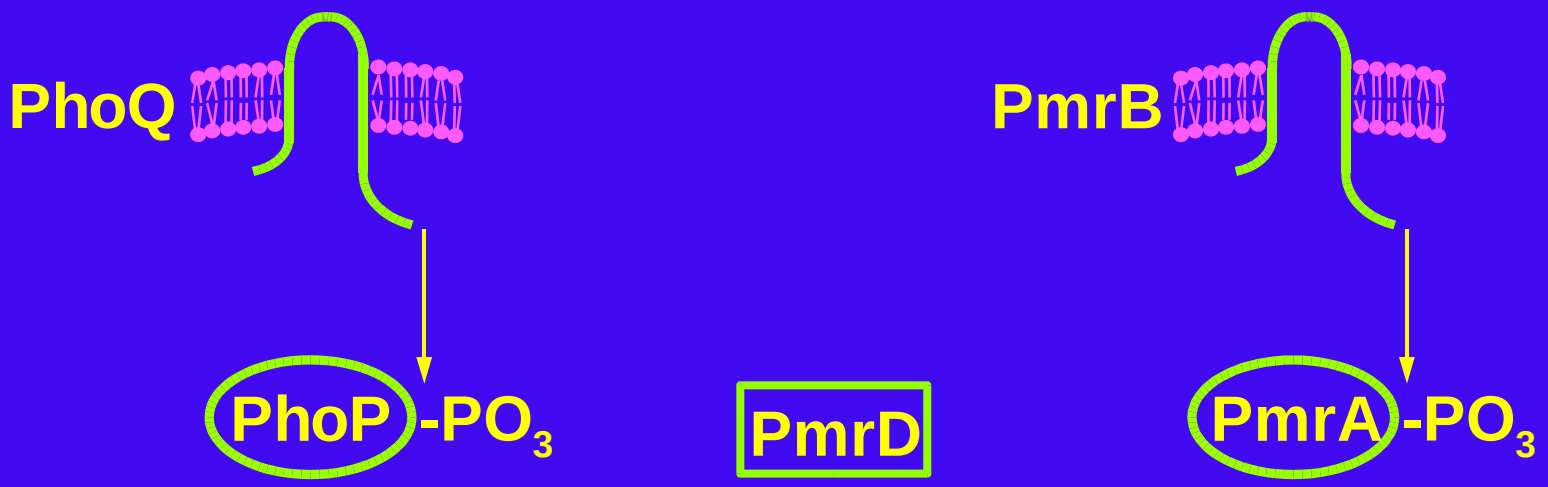
TRANSCRIPTIONAL REGULATION:

RECRUITING AND COOPERATING

AT***TGTTTA***TTTAC***TGTTTA***GCGCGCGCGC***TTGACA***ATTTT***TATAAT***



THE PMRD PROTEINS OF *SALMONELLA* AND *E. COLI* EXHIBIT UNUSUALLY LOW AMINO ACID IDENTITY



85.4% 93.3% 55.3% 84.0% 90.5%

(the median amino acid identity between *Salmonella* and *E. coli* proteins is 90%)

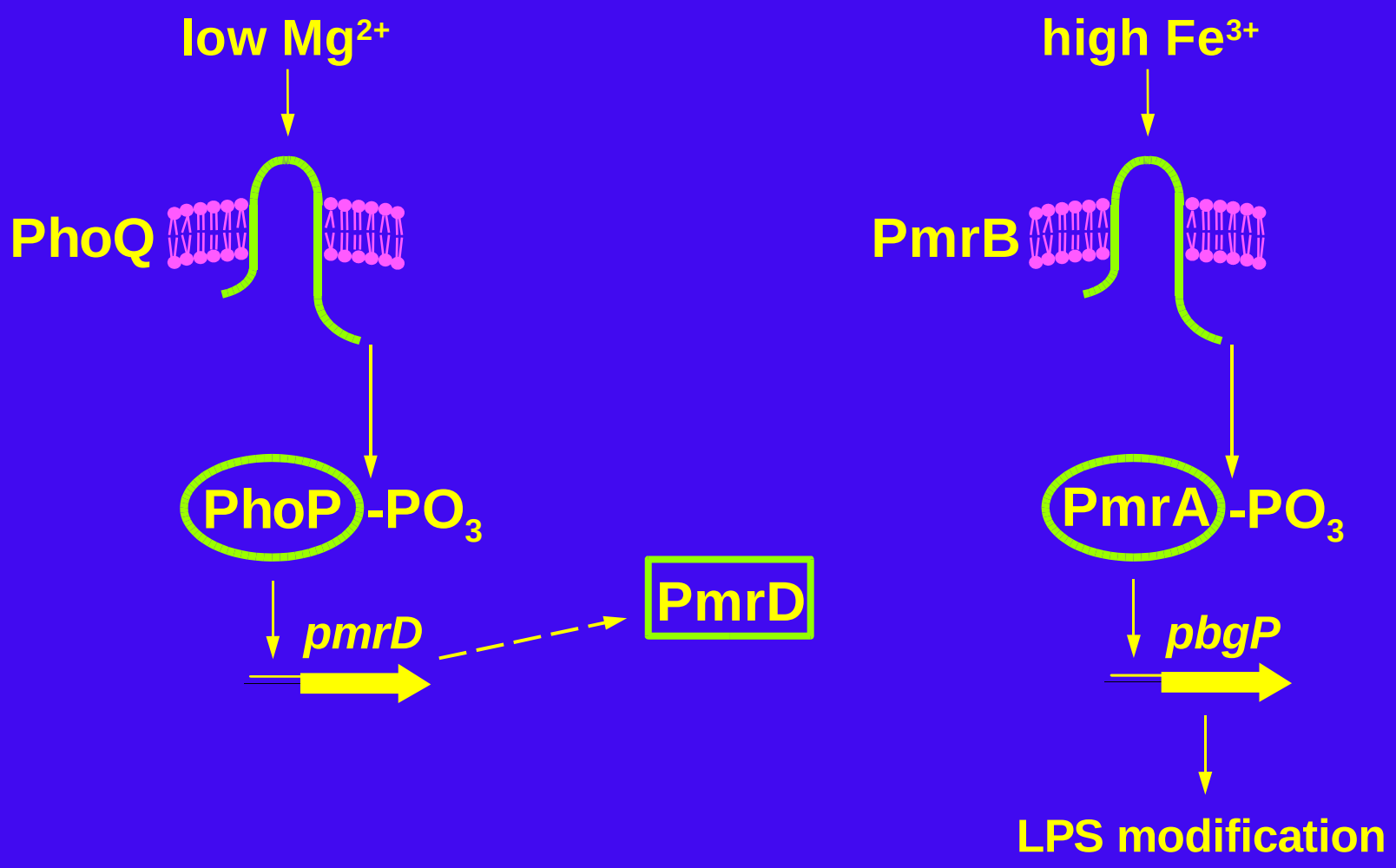
THE *ugd* PROMOTER OF *SALMONELLA*, BUT NOT OF *E. COLI*, HARBORS A PHOP BOX

Salmonella

E. coli

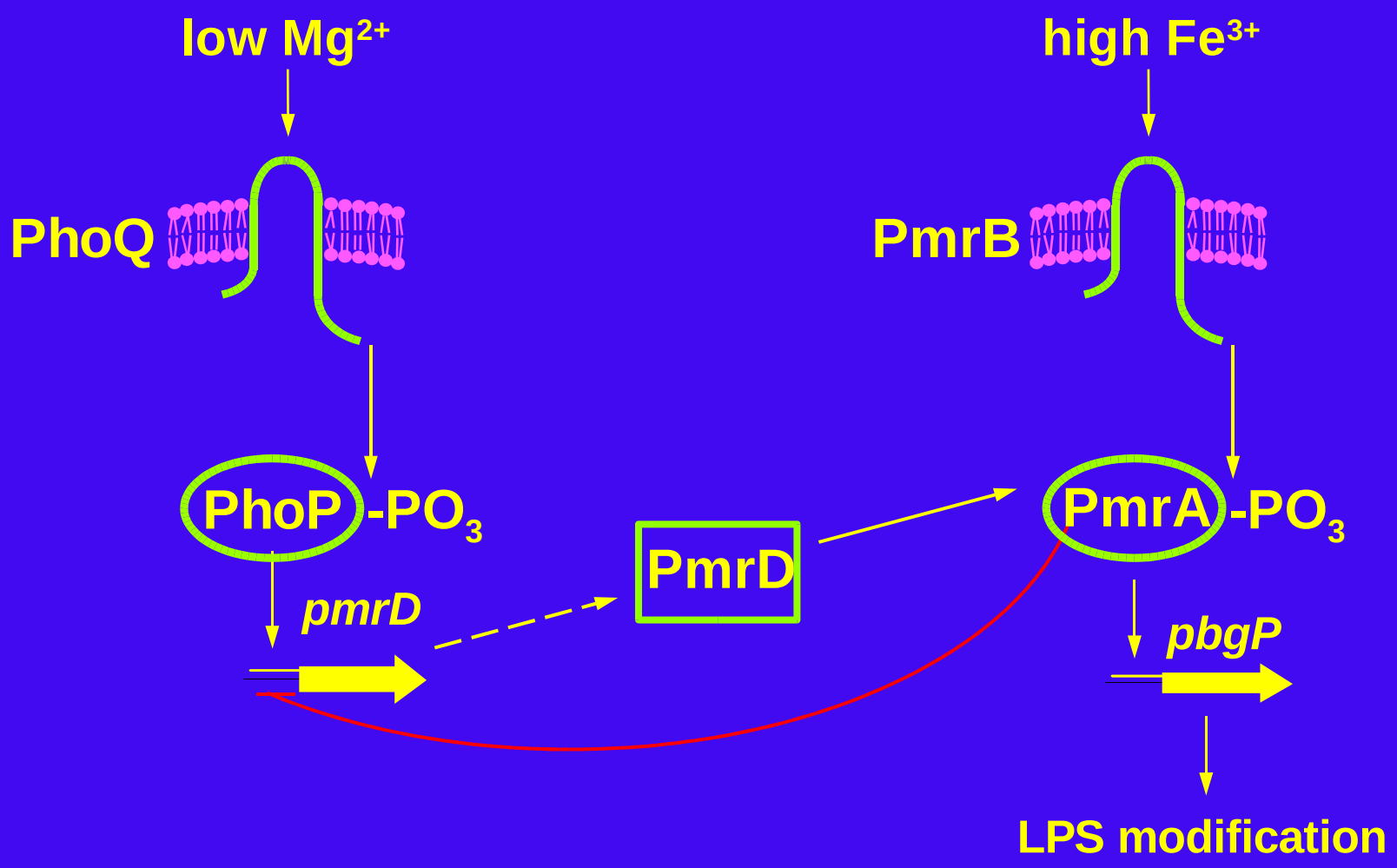
THE *E. COLI* PMRA/PMRB SYSTEM

RESPONDS TO Fe^{3+} BUT NOT TO LOW Mg^{2+}



THE *SALMONELLA* PMRA/PMRB SYSTEM

RESPONDS TO Fe^{3+} AND LOW Mg^{2+}

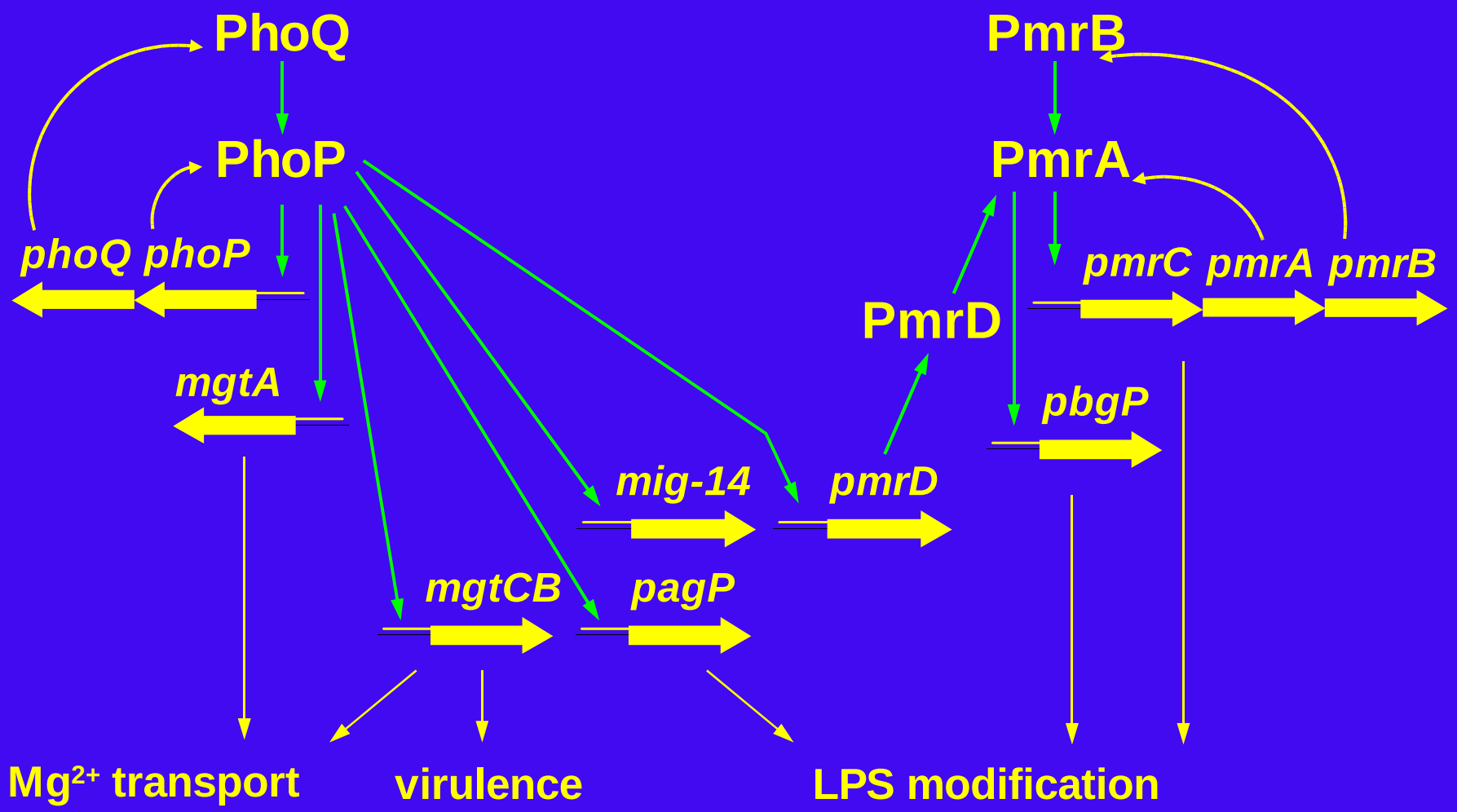


THE *pmrD* PROMOTER OF *SALMONELLA*, BUT NOT OF *E. COLI*, HARBORS A PMRA BOX

Salmonella

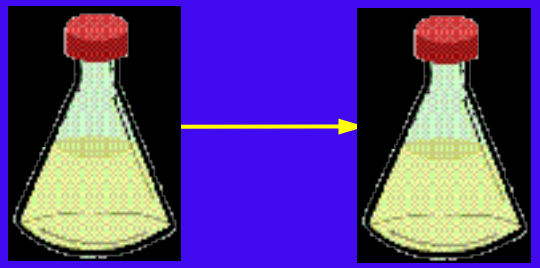
E. coli

SEQUENTIAL ACTIVATION OF THE PHOP REGULON

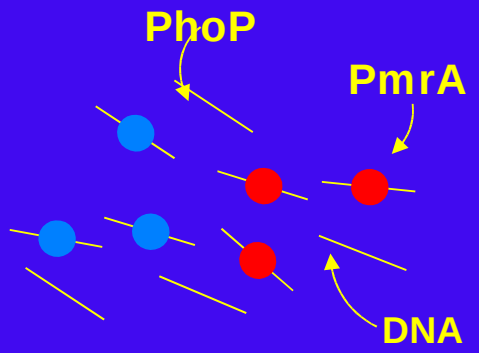


CHROMATIN IMMUNOPRECIPITATION (ChIP) STUDY OF PHOP- AND PMRA-REGULATED PROMOTERS

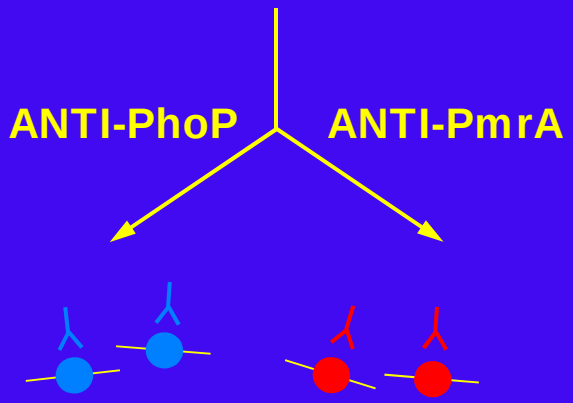
high Mg^{2+} → low Mg^{2+}



AT DIFFERENT TIMES, ADD CROSSLINKER, LYSE CELLS, AND FRAGMENT DNA



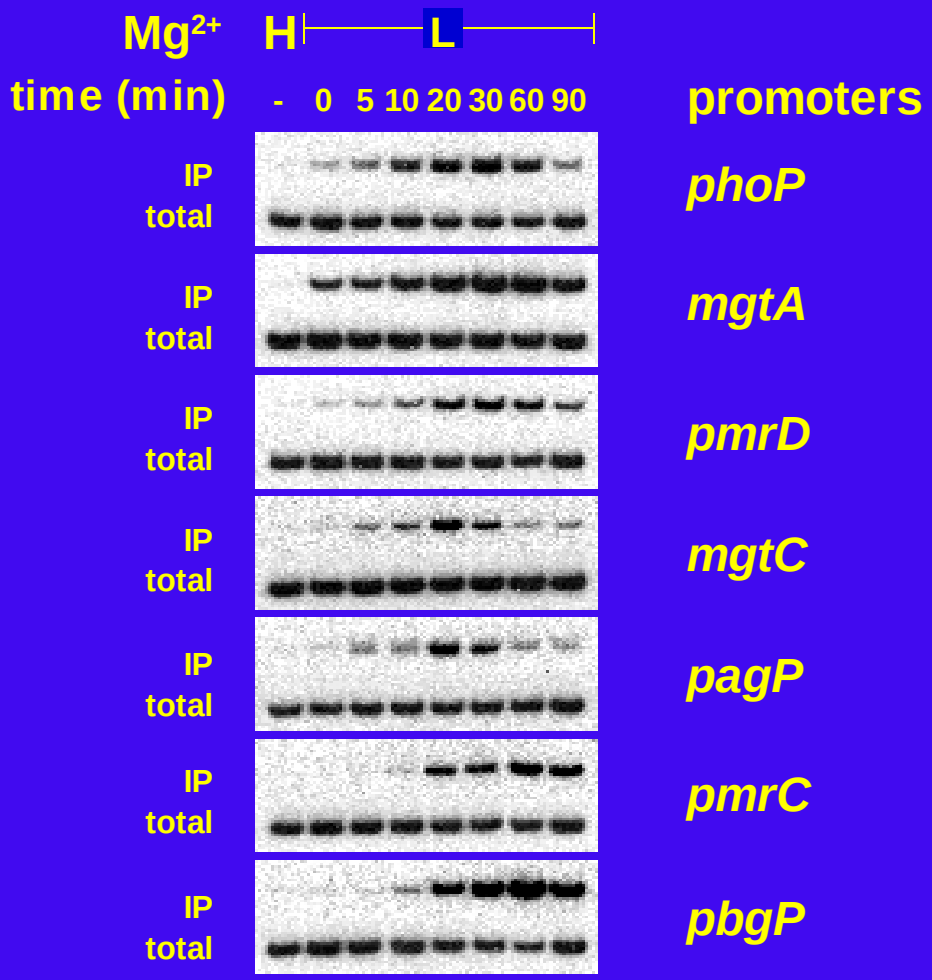
IMMUNOPRECIPITATION



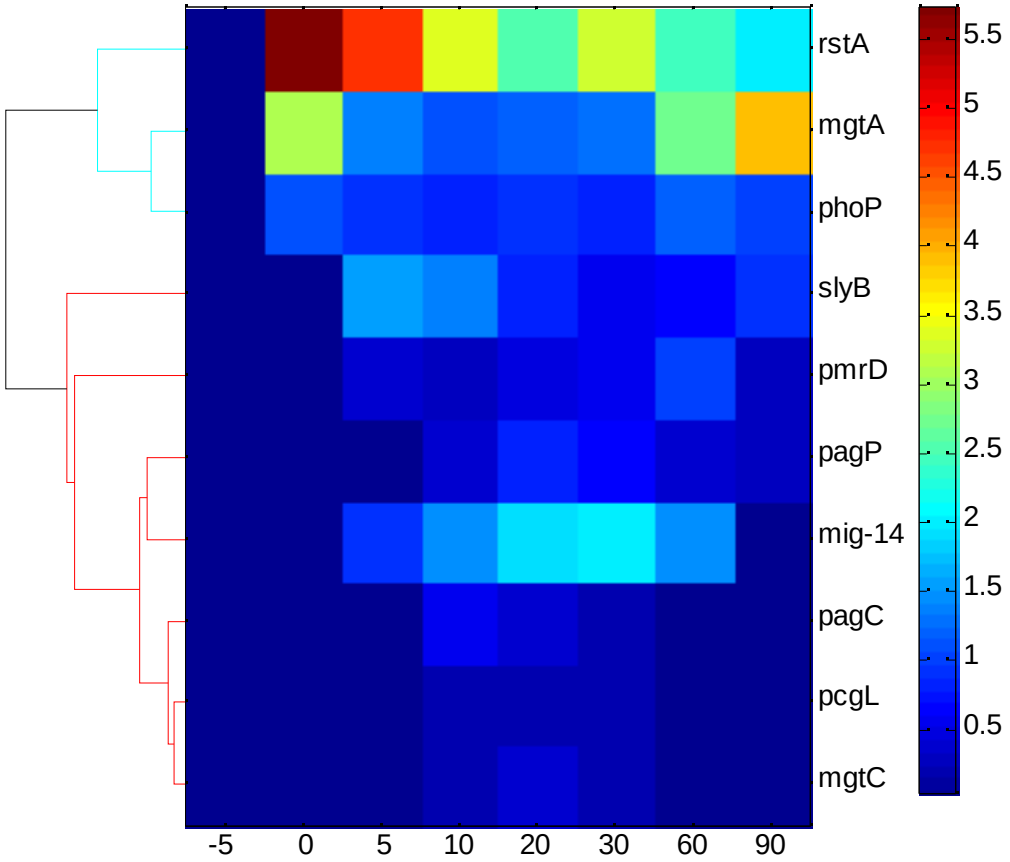
CROSS-LINKING REVERSAL AND DNA PURIFICATION

PCR SPECIFIC FOR PhoP- OR PmrA-REGULATED PROMOTERS

ORDERED BINDING OF THE PHOP AND PMRA PROTEINS TO THEIR TARGET PROMOTERS



TEMPORAL ORDER OF BINDING



ANATOMY OF BACTERIAL PROMOTERS

Class I promoter

Class II promoter

THE PHOP PROTEIN CONTROLS TRANSCRIPTION FROM DIFFERENT CLASSES OF PROMOTERS

FACTORS DETERMINING TRANSCRIPTION FROM A BACTERIAL PROMOTER

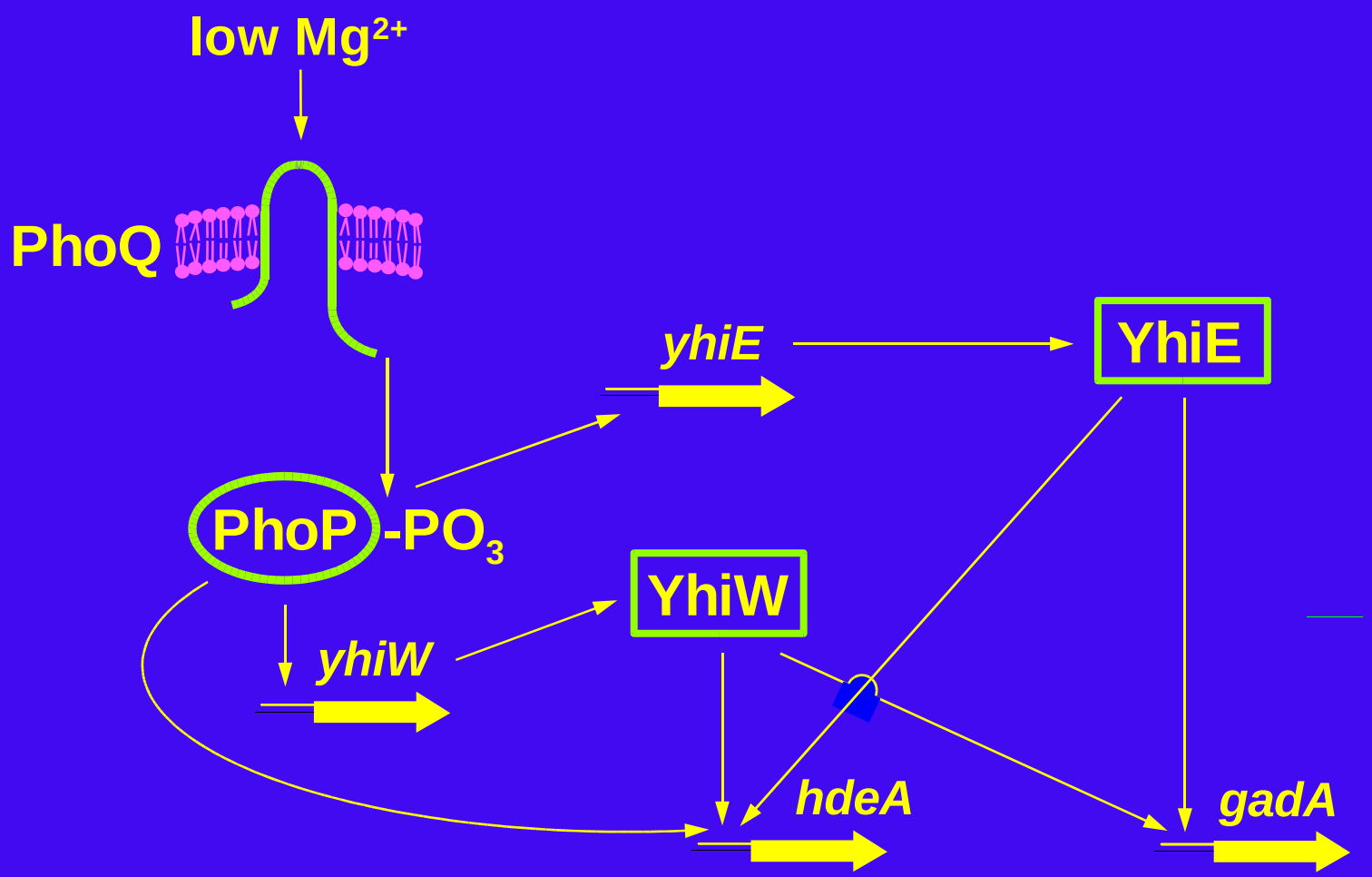
. Quality of the binding site for the transcription factor

. Distance and orientation of the binding site relative to the RNA polymerase binding site

. Quality of the binding site for RNA polymerase

. Presence of binding sites for other regulatory proteins

THE PHOP/PHOQ SYSTEM REGULATES EXPRESSION OF ACID pH RESISTANCE GENES IN *E. COLI*

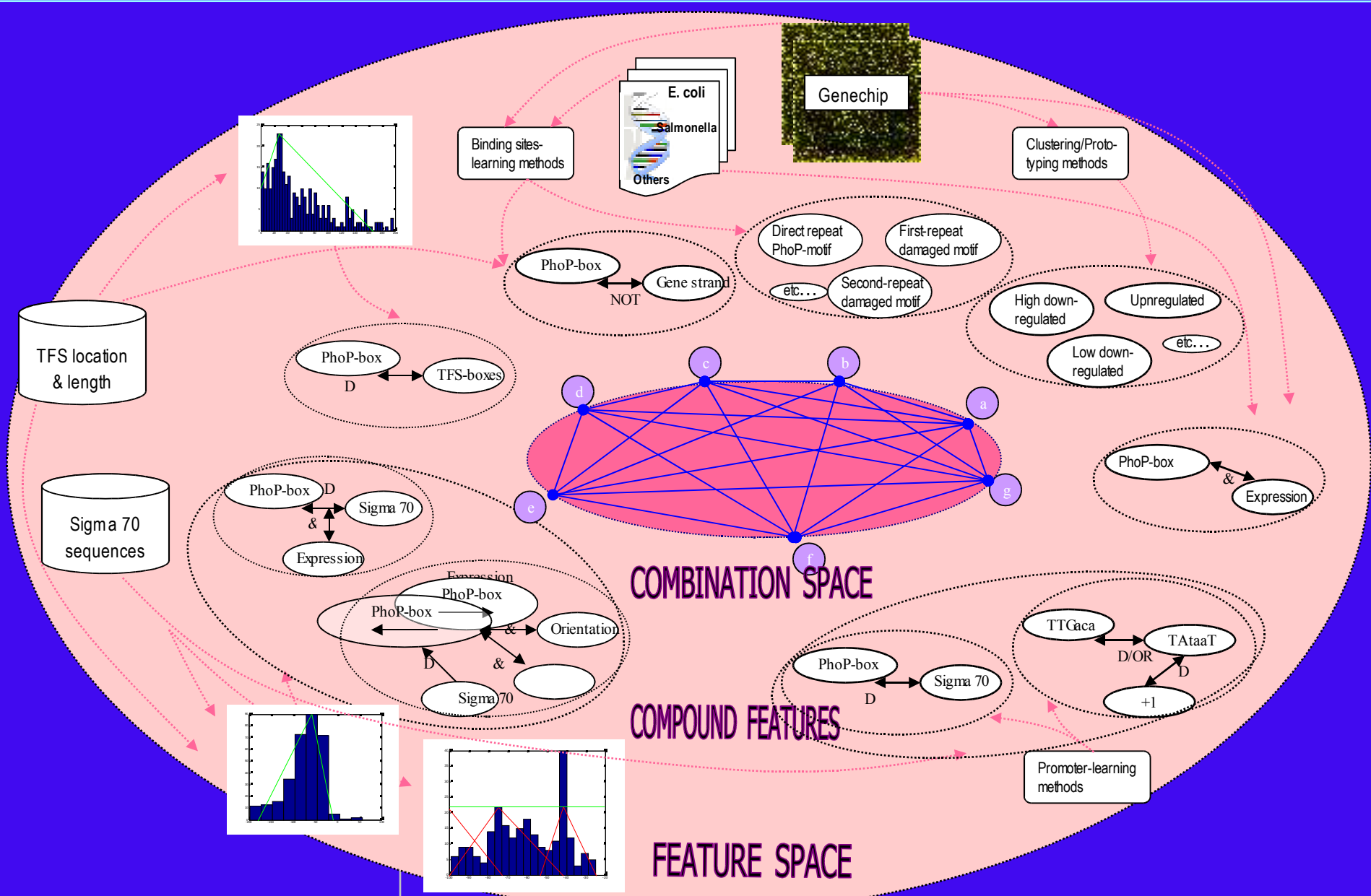


Gene Promoter Scan (GPS)

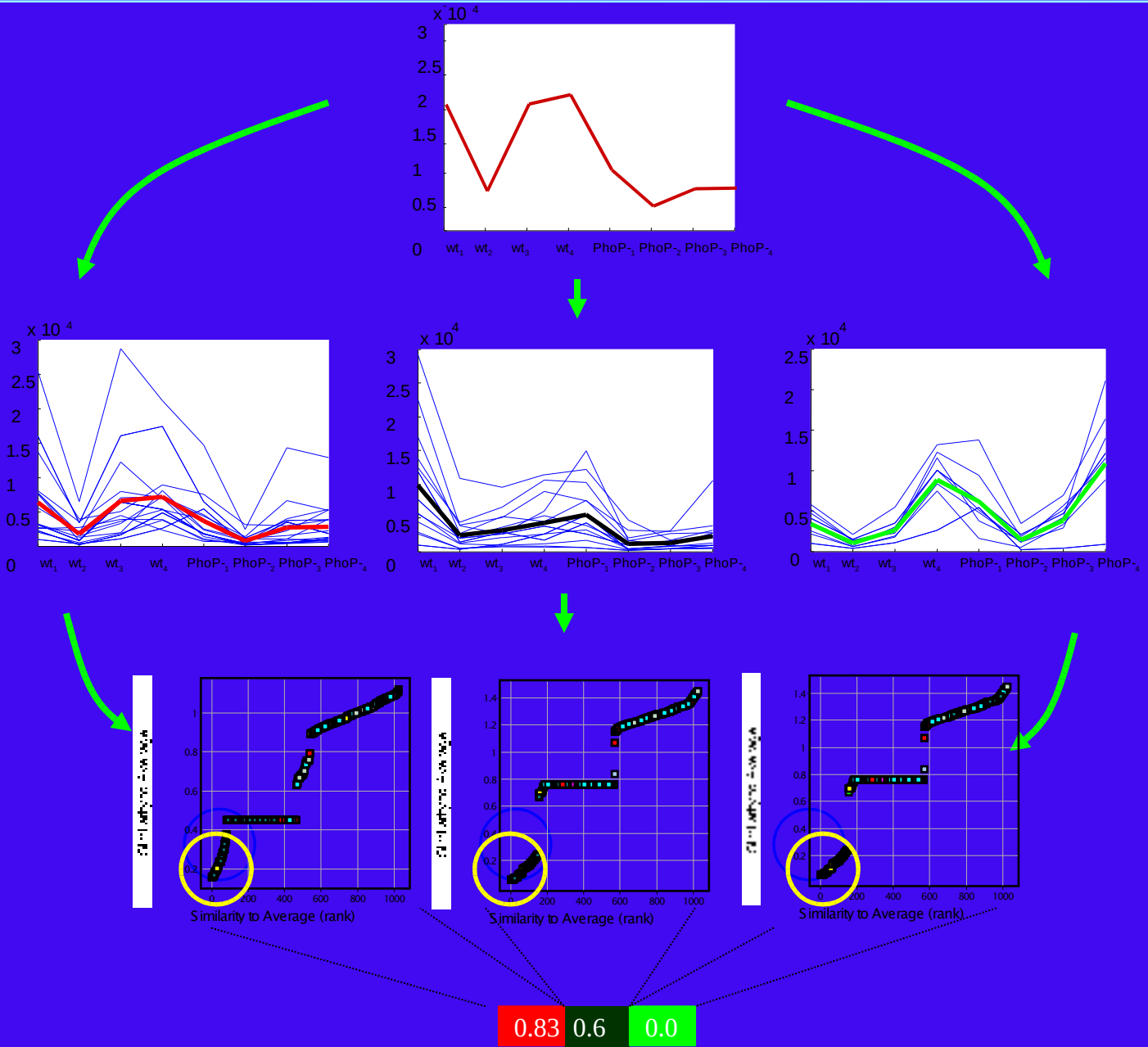
OUR ULTIMATE GOAL IS TO DEFINE THE MOLECULAR LOGIC OF PHOP REGULATION BY …

- Identifying regulatory profiles that summarize the behavior of the PhoP regulon
- Understanding the subjacent biological properties of the regulatory profiles
- Uncovering interactions with other regulatory systems and proteins

REGULATORY FEATURES CONSIDERED INCLUDE …



GENE EXPRESSION FEATURES



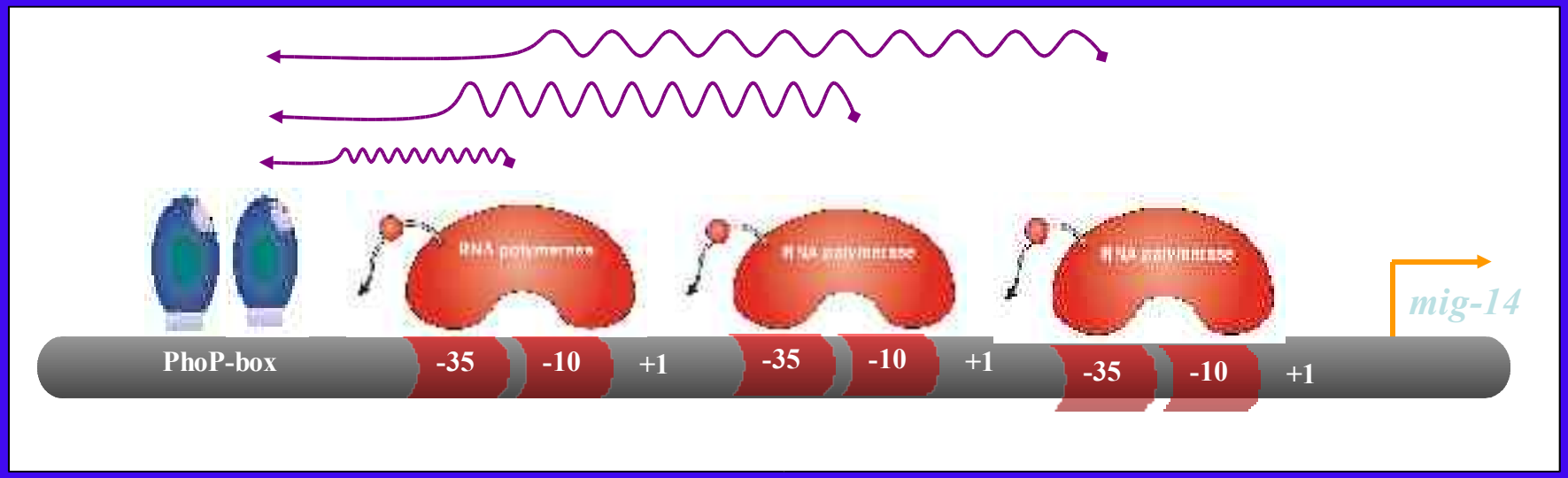
BINDING SITES SUBMOTIFS

```
#5 TTGTTTATGATTTGTTTA
A |   0   6   3   1   0  21   25   25  25   25   25   5   4   1   1   6  21
C |   3   0   0   3   1   3   25   25  25   25   25   1   0   2   0   0   2
G |   6  19   2   0   4   0   25   25  25   25   25   9  19   0   4   4   2
T |  16   0  20  21  20   1   25   25  25   25   25  10   2  22  20  15   0
```

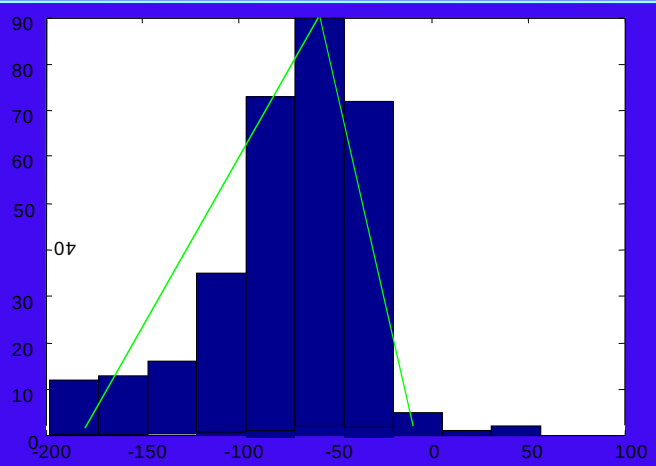
TGTTTA_ _ _ _ _ TGTTTA

M_1	T	TATT$^{G}_{T}$A	C	GTT	C	TGTTTA	T
M_2	A	$^{G}_{T}{}^{C}$GTTTA	T	G$^{A}_{T}$T	T	TGTTTA	A
M_3	A	TGTTTA	$^{A}_{G}$	A$^{A}_{T}$A	$^{C}_{T}$	$^{A}_{T}{}^{G}$GTTTA	$^{A}_{T}$
M_4	T	TGTTTA	T	AAT	T	TGTTGA	T

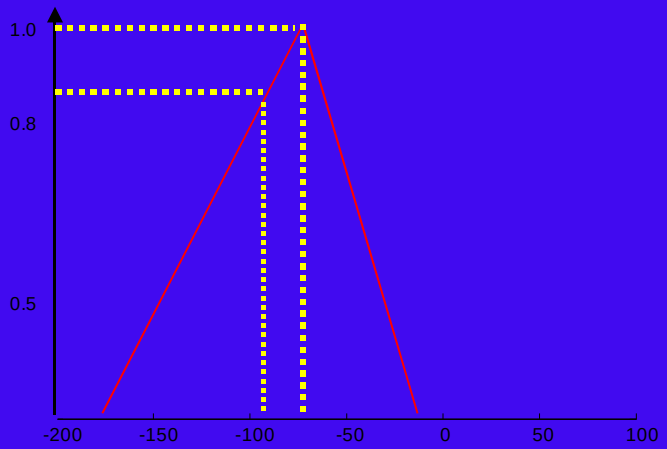
PROMOTER FEATURES: RNA POLYMERASE, CLASS AND LOCATION



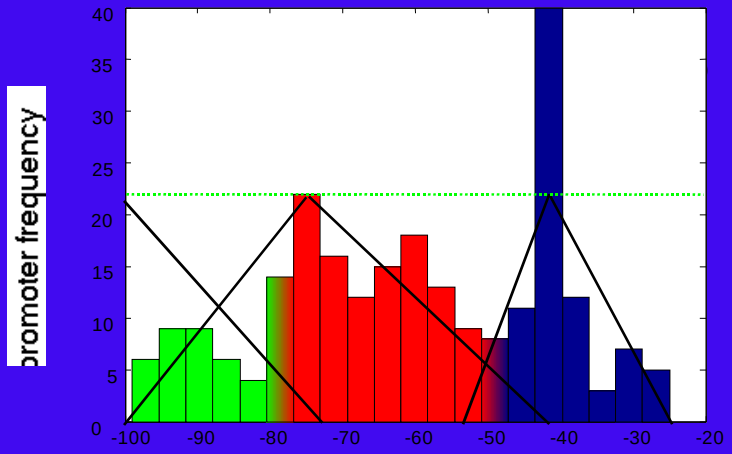
PROMOTER CONDITION: ACTIVATED/REPRESSED AND DISTANCE RELATIONSHIPS



Distance from binding sites to +1

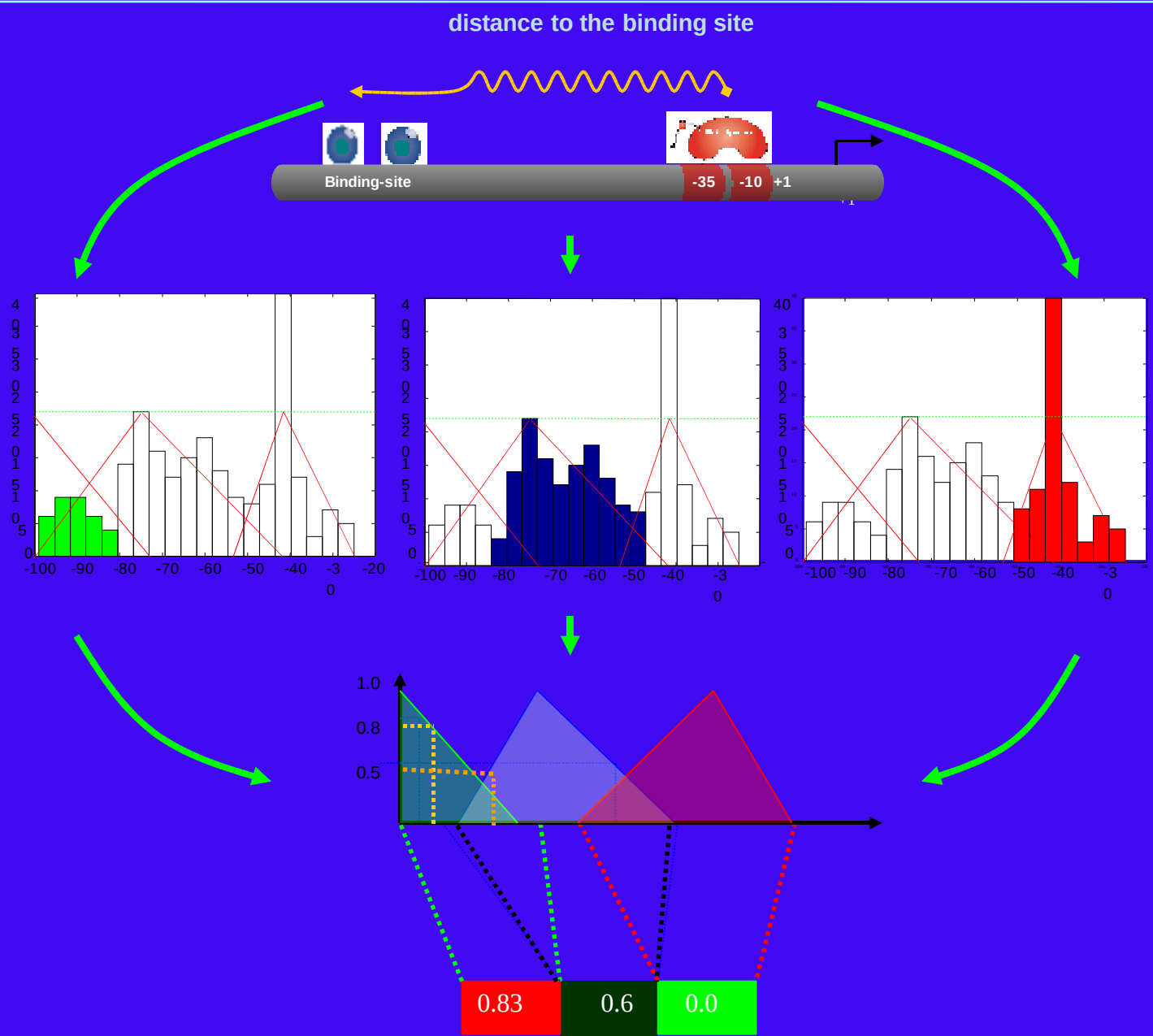


Distance from binding sites to +1

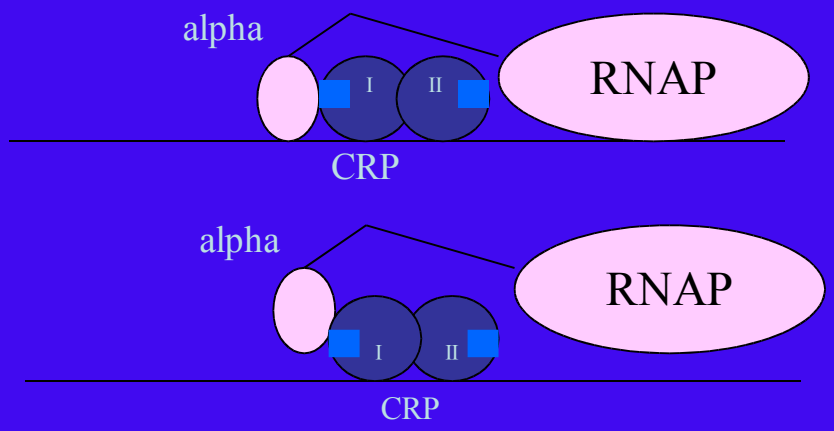


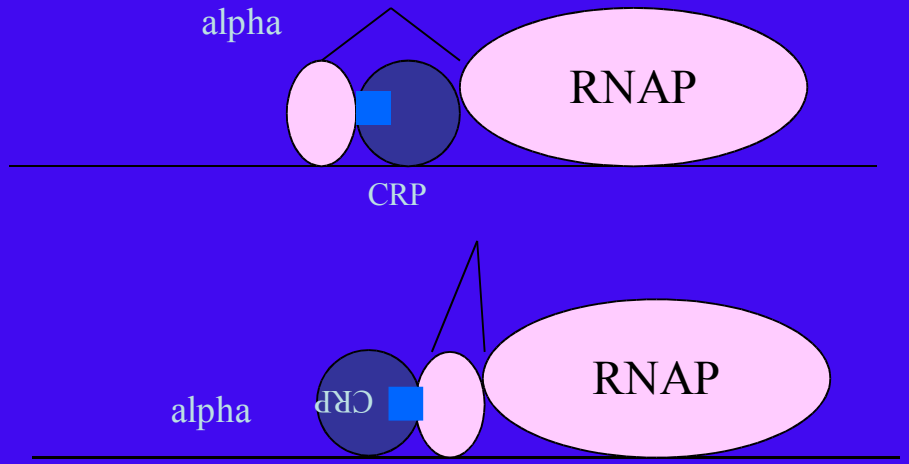
Distance from binding sites to +1

CLOSE, MEDIUM AND REMOTE PROMOTERS

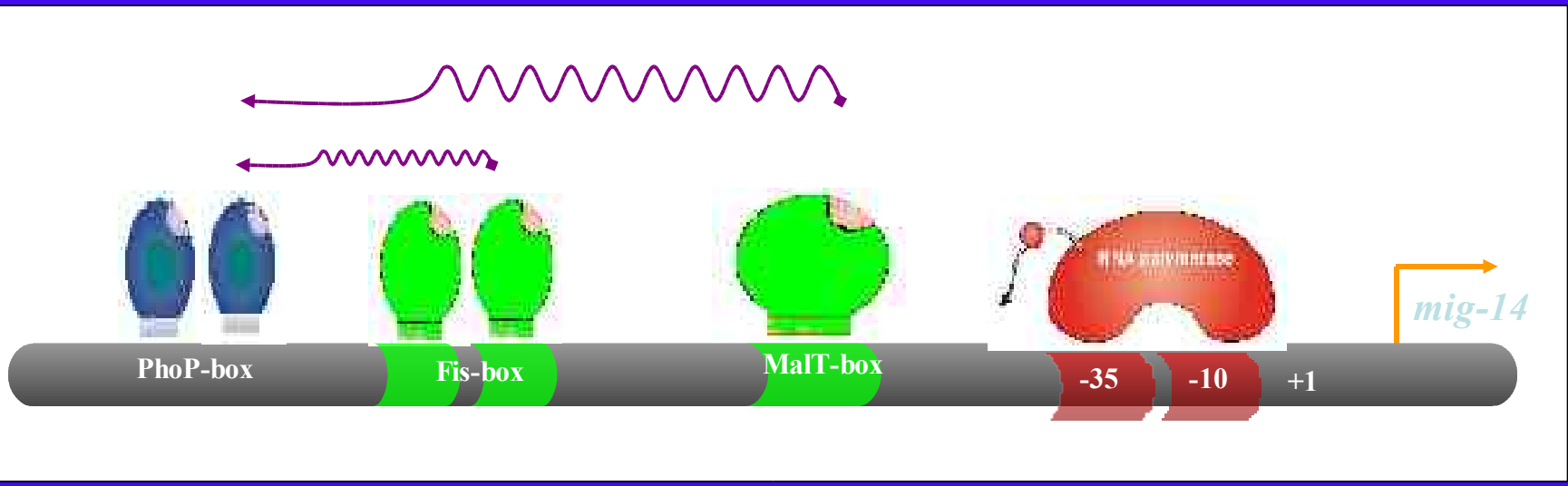


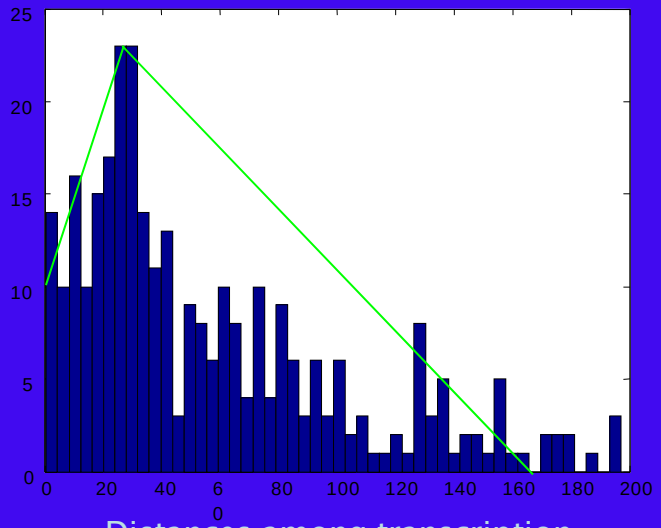
BINDING SITE ORIENTATION AND POSSIBLE TOPOLOGIES





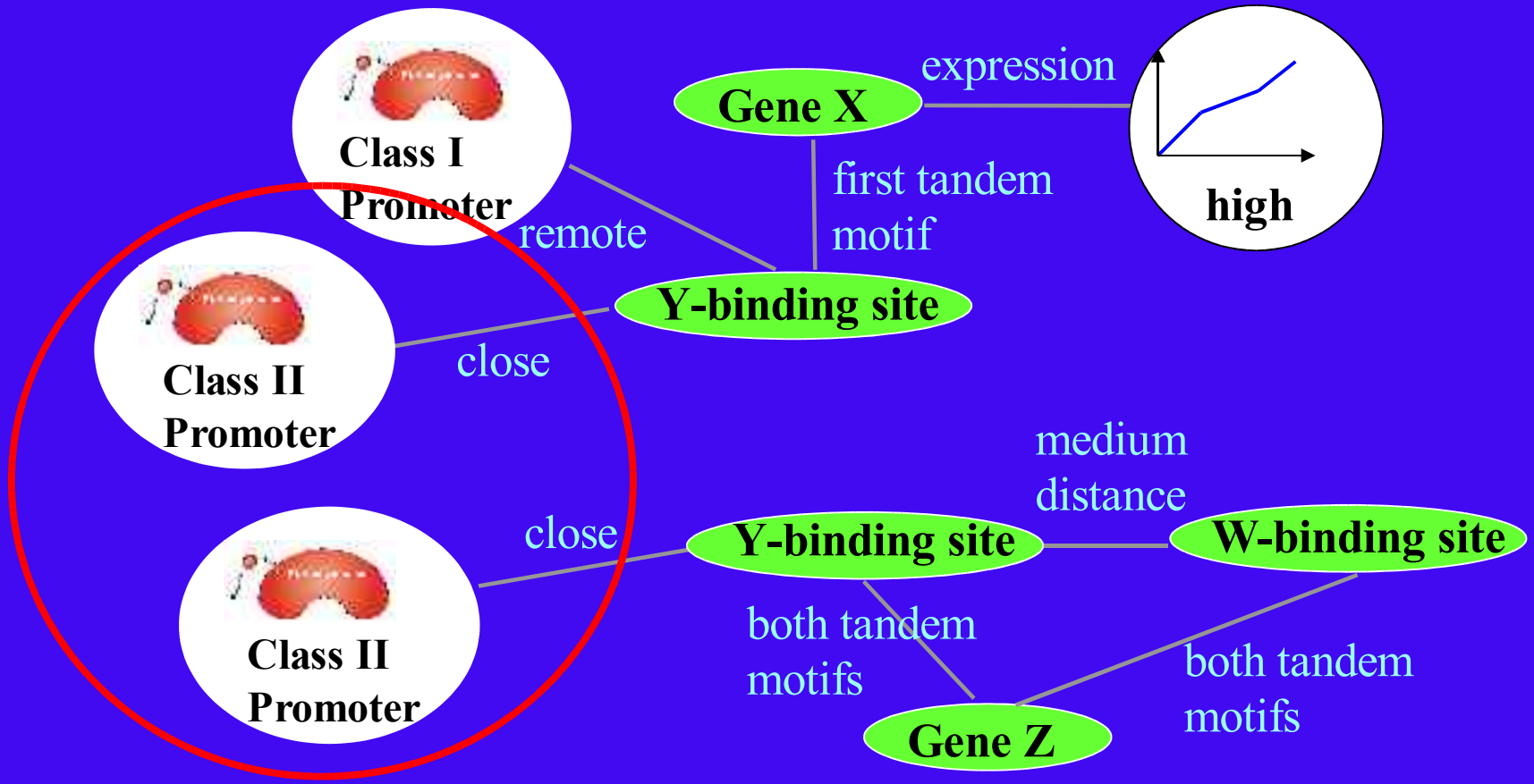
TRANSCRIPTION FACTOR INTERACTIONS





Distances among transcription factor binding sites

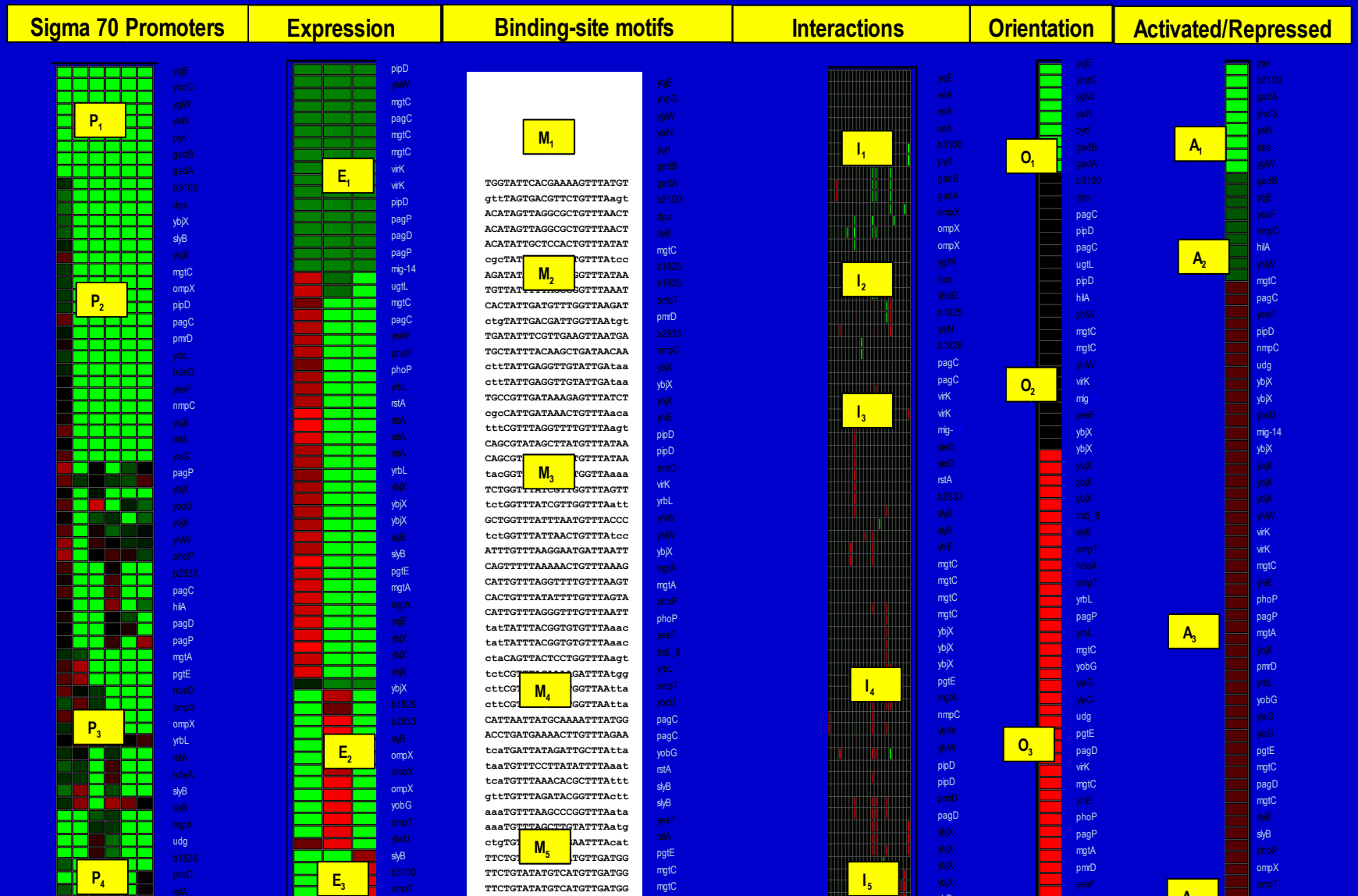
A CONCEPTUAL CLUSTERING APROACH FOR DISCOVERY OF REGULATORY PROFILES



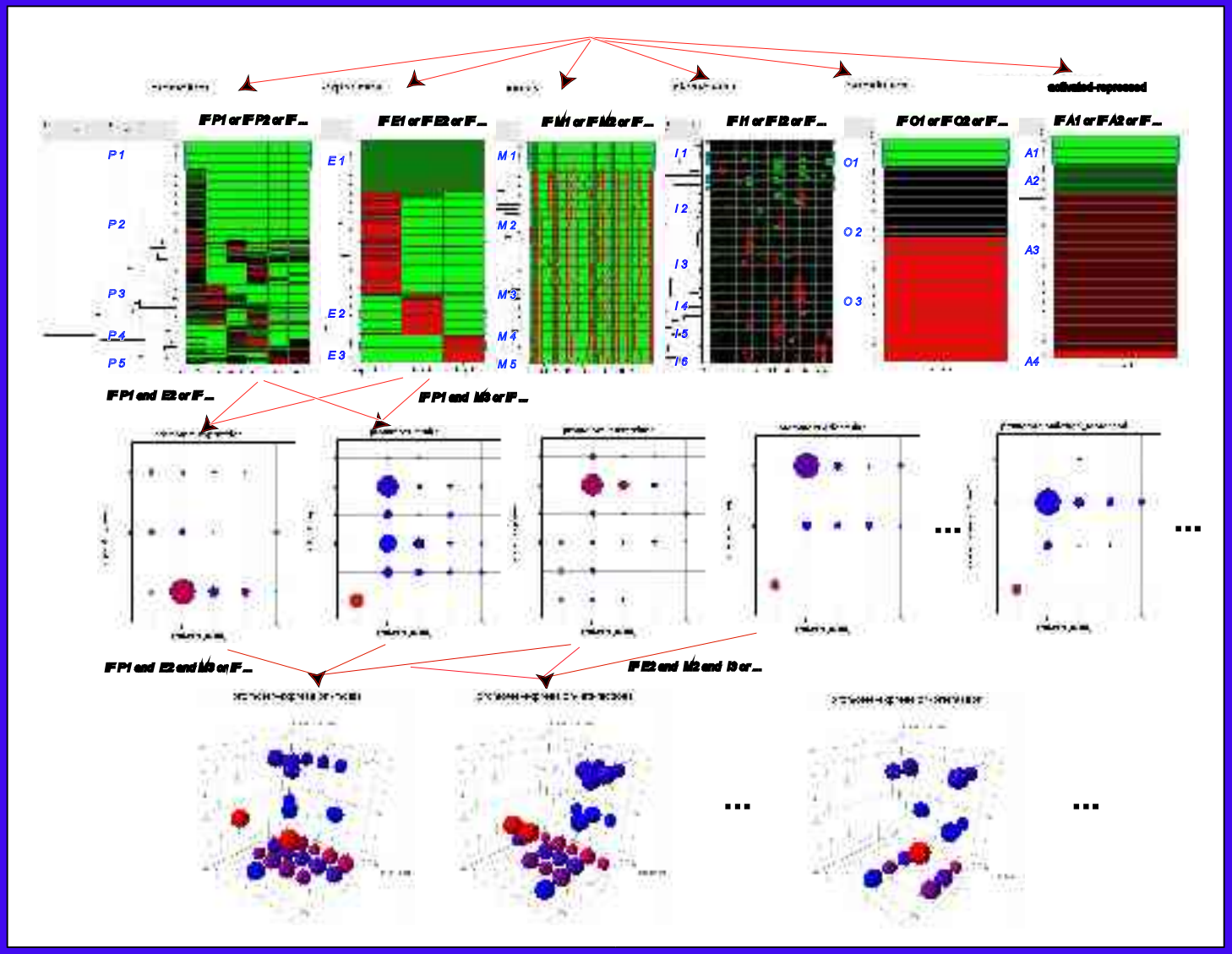
ADVANTAGES OF CONCEPTUAL CLUSTERING (Michalsky, Chesseman, Cook, Ruspini, etc.)

- Hypothesis generation – data exploration – model fitting – finding true topologies
- Deal with structural data
- Iterative and incremental search of interesting repetitive substructures
- Attributes and relations belong to more than one substructure
- Attributes and relations are flexible
- Ability to deal with missing values
- DISCOVER → DESCRIBE -> PREDICT

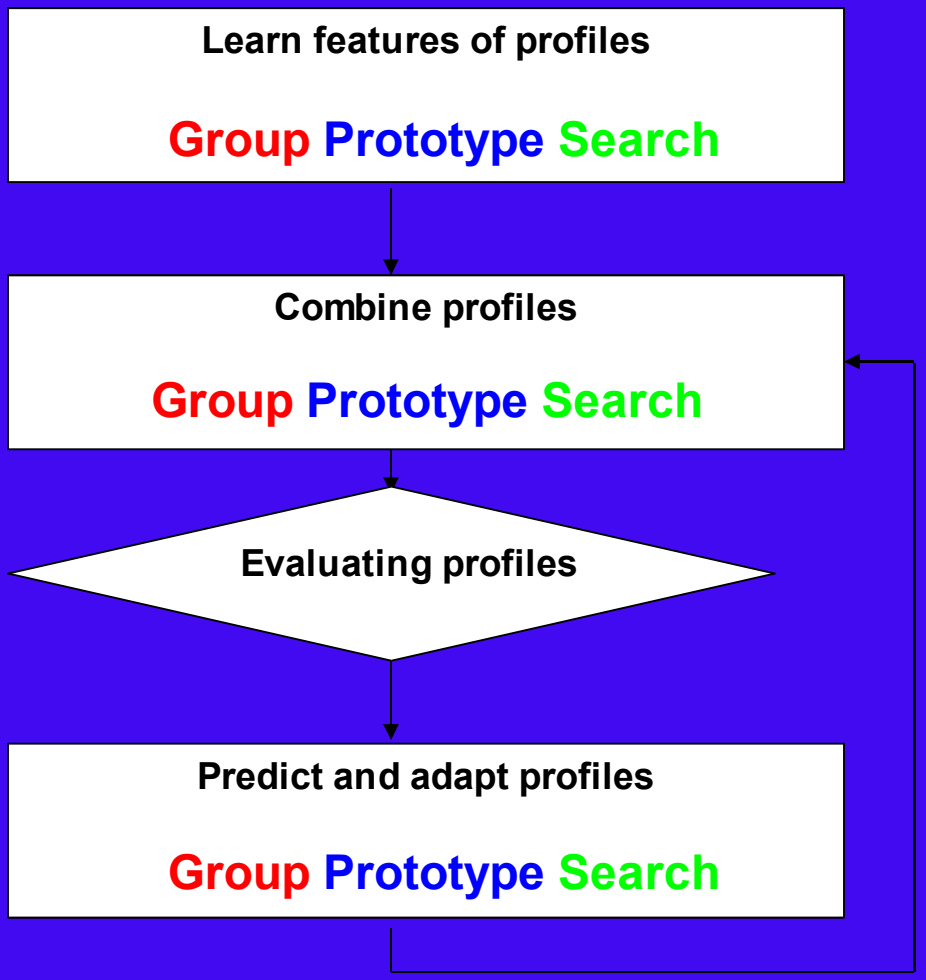
THE BUILDING BLOCK PROFILES OF PHOP-REGULATED GENES



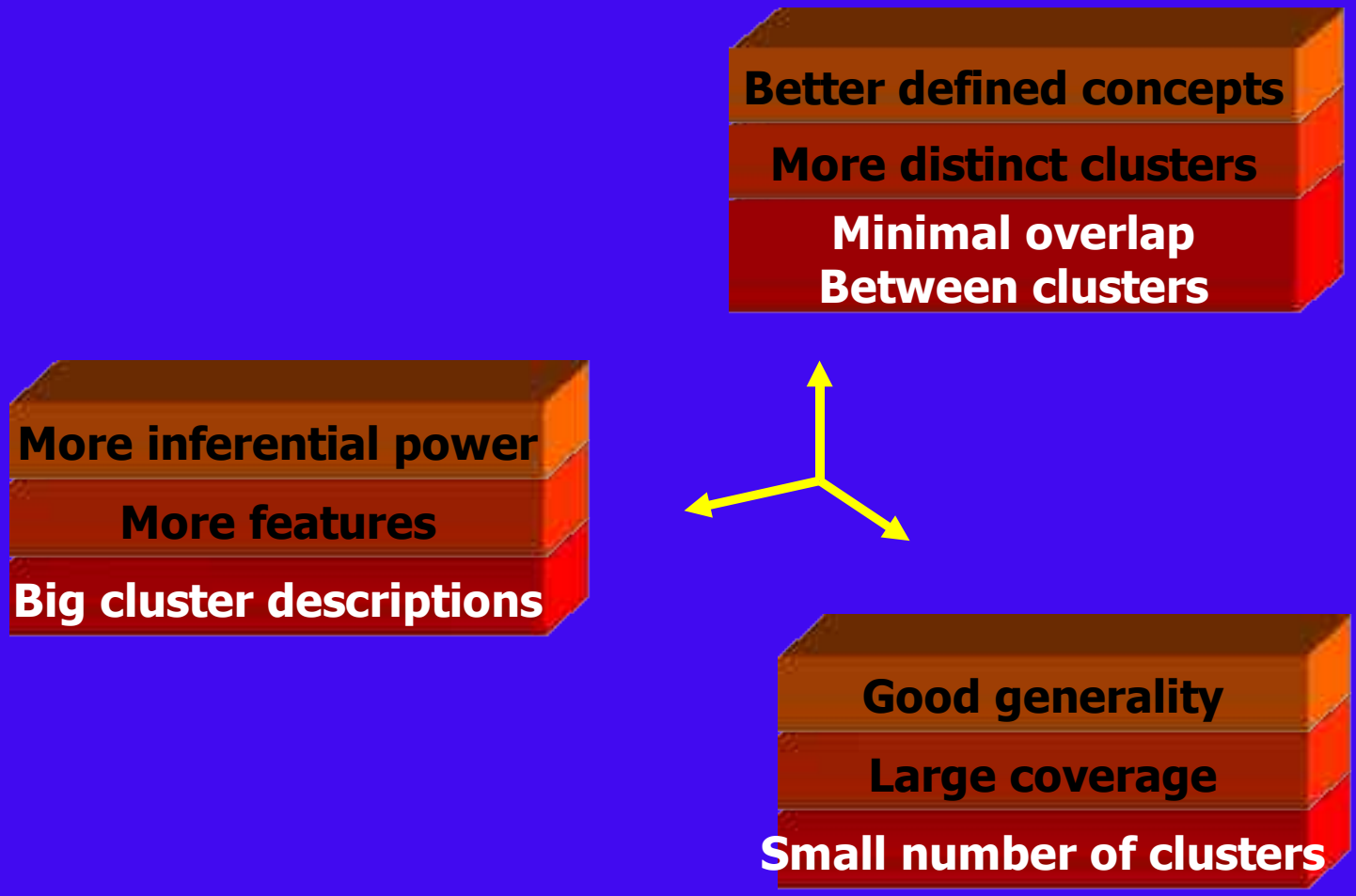
THE COMBINED PROFILES OF PHOP-REGULATED GENES



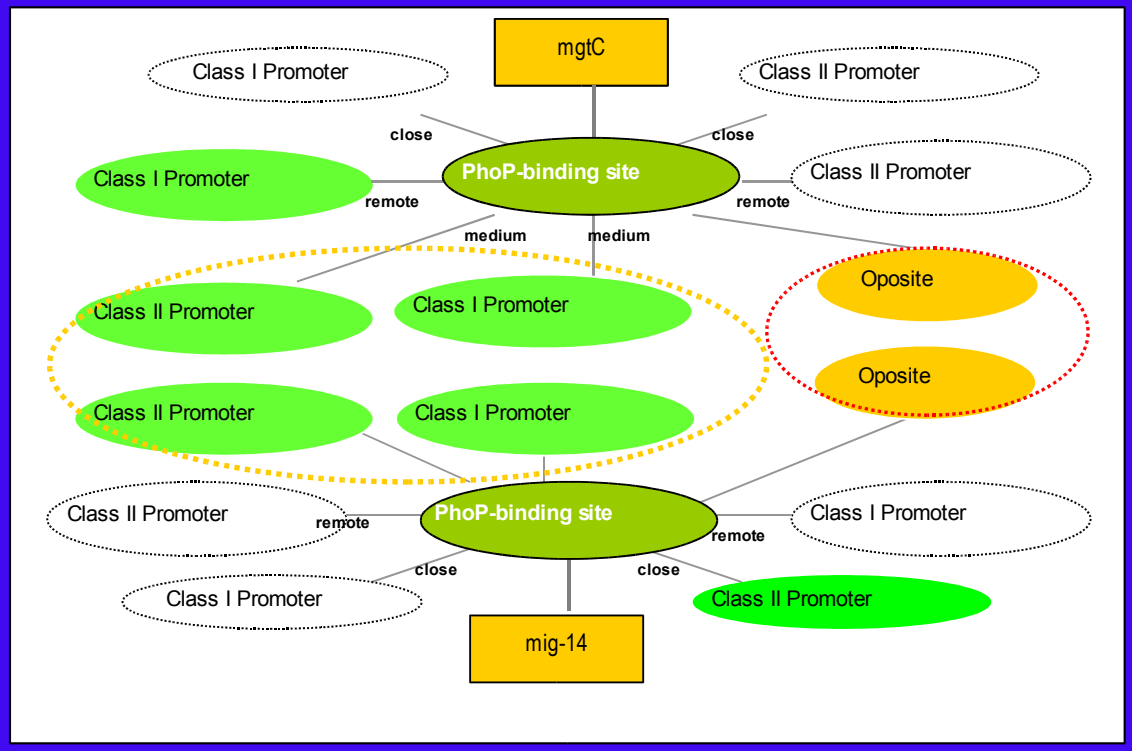
Gene Promoter Scan: ALGORITHM

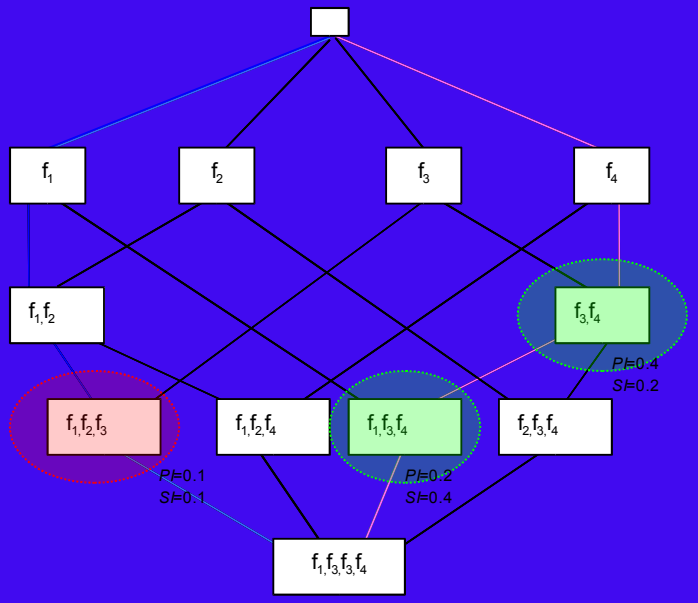


AN EVALUATION OF THE GENERATED COMBINED PROFILES

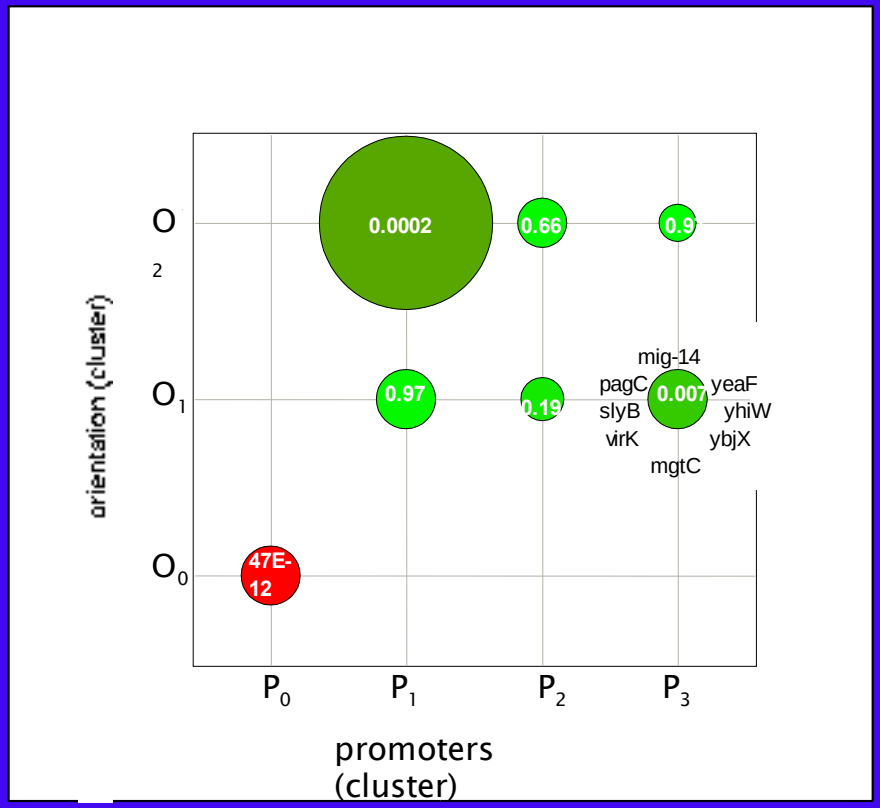


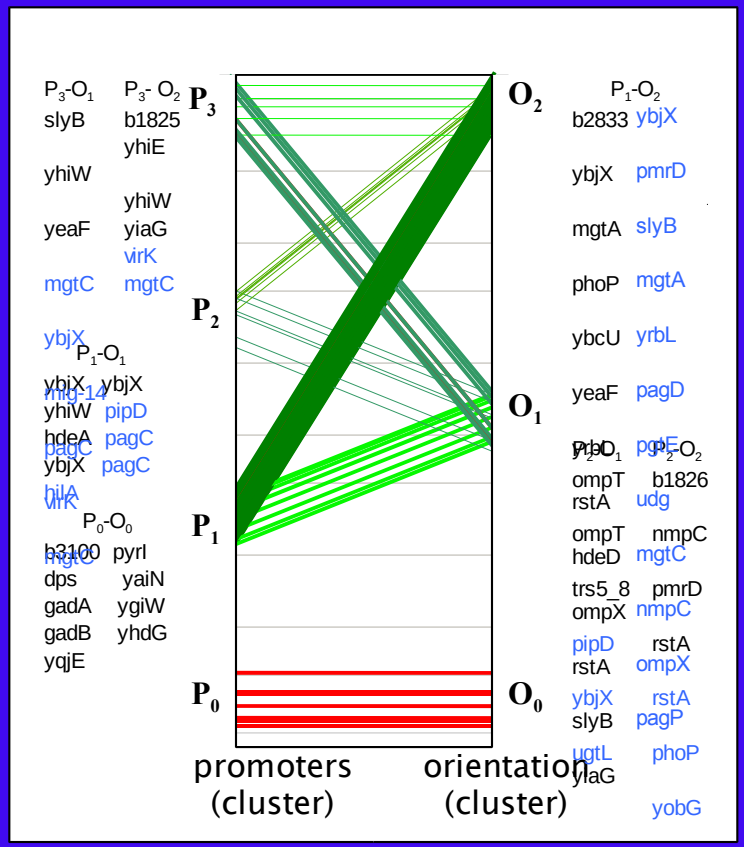
GPS: A CONCEPTUAL CLUSTERING METHOD

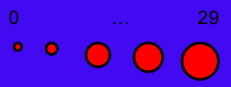




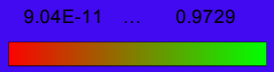
GPS EVALUATION: PROBABILITY OF INTERSECTION





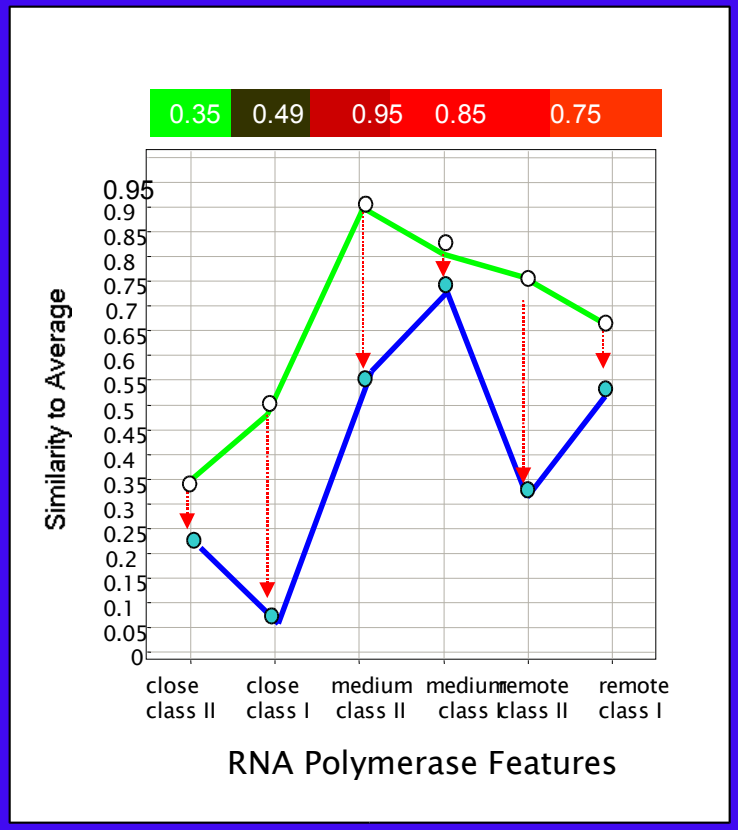


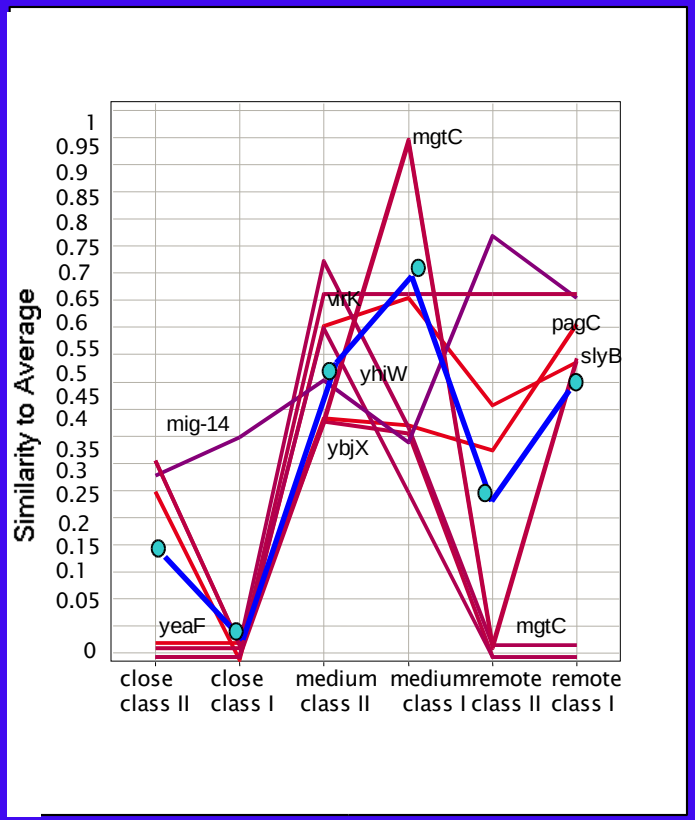
Size by records count
Line with by count



Color by probability of interaction

GPS EVALUATION: SIMILARITY OF INTERSECTION

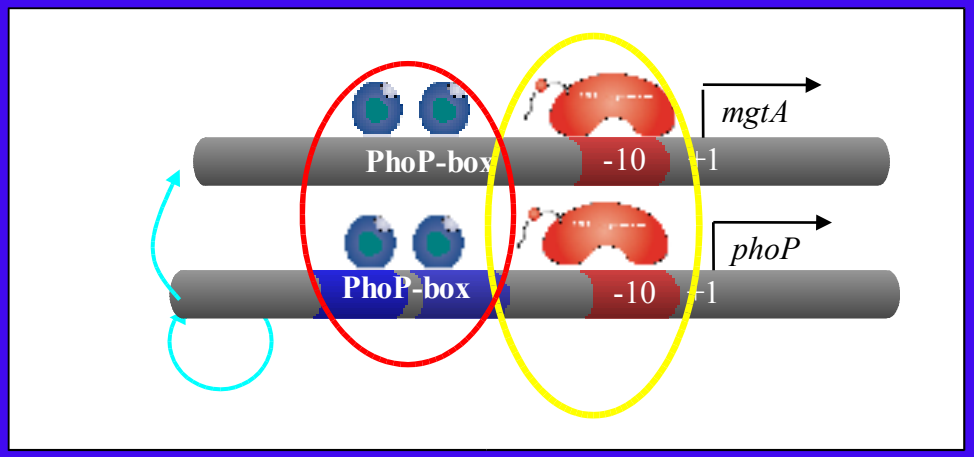


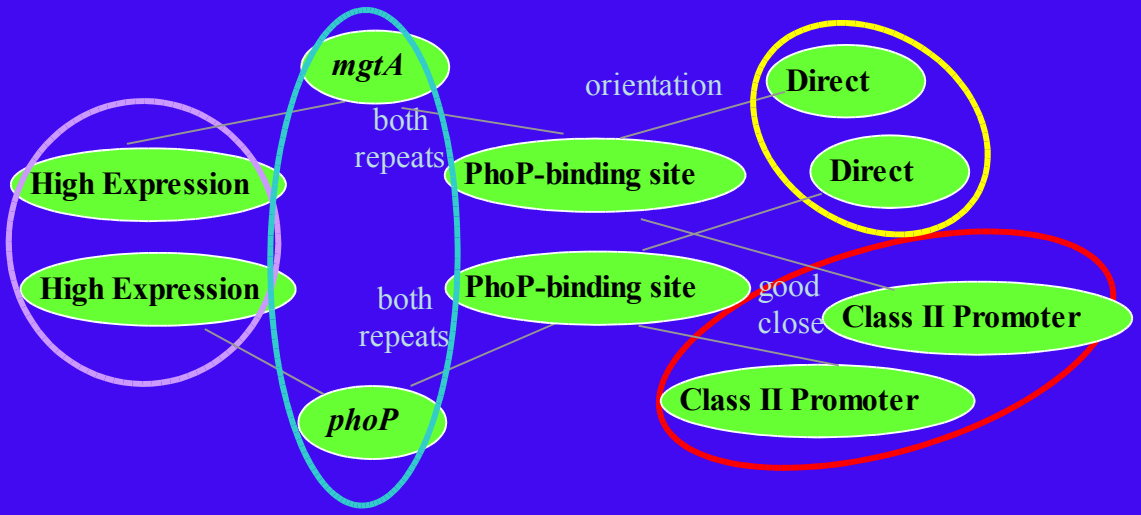


THE PUBLISHED PHOP-REGULATED PROMOTER (Yamamoto et al., 2002; Minagawa et al., 2003)

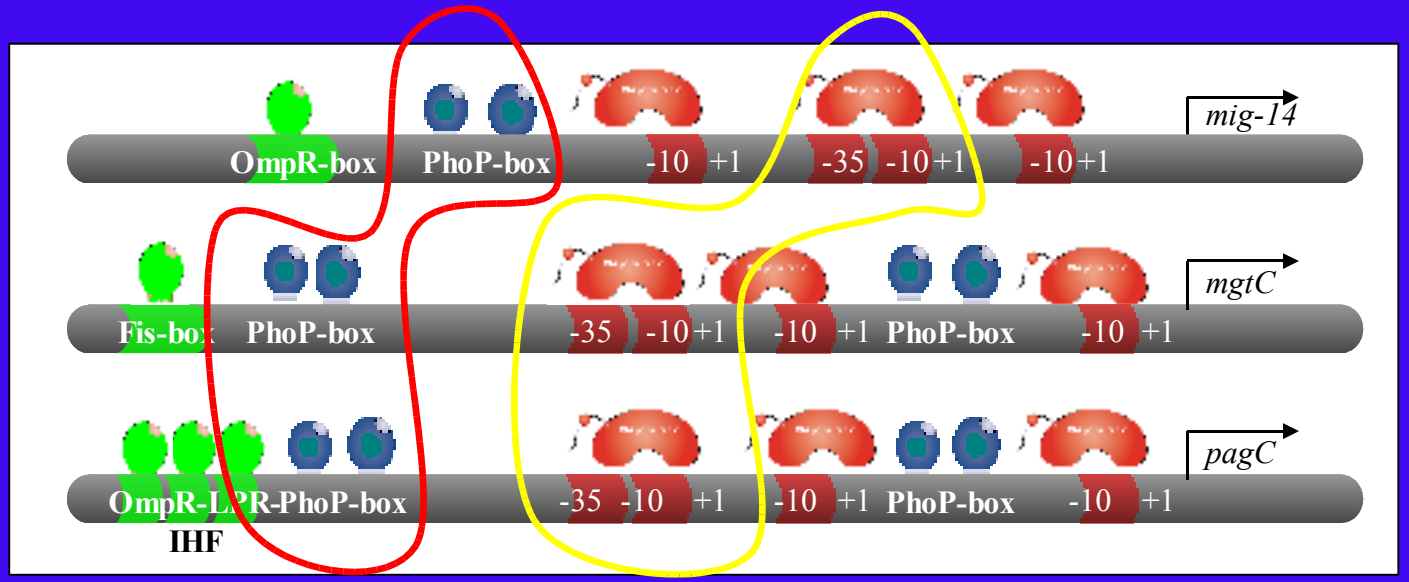
- Class II
- PhoP-binding sites consensus
  TGTTTANNNNNTGTTTA
- Direct binding motif repeat

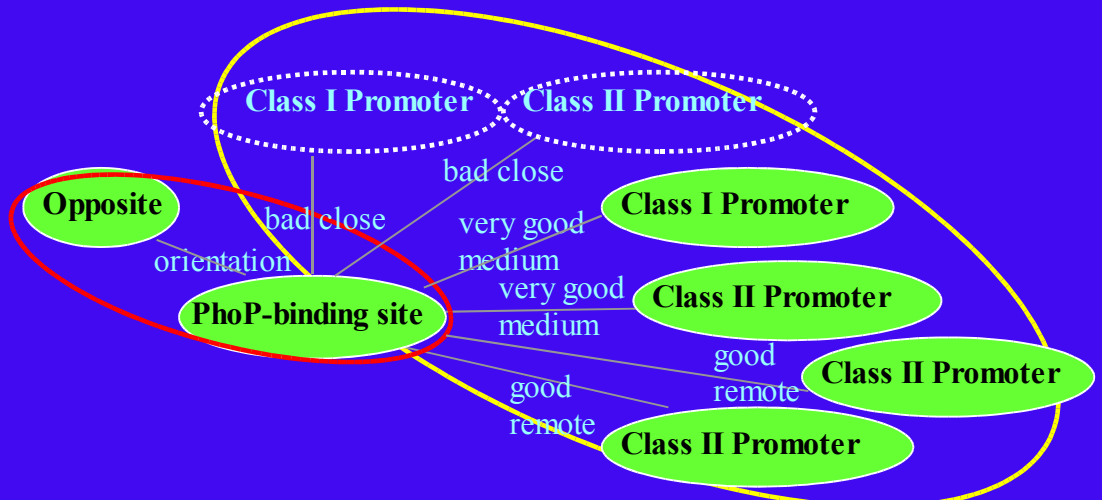
OUR ANALYSIS RECOVERED THE TYPICAL PHOP-REGULATED PROMOTER



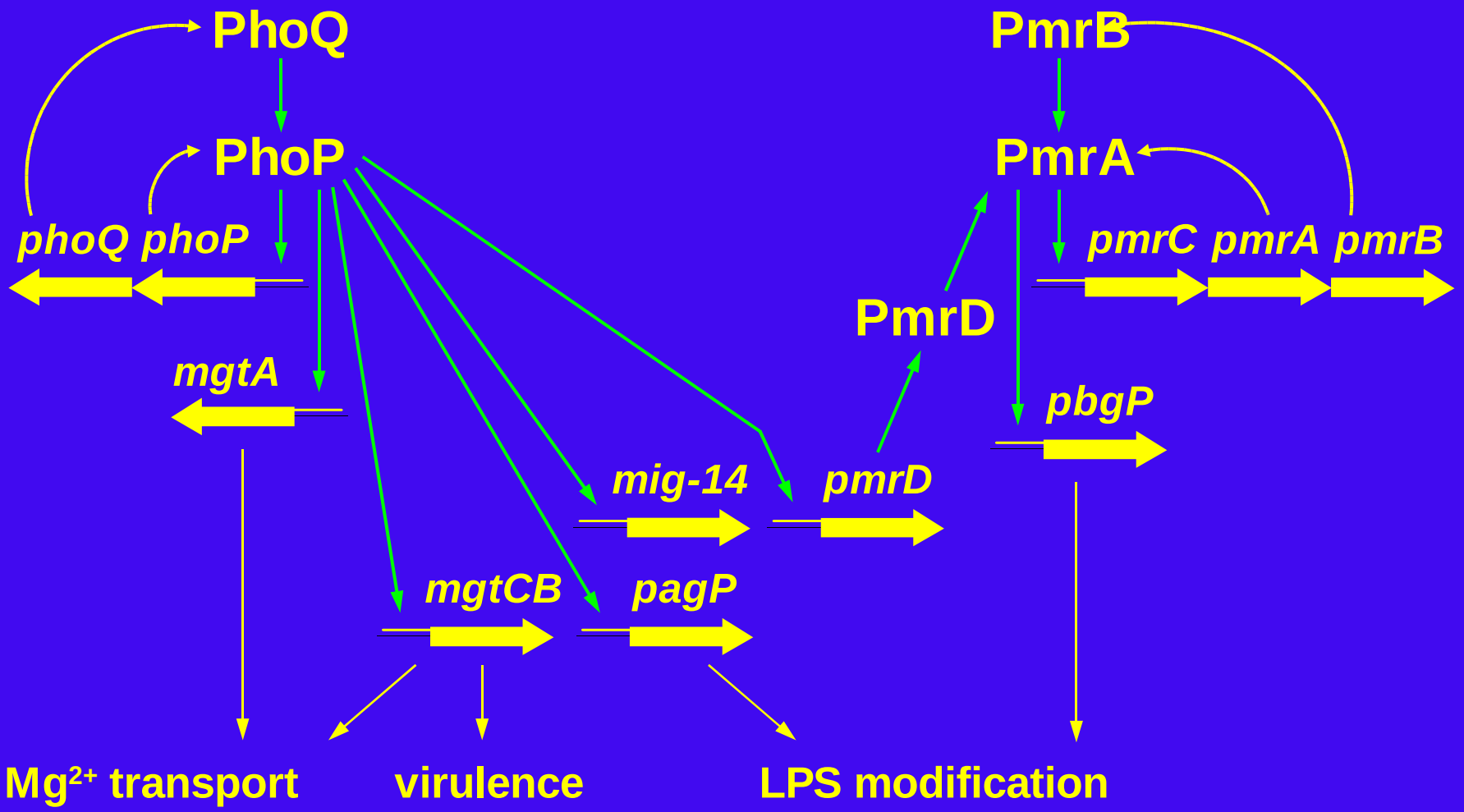


OUR ANALYSIS DISCOVERED ATIPYCAL PHOP-REGULATED PROMOTERS

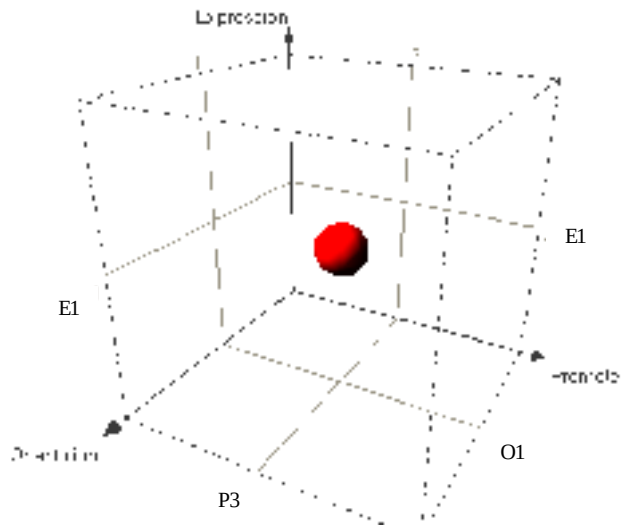


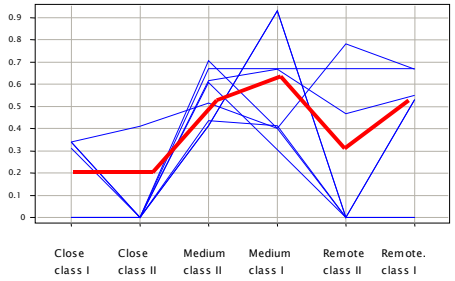


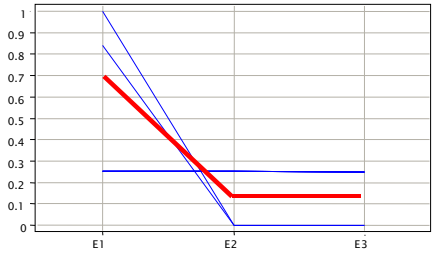
SEQUENTIAL ACTIVATION OF THE PHOP REGULON



PROMOTER –EXPRESSION –ORIENTATION PROFILES

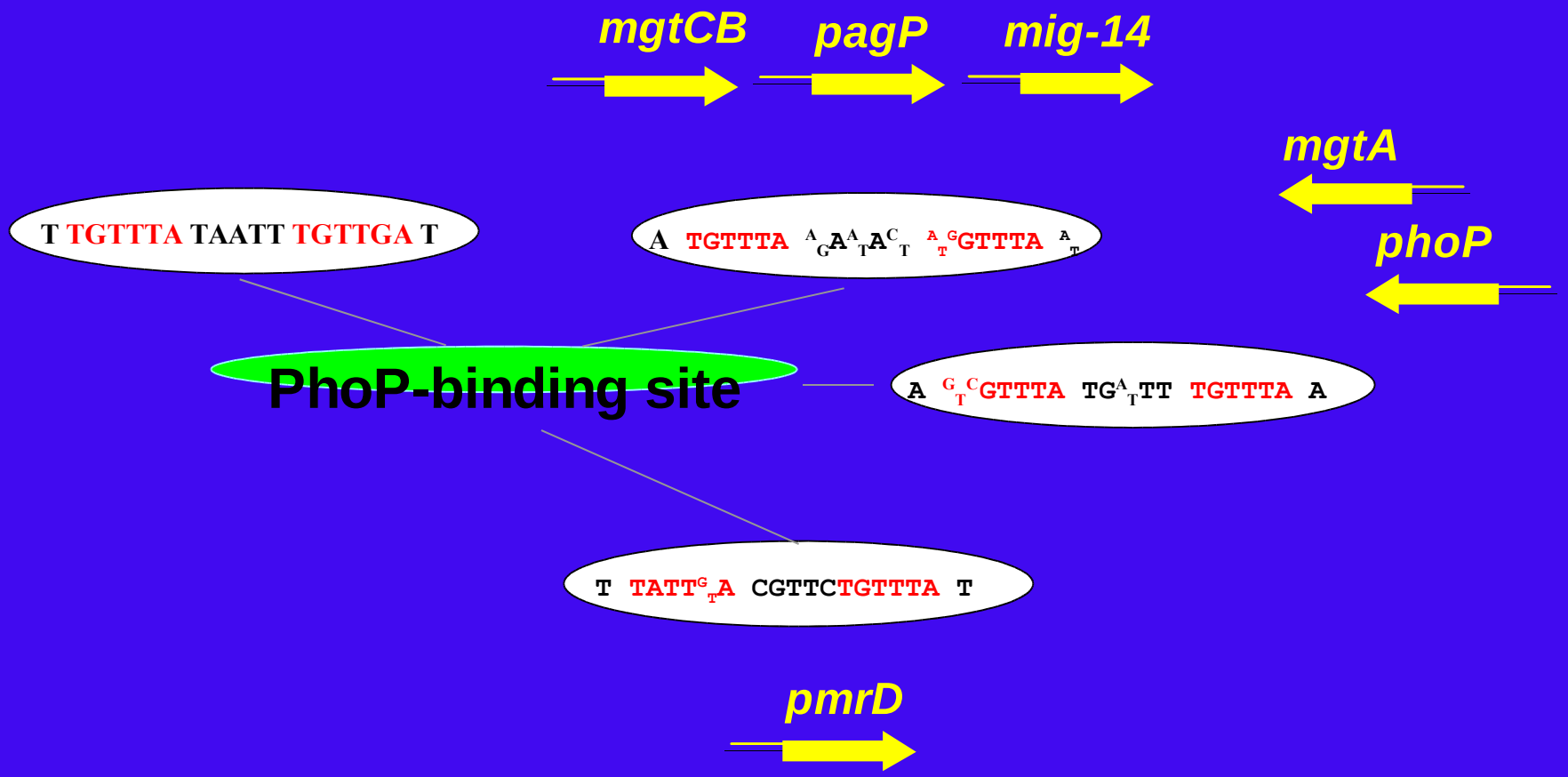




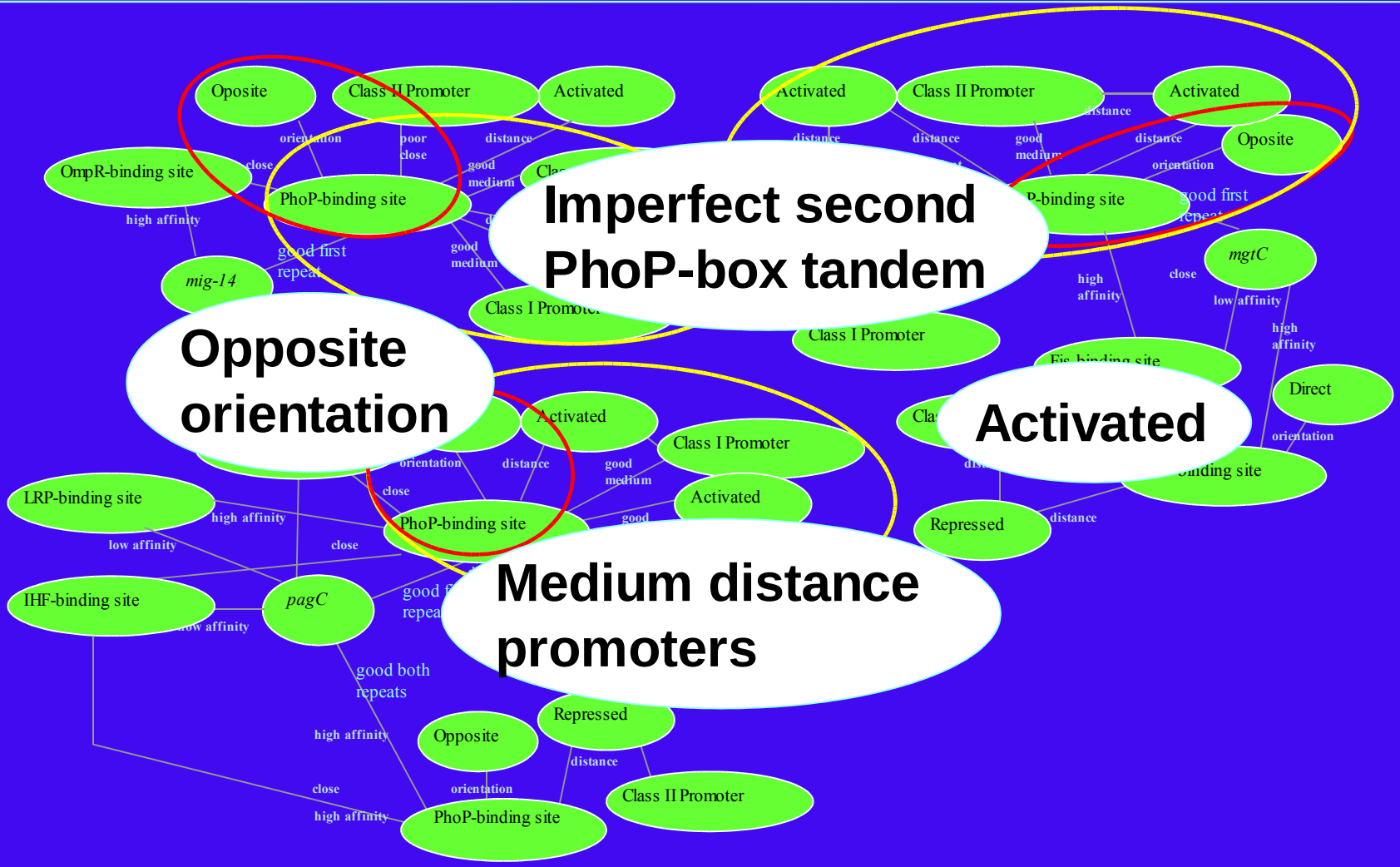


mgtC
ybjX
mig-14
slyB
pagC
yhiW
virK

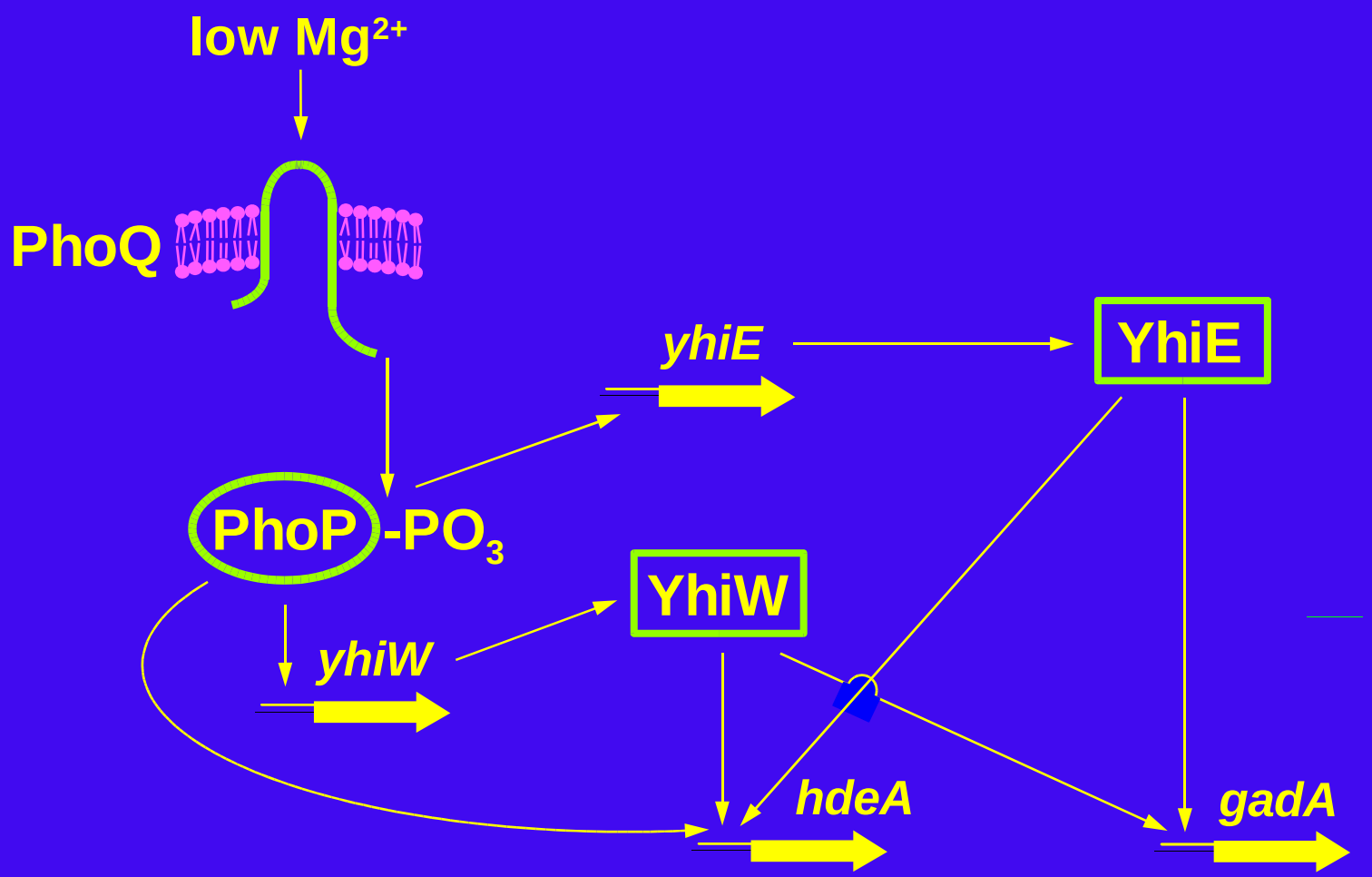
THE MINIMAL PROFILES THAT DESCRIBES THE SEQUENTIAL ACTIVATION OF THE PHOP REGULON



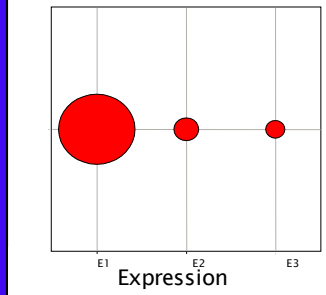
THE MAXIMAL PROFILES THAT DEFINES THE SEQUENTIAL ACTIVATION OF PHOP



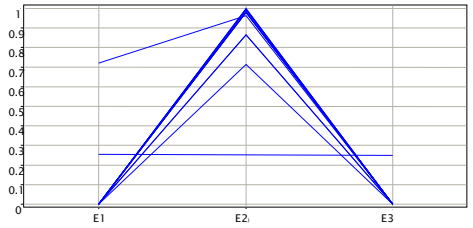
THE PHOP/PHOQ SYSTEM REGULATES EXPRESSION OF ACID pH RESISTANCE GENES IN *E. COLI*

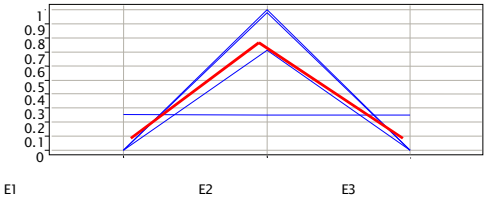


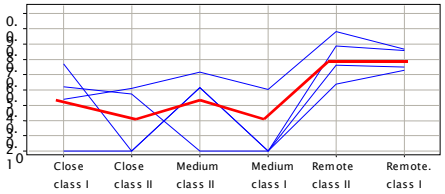
EXPRESSION AND EXPRESSION-PROMOTER PROFILES



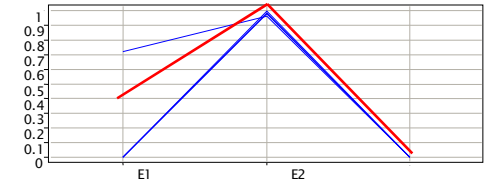
Expression-Promoter





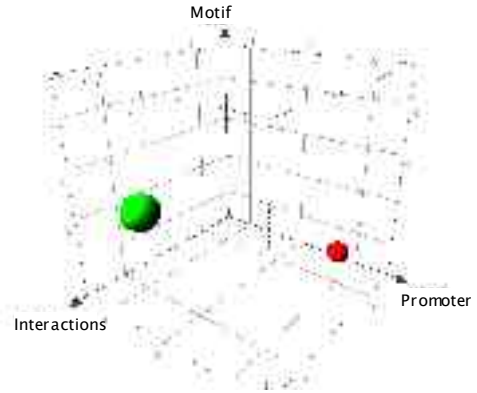


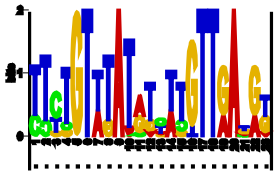
b1825
yhiE
yhiW
yiaG

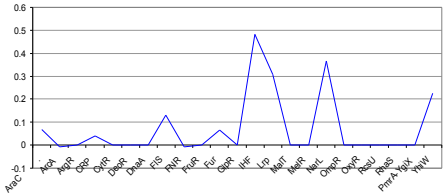


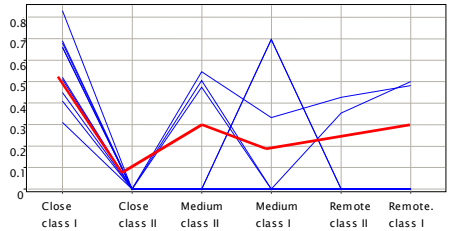
dps
gadA
gadB
ygiW

MOTIFS-PROMOTER-INTERACTION PROFILE



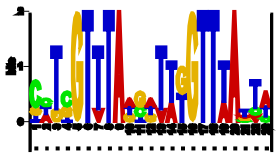


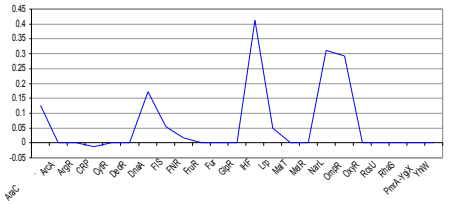


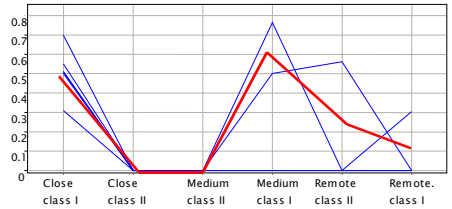


hilA
nmpC
pagP
ybjX

hdeA
hdeB
hdeD

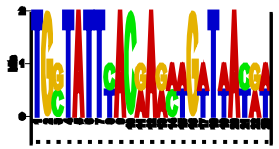


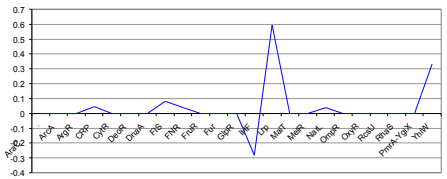


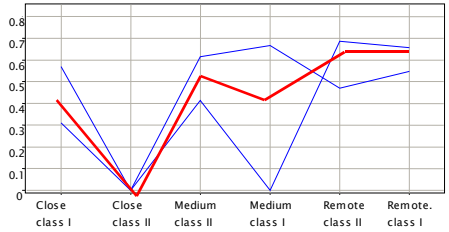


slyB
ybcU
yhiW
mgtA

phoP

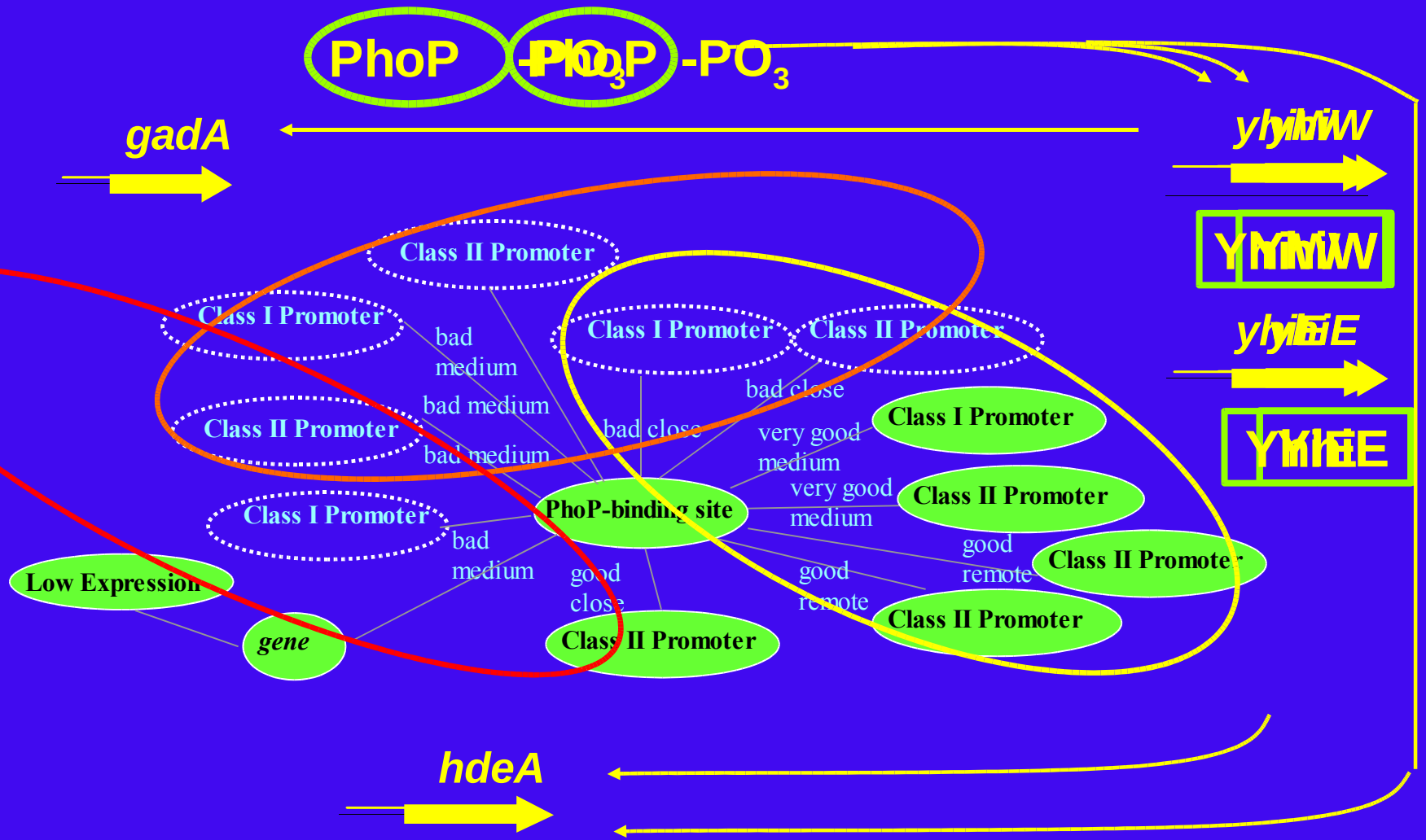






slyB
yhiE

THE PROFILE THAT DEFINES THE PHOP/PHOQ REGULATION OF ACID pH RESISTANCE GENES

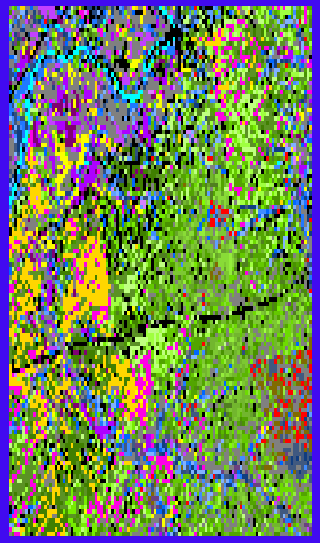


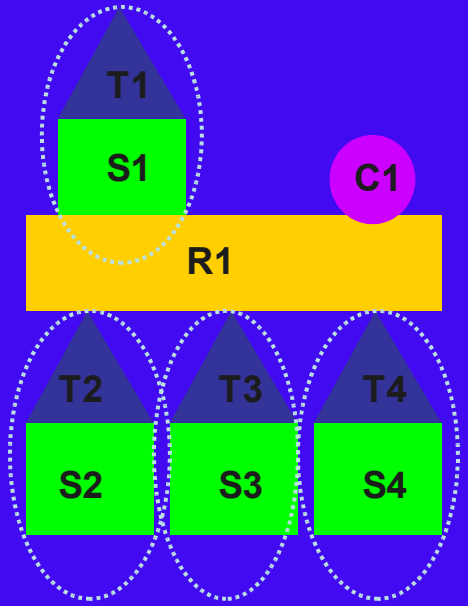
ADVANTAGES OF CONCEPTUAL CLUSTERING (Michalsky, Chesseman, Cook, Ruspini, etc.)

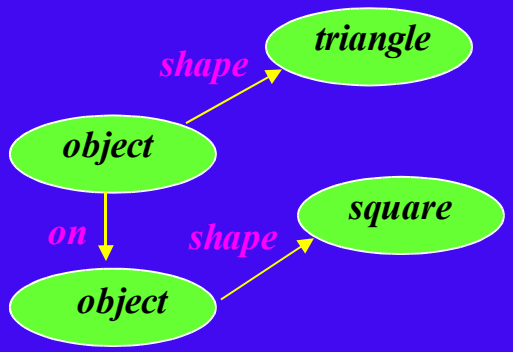
- Hypothesis generation – data exploration – model fitting – finding true topologies
- Deal with structural data
- Iterative and incremental search of interesting repetitive substructures
- Attributes and relations belong to more than one substructure
- Attributes and relations are flexible
- Ability to deal with missing values
- DISCOVER → DESCRIBE -> PREDICT

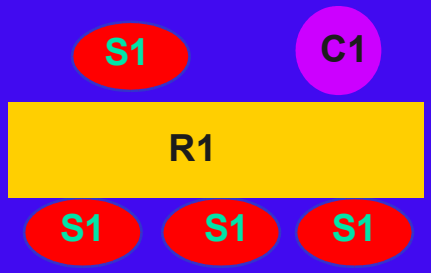
Conceptual Clustering: Substructure Discovery

Input Problem ➡ *Input Database* ➡ *Substructure* ➡ *Compressed Database*







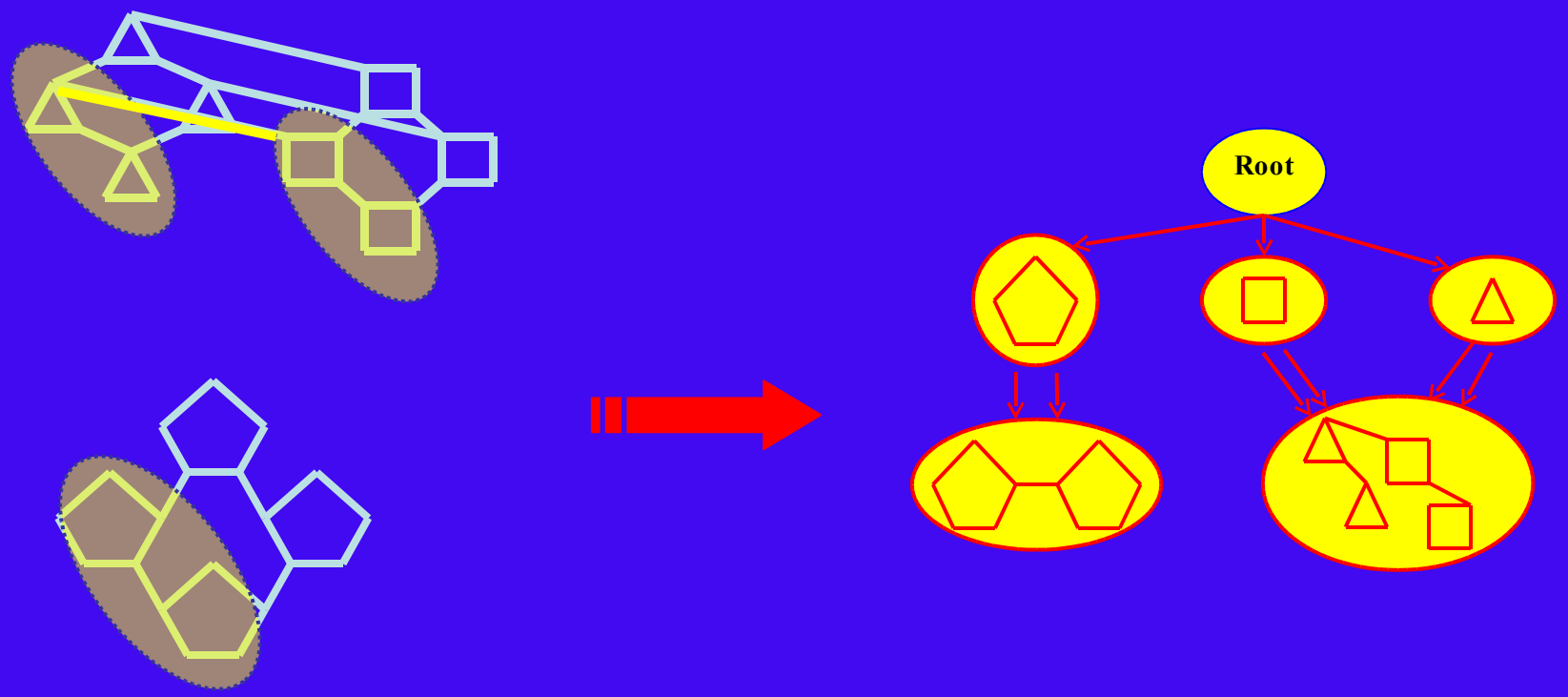


Data exploration | Group predictions | Data reduction

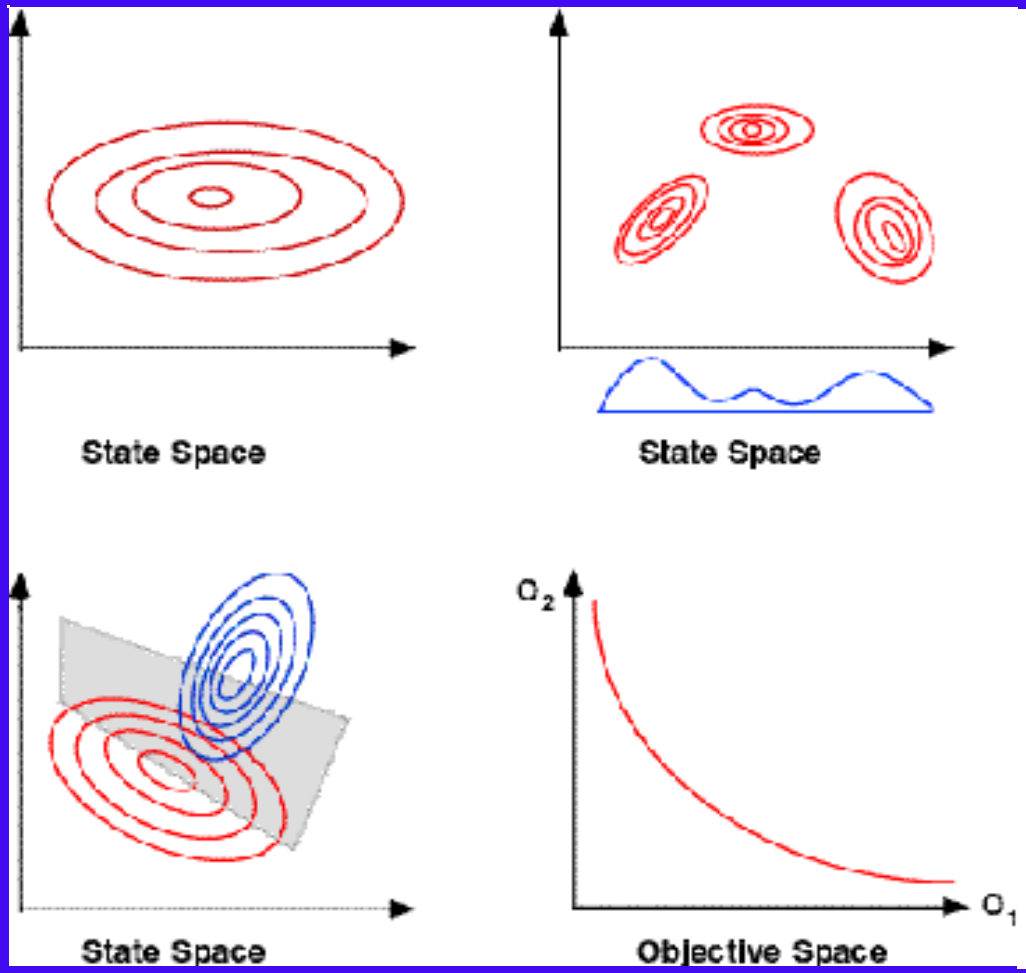
Model fitting | Hypothesis generation and testing | Finding true topologies

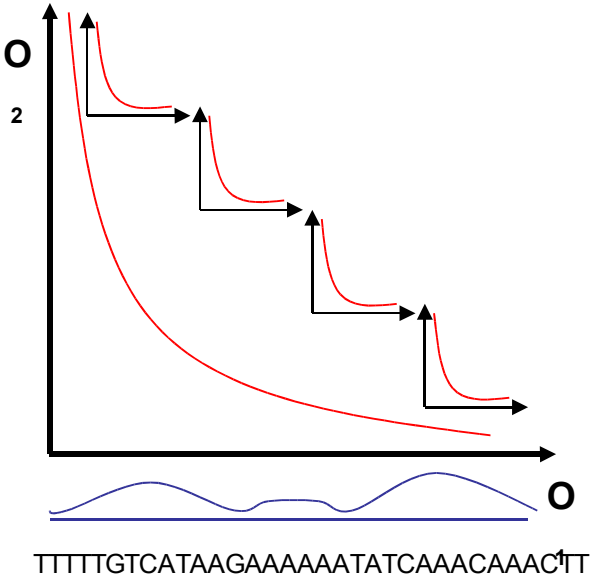
Recognize repetitive relations among attributes that describe objects in structural databases (substructure discovery)

Substructure Discovery and Evaluation Strategy by MOGA CC



MOGA CC: Local and Global Evaluation





Hybrid Promoter Analysis Methodology (HPAM)

Time Delayed Neural Network (TDDN)

+

Multi-Objective Scatter Search genetic algorithm (MOSS)

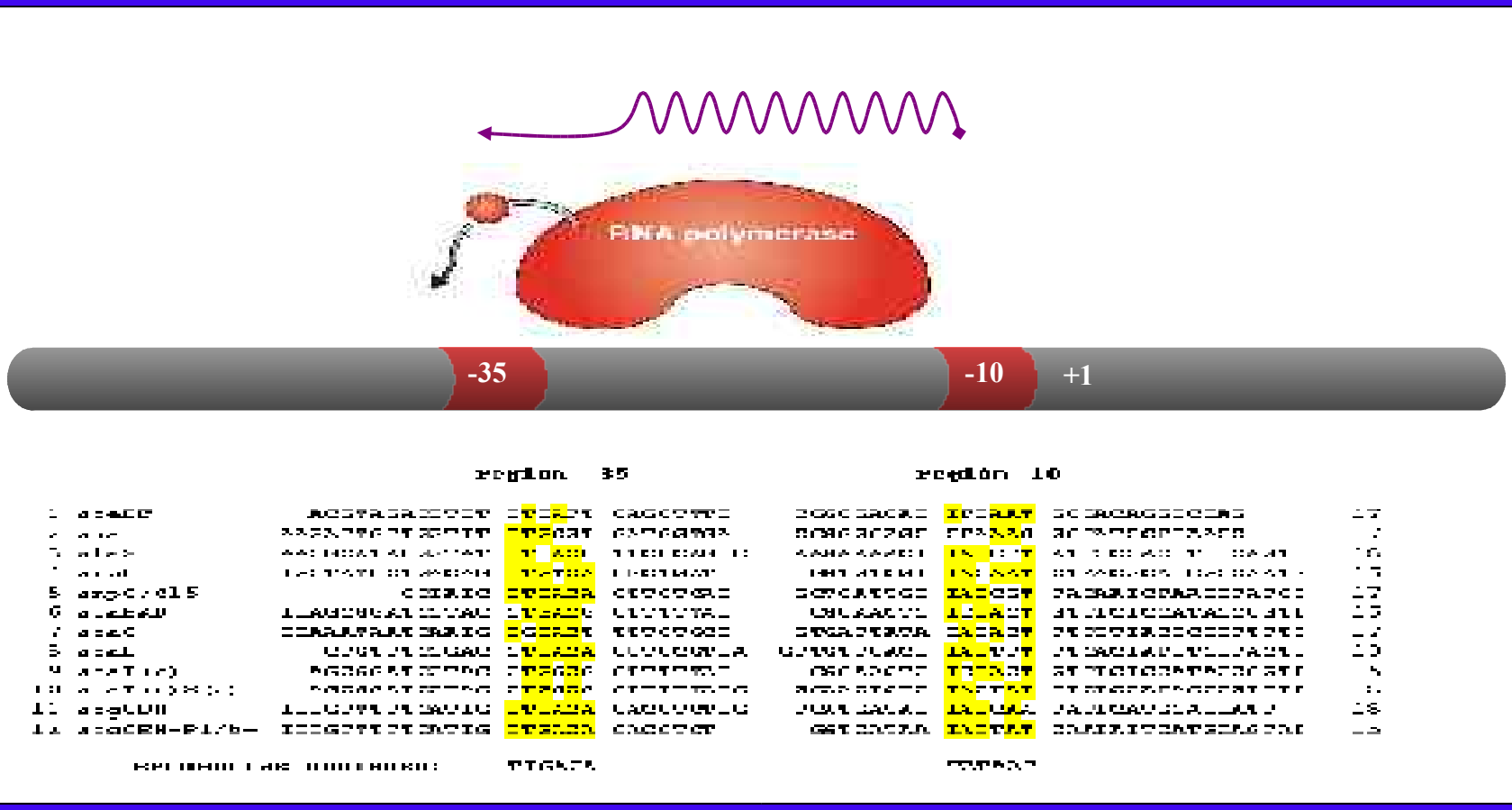
BINDING SITES SUBMOTIFS

```
#5 TTGTTTATGATTTGTTTA
A |   0   6   3   1   0  21   25   25  25   25   25   5   4   1   1   6  21
C |   3   0   0   3   1   3   25   25  25   25   25   1   0   2   0   0   2
G |   6  19   2   0   4   0   25   25  25   25   25   9  19   0   4   4   2
T |  16   0  20  21  20   1   25   25  25   25   25  10   2  22  20  15   0
```

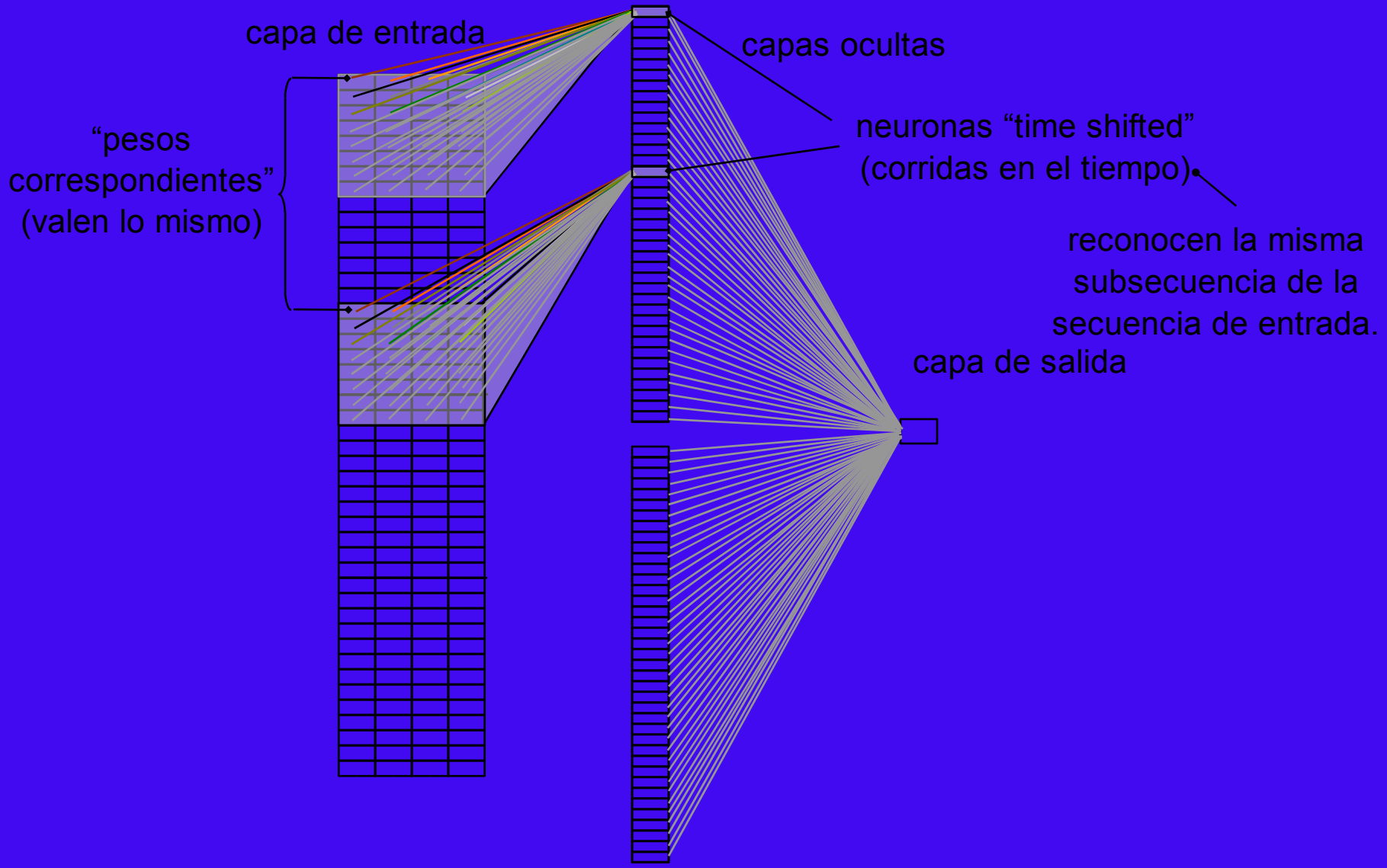
TGTTTA_ _ _ _ _ TGTTTA

M_1	T	TATT$^{G}_{T}$A	C	GTT	C	TGTTTA	T
M_2	A	$^{G}_{T}{}^{C}$GTTTA	T	G$^{A}_{T}$T	T	TGTTTA	A
M_3	A	TGTTTA	$^{A}_{G}$	A$^{A}_{T}$A	$^{C}_{T}$	$^{A}_{T}{}^{G}$GTTTA	$^{A}_{T}$
M_4	T	TGTTTA	T	AAT	T	TGTTGA	T

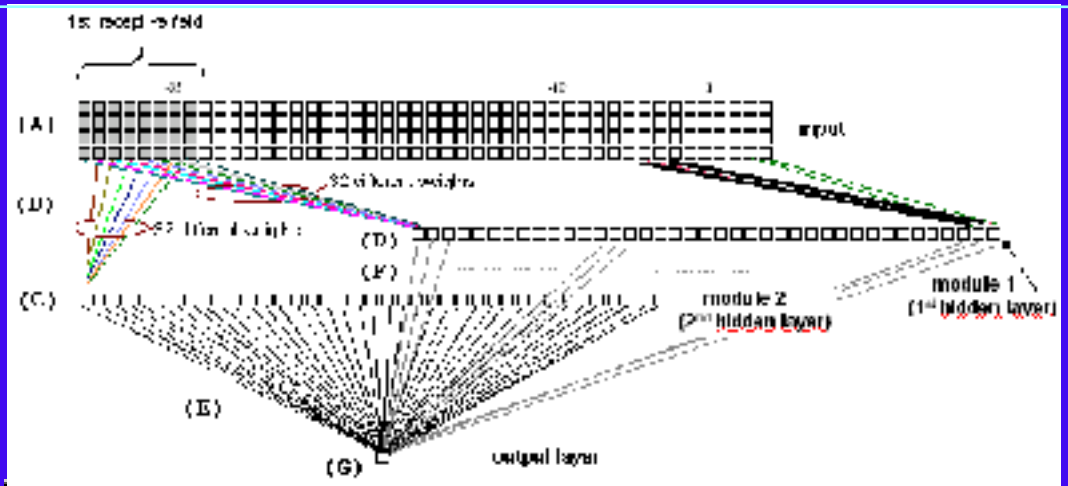
PROMOTER FEATURES: RNA POLYMERASE, CLASS AND LOCATION

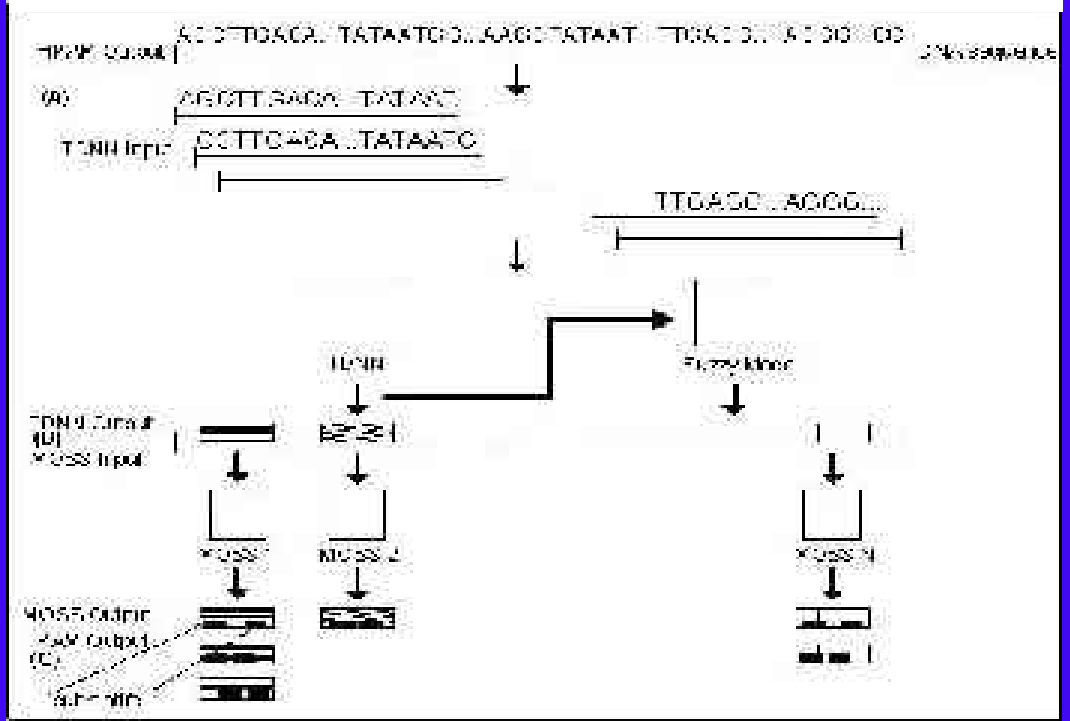


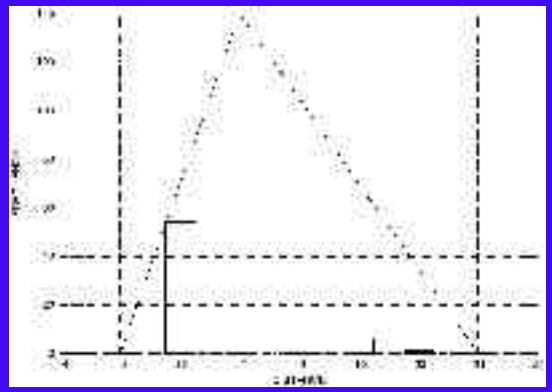
TDNN: A WEIGHT SHARING NEURAL NETWORK

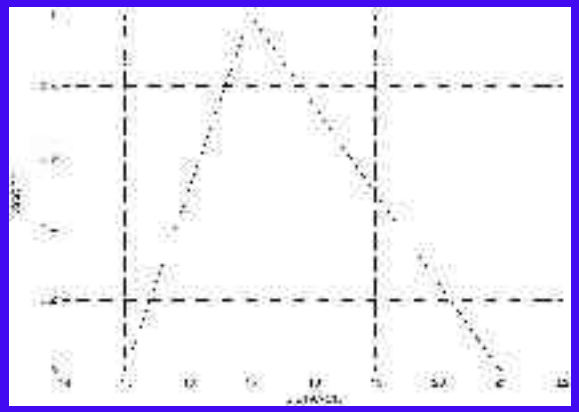


MOSS: A MULTIOBJECTIVE GENETIC FUZZY SYSTEM



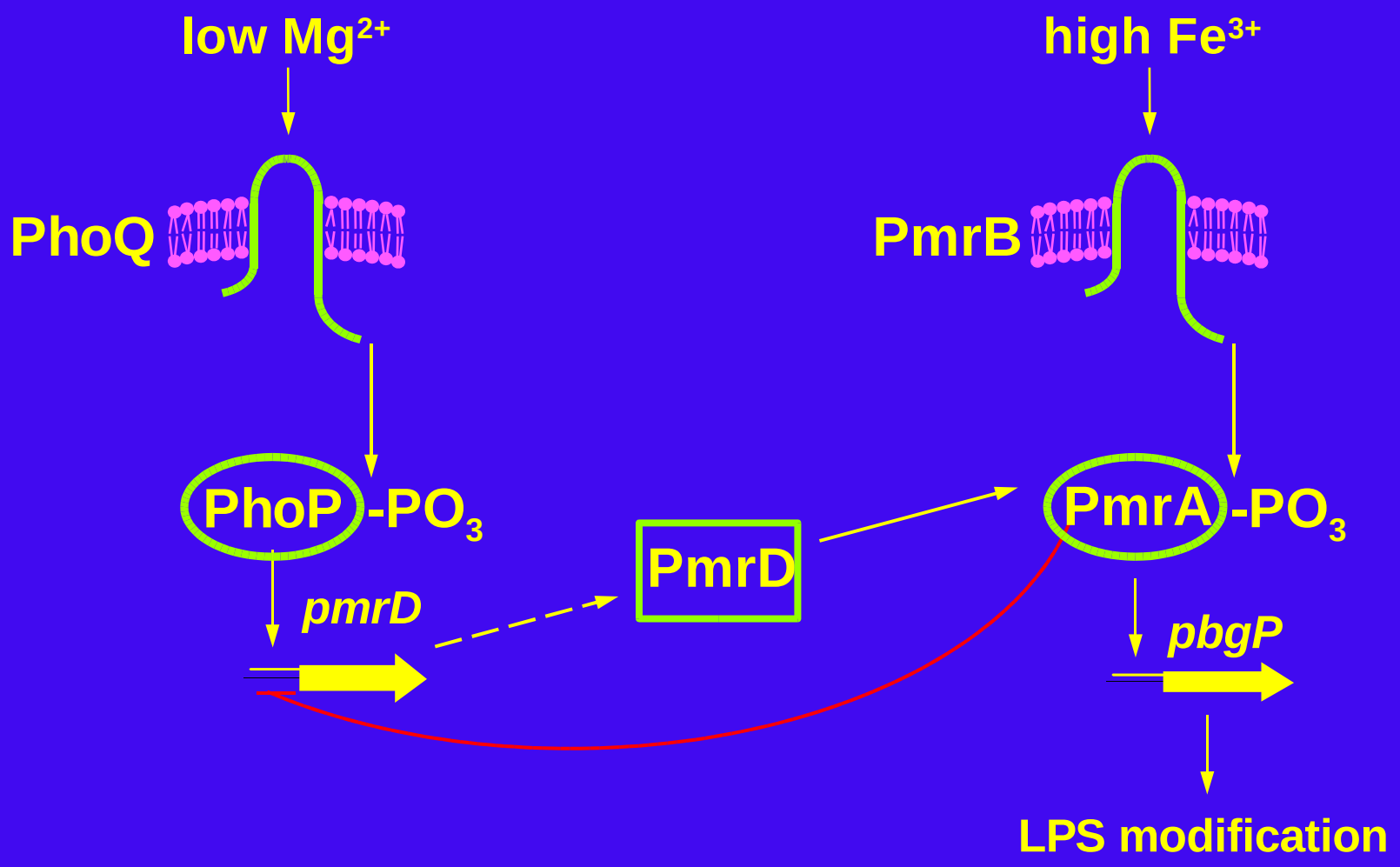




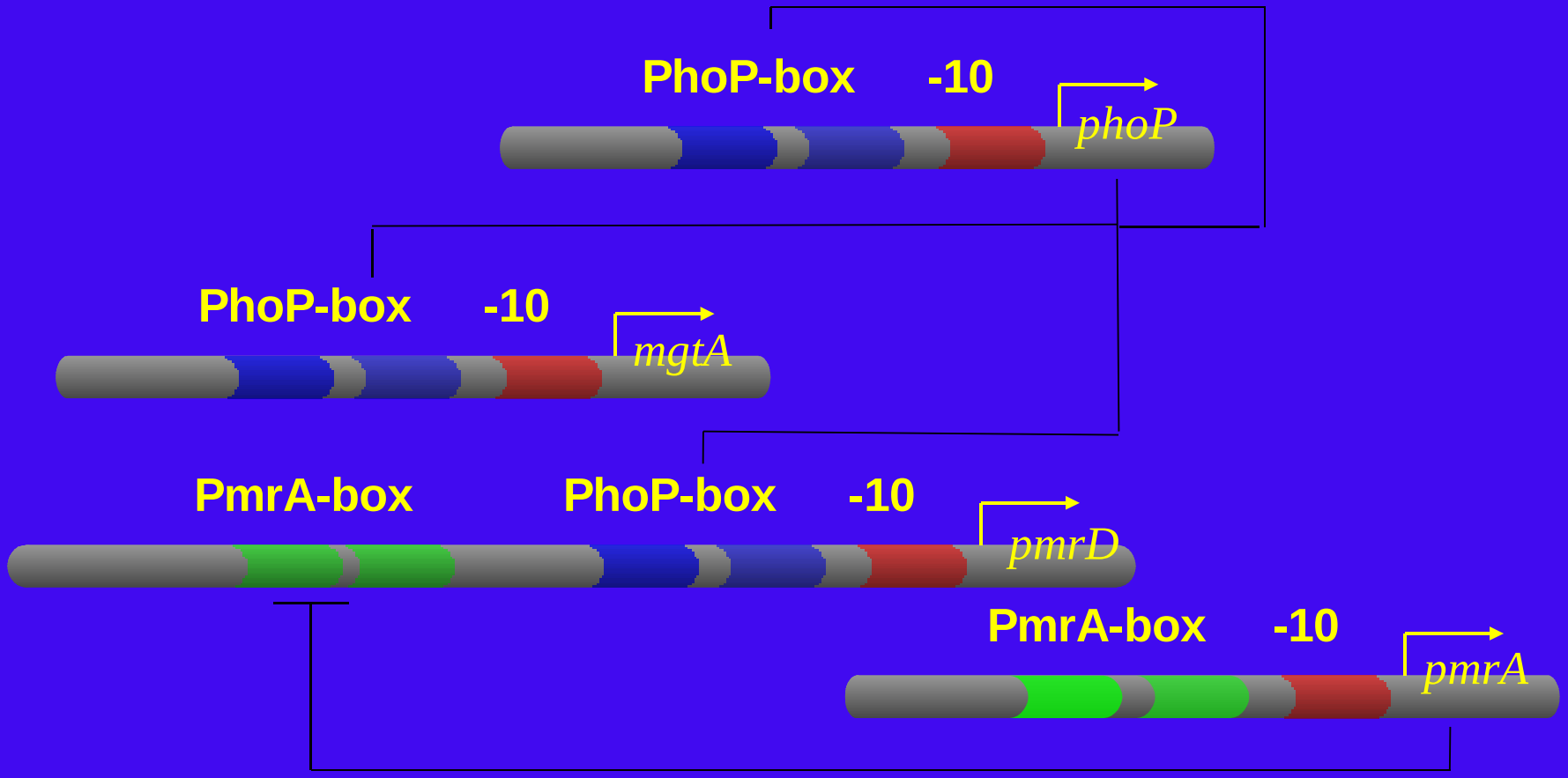


Gene Expression Networks Iterative Explorer (GENIE)

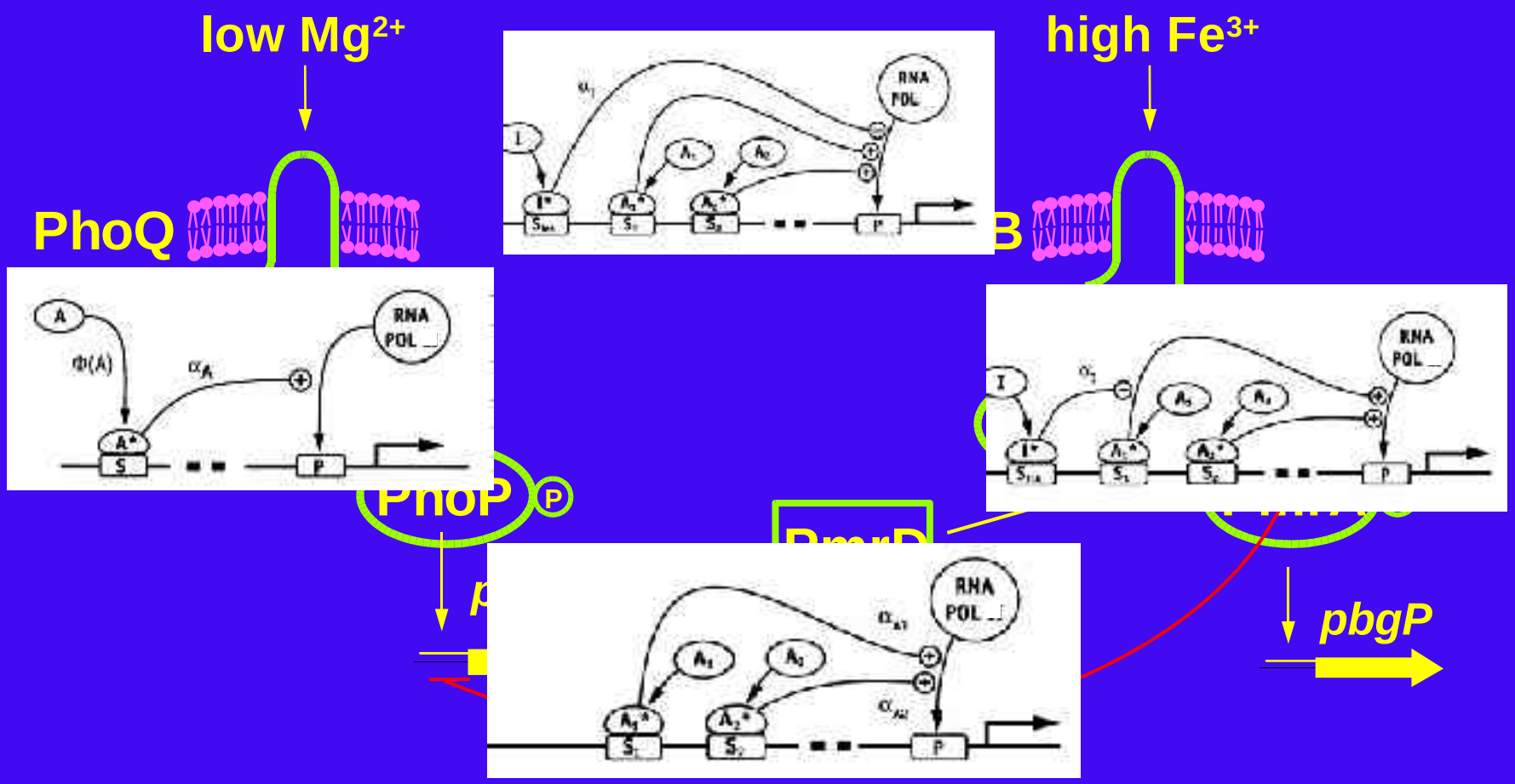
Promoter binding dynamics reveals architecture of small bacterial regulatory networks



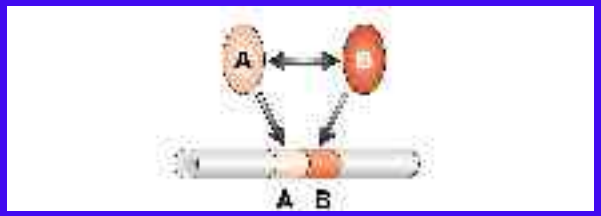
GENERATING GENETIC CIRCUITS BY GPS NAVIGATION

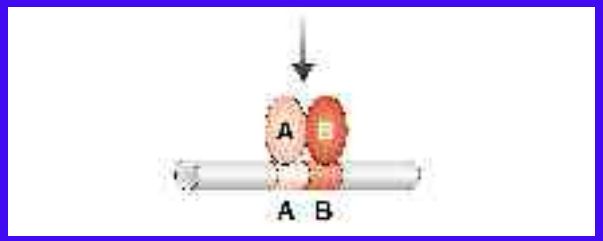


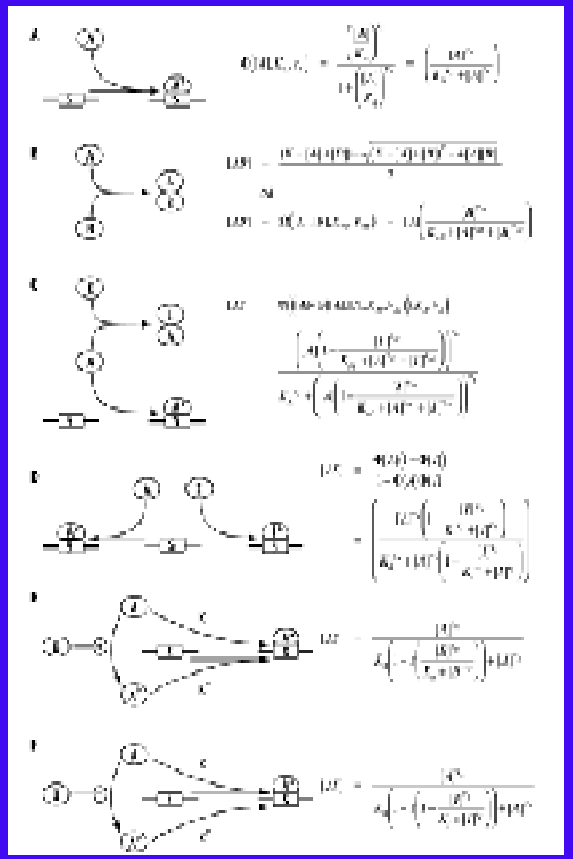
THE USE OF CONCEPTUAL CLUSTERING FINDINGS AS A HYPOTESIS GENERATOR



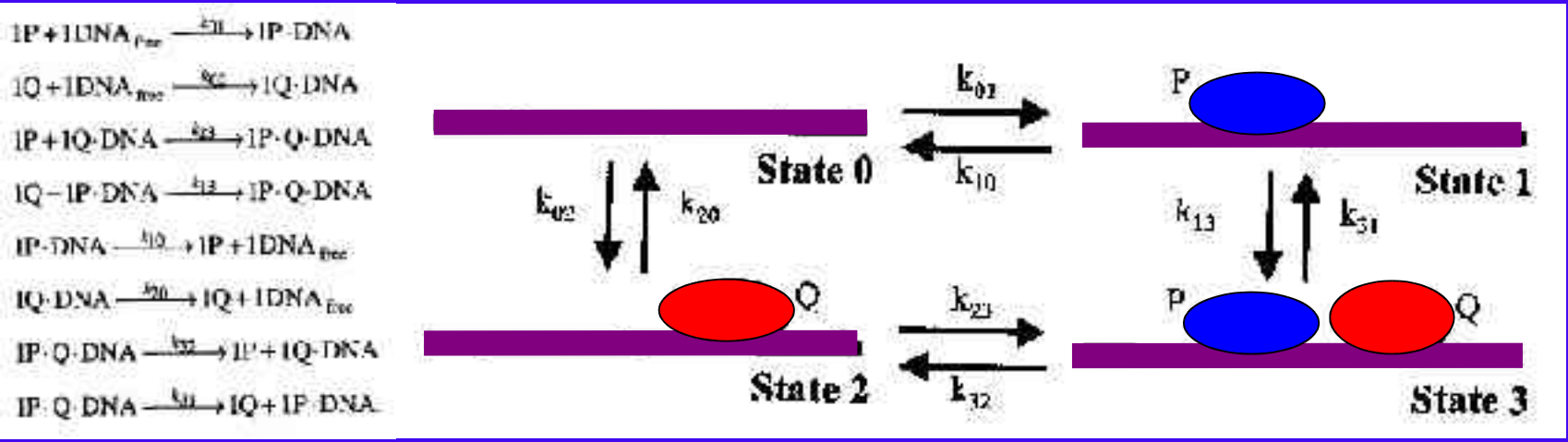
GENE PROMOTER ACTIVITY AS BUILDING BLOCK INTERACTIONS





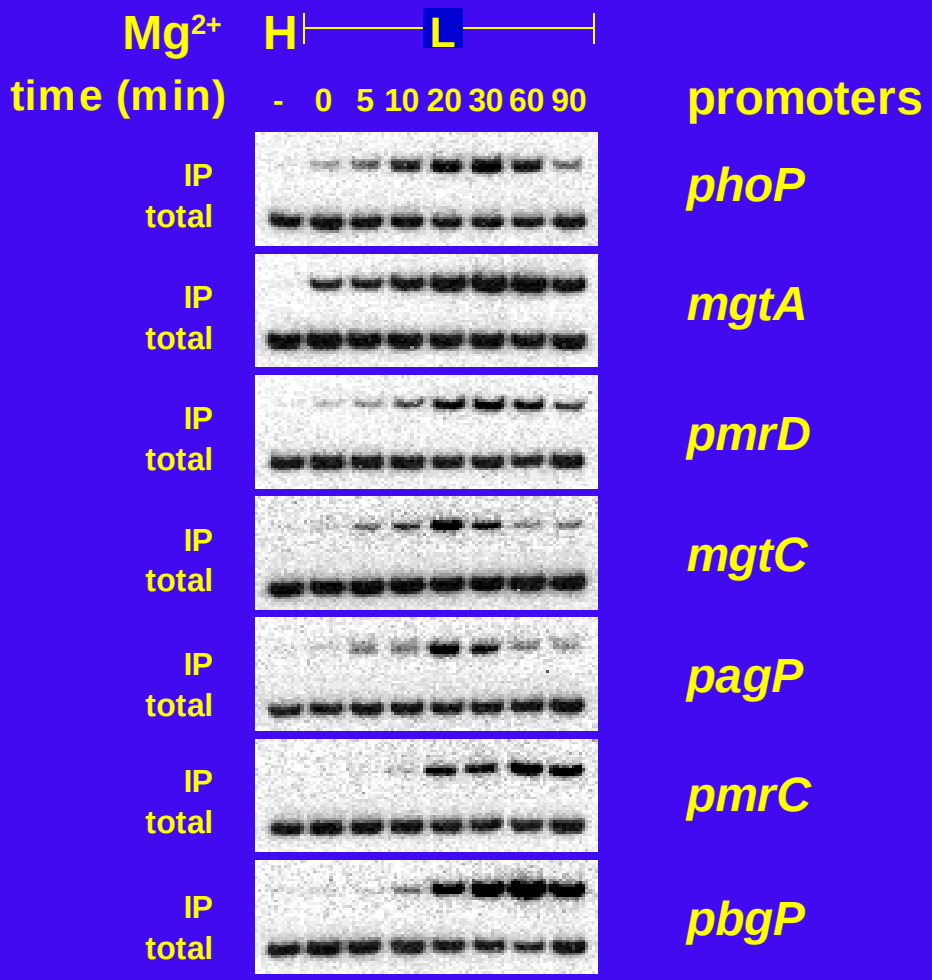


KINETIC MODEL OF TWO PROTEINS P AND Q BINDING TO DNA

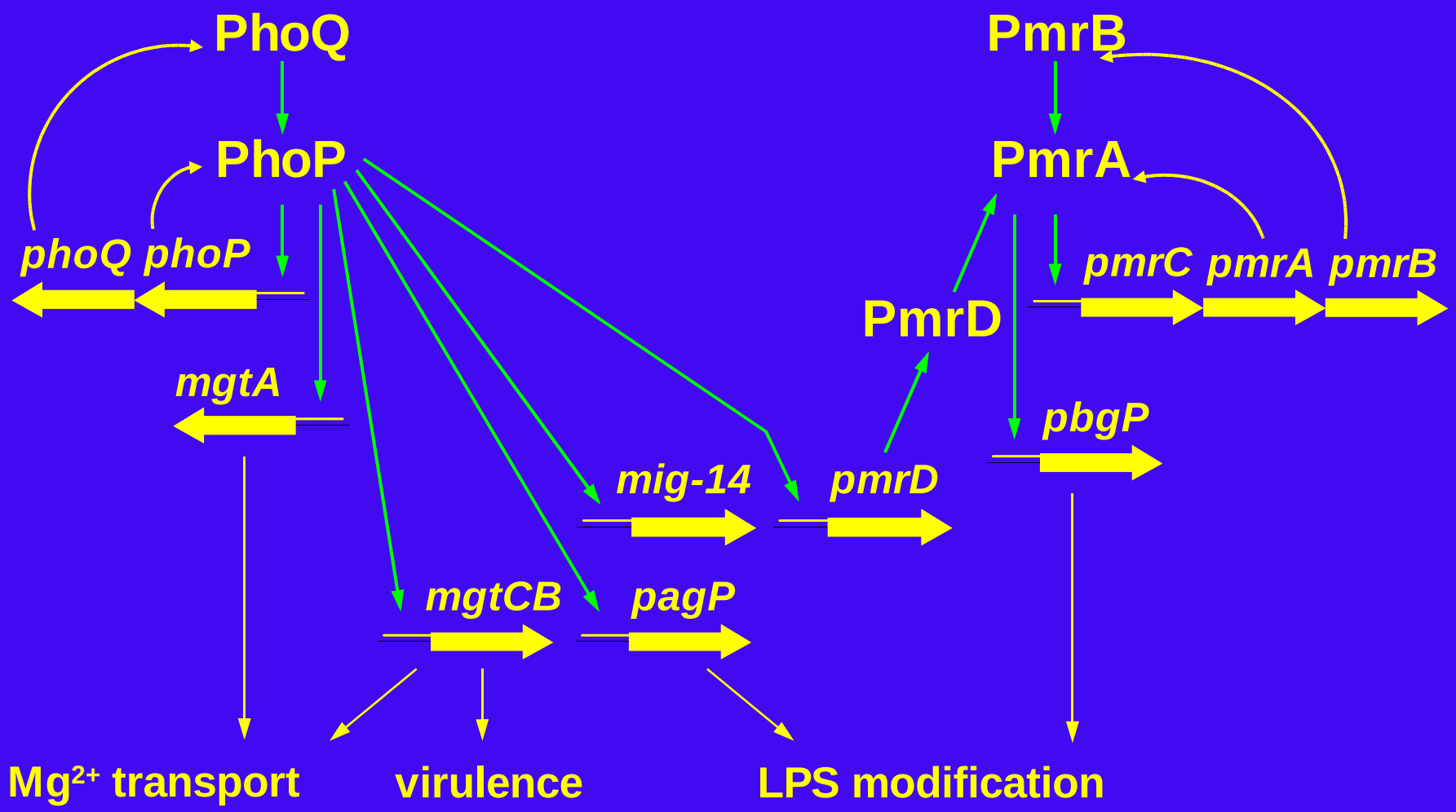


(Bower and Bolouri, 2001)

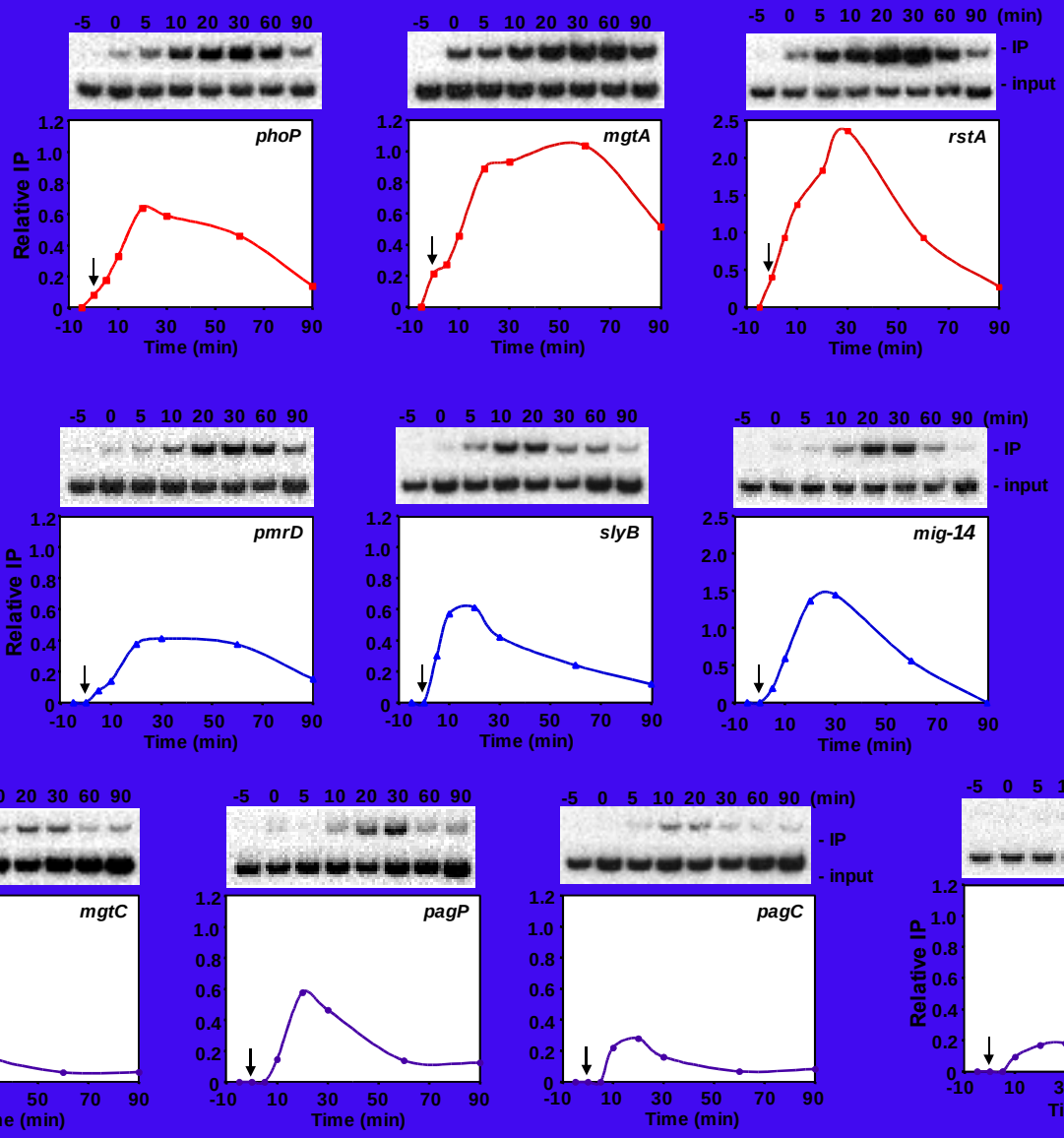
ORDERED BINDING OF THE PHOP AND PMRA PROTEINS TO THEIR TARGET PROMOTERS



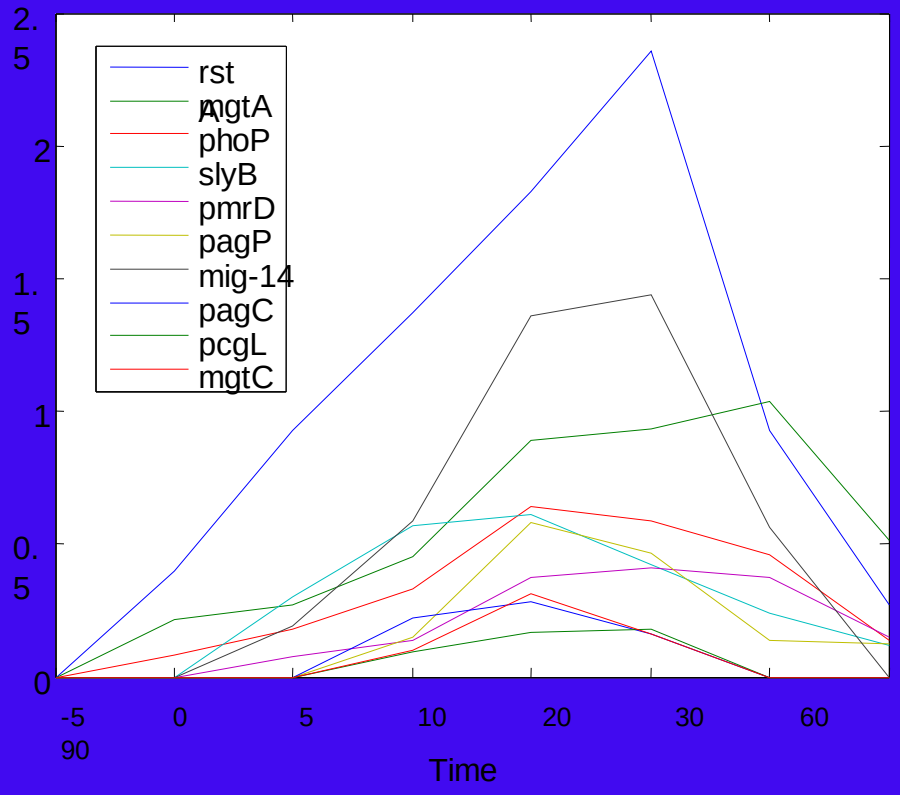
SEQUENTIAL ACTIVATION OF THE PHOP REGULON

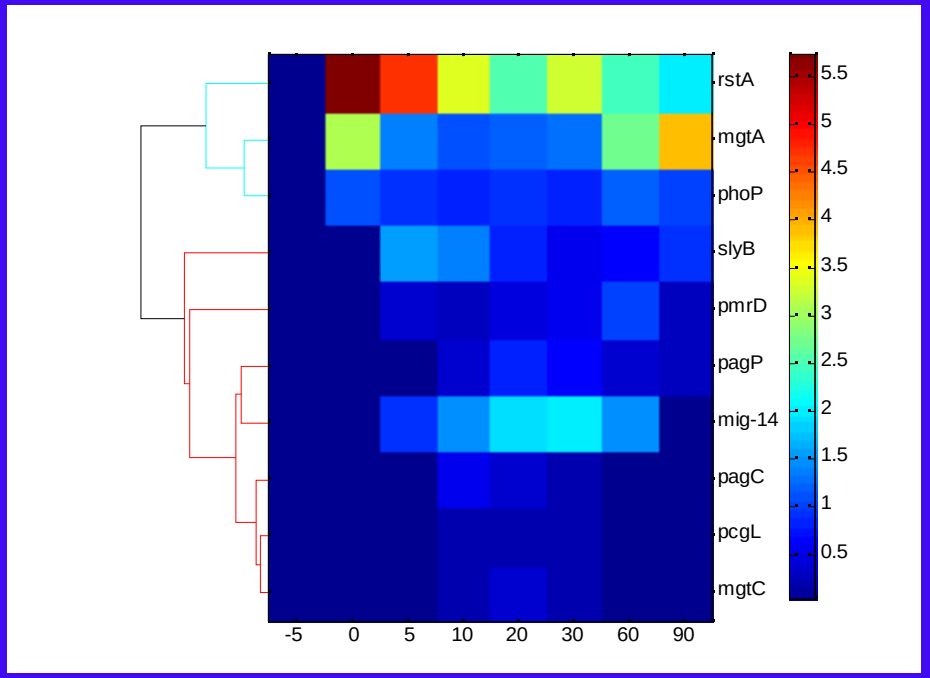


ORDERED PHOP BINDING AND TRANSIENT OCCUPANCY OF PROMOTERS



TEMPORAL ORDER OF BINDING





PROMOTER ACTIVITY AND PREDICTIONS FROM MECHAELIS-MENTEN KINETIC MODEL

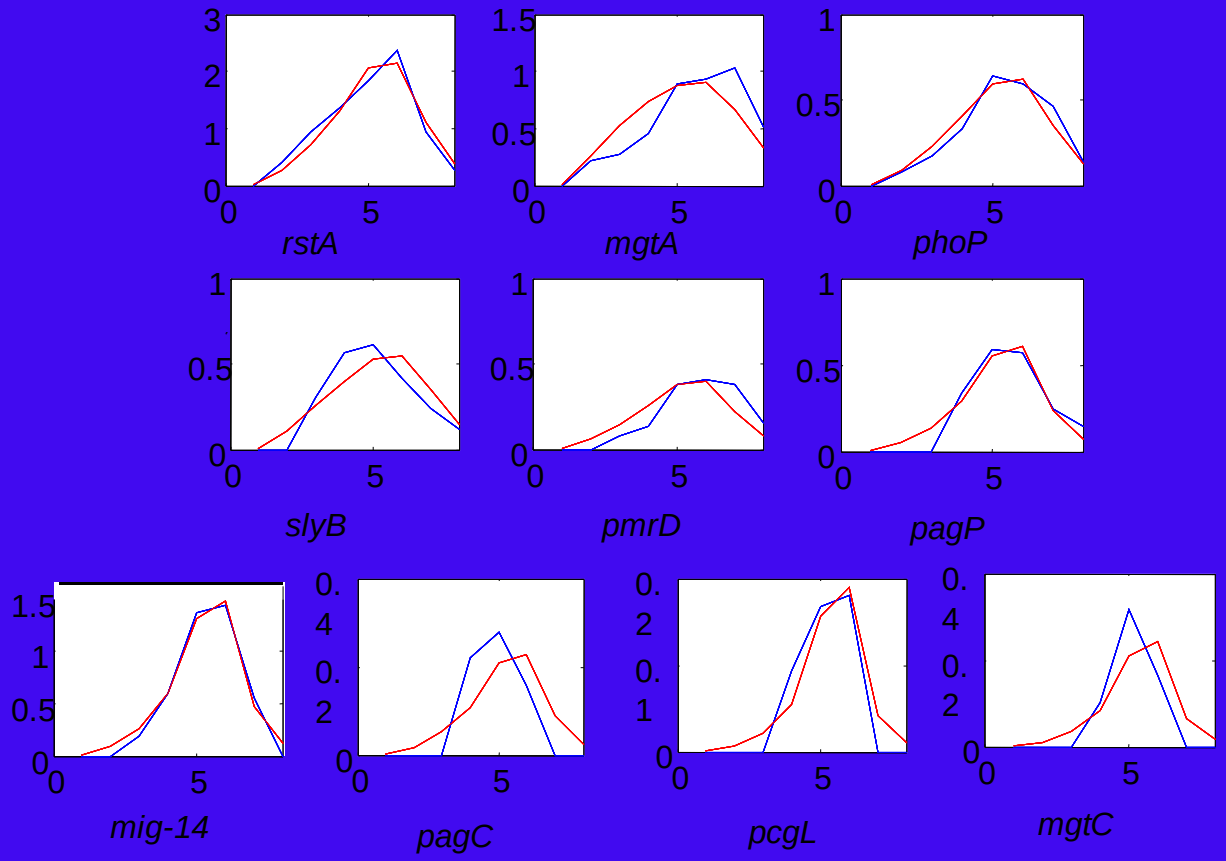
$$X_{ij}(t) = \frac{\beta_i A_j(t) / k_i}{1 + A_j(t) / k_i}$$

rate of the unrepressed promoter β_i

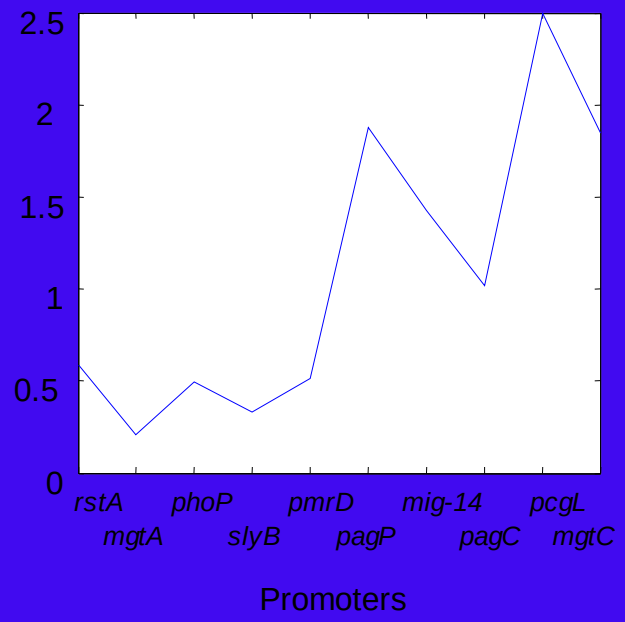
the effective affinity of the activator (concentration at half maximal activation) k_i

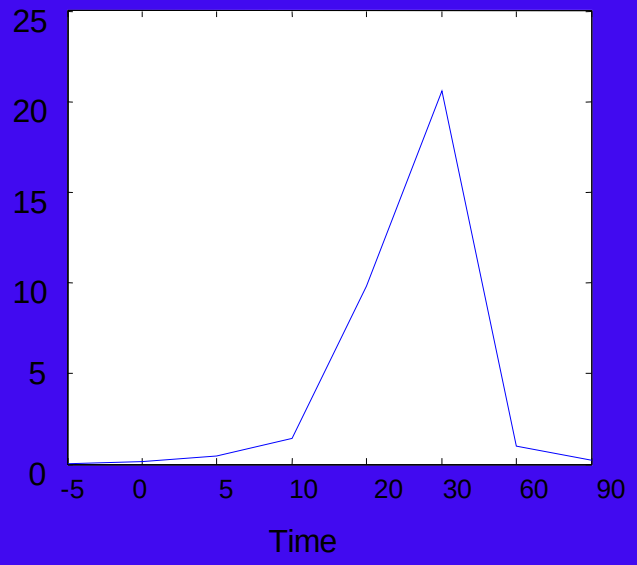
the effective activation concentration in experiment *j* for all promoters $A_j(t)$

TEMPORAL ORDER OF BINDING WITH LEARNED ACTIVATION CONCENTRATION

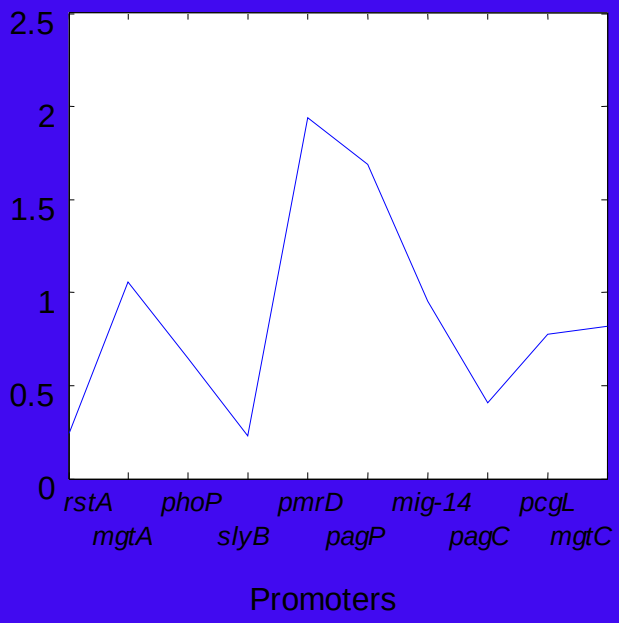


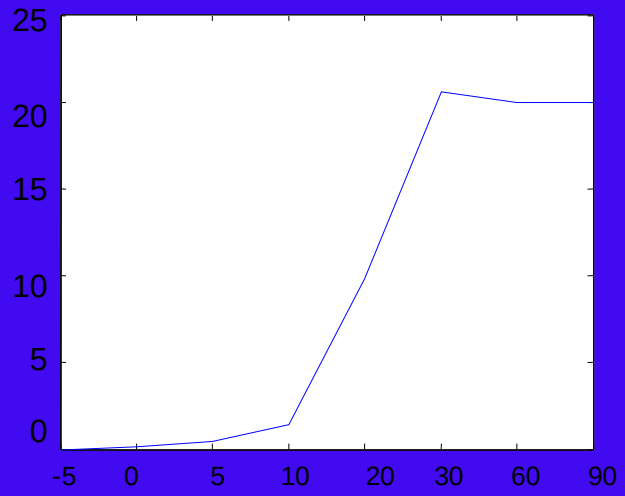
EFECTIVE PROMOTER RATE: ORDERED RESULTS



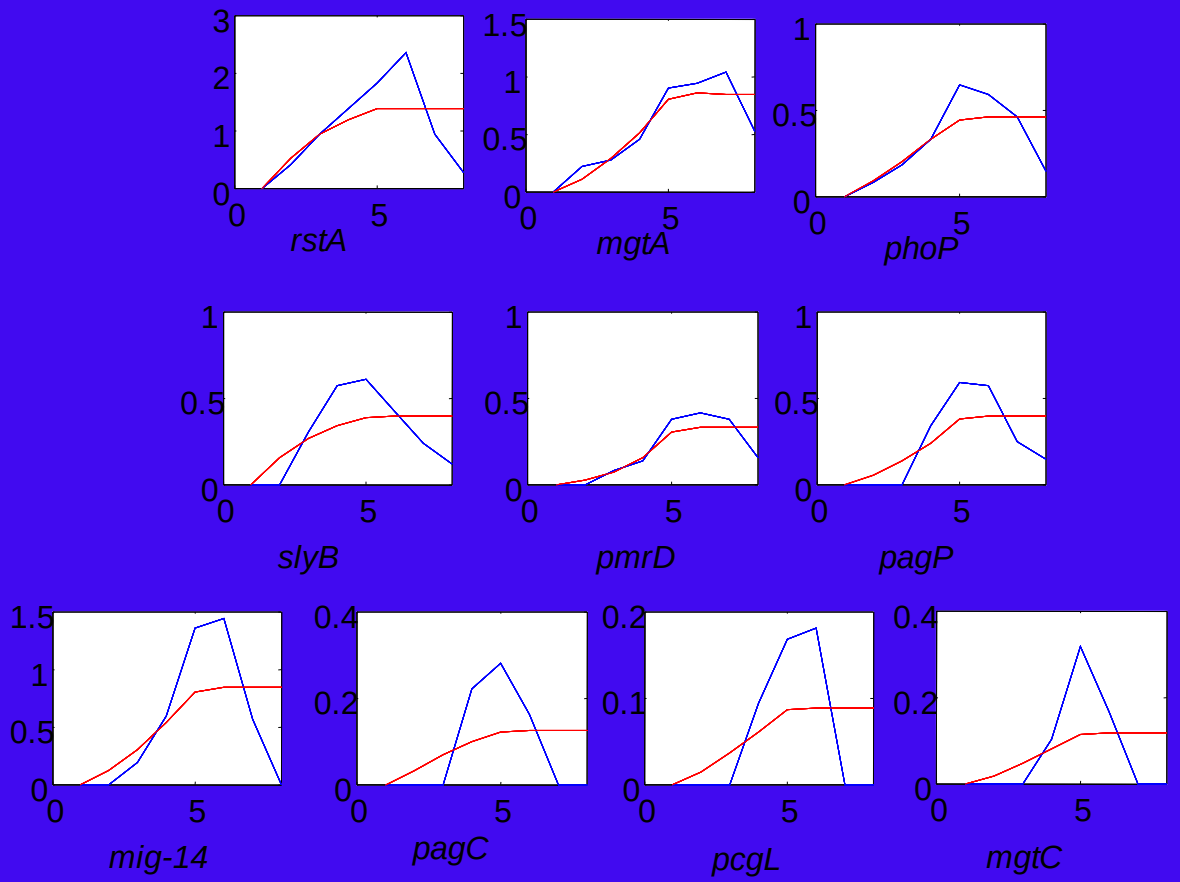


EFFECTS OF THE ACTIVATION CONCENTRATION

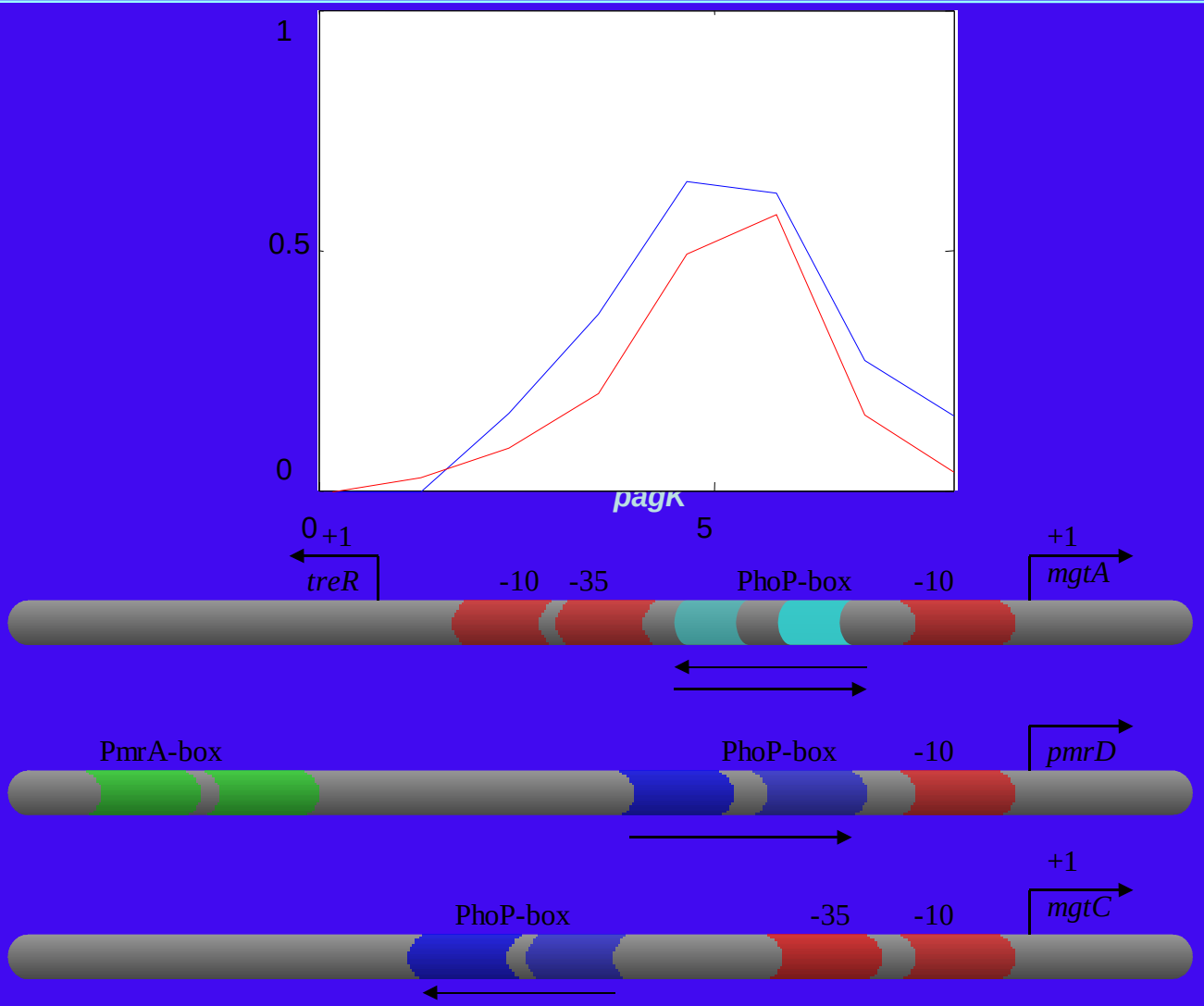




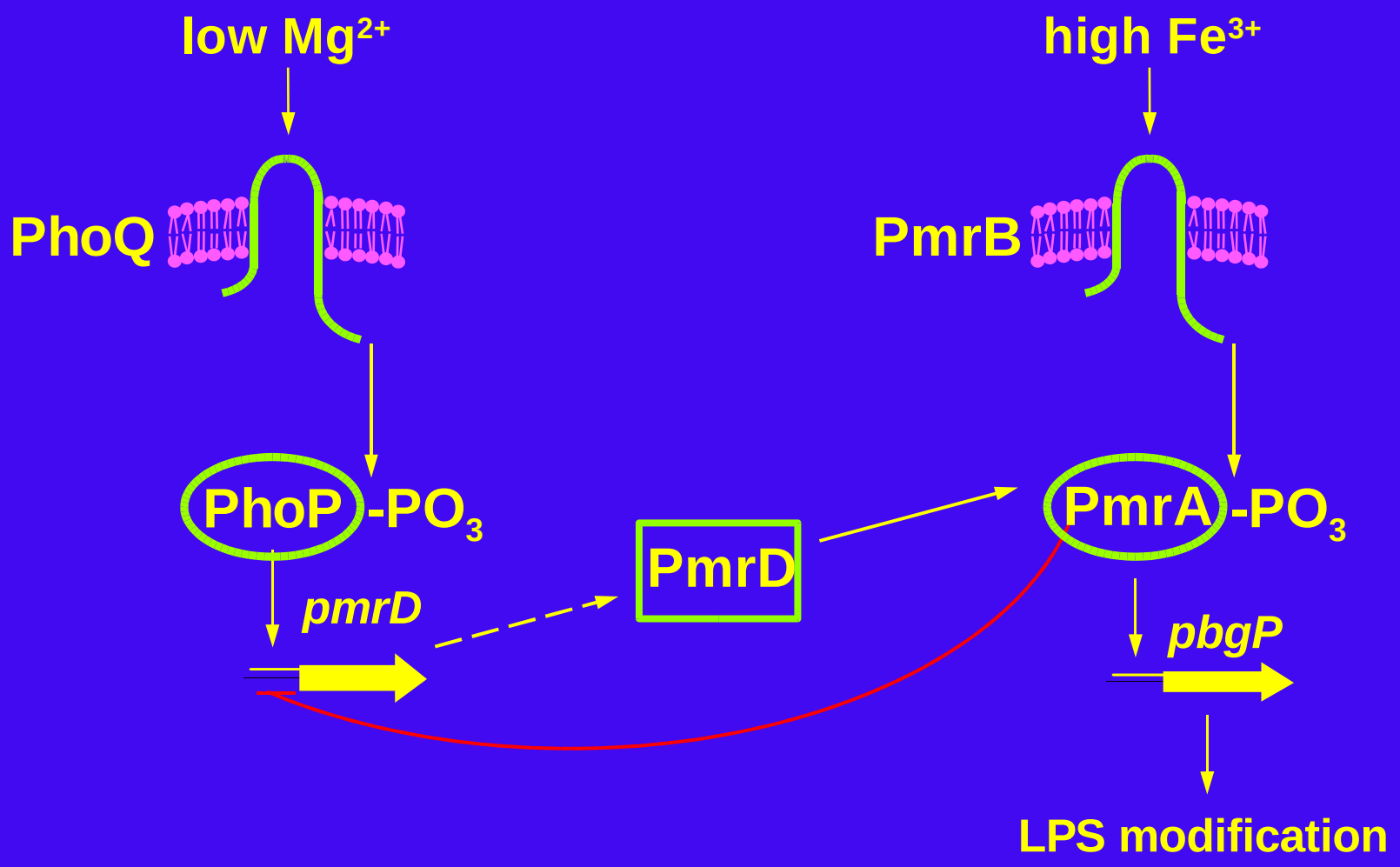
TEMPORAL ORDER OF BINDING WITH EXPERT-BASED CONCENTRATION



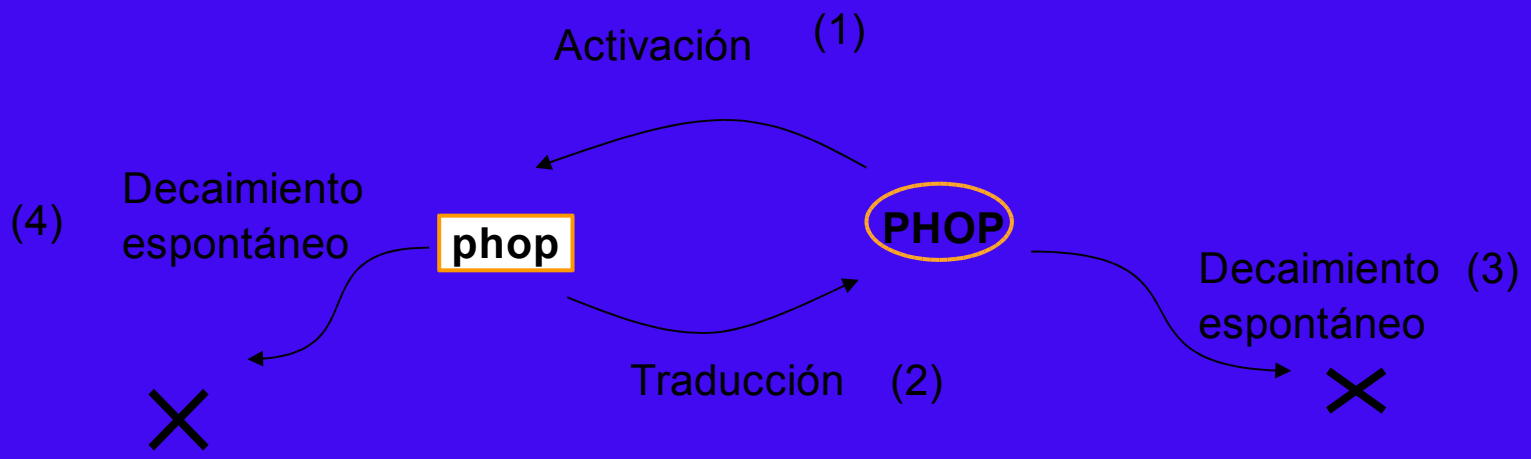
TEMPORAL ORDER OF BINDING: PREDICTION BASED ON REGULATORY FEATURES



Fe^{3+} REPRESSES *pmrD* EXPRESSION VIA THE PMRA/PMRB SYSTEM



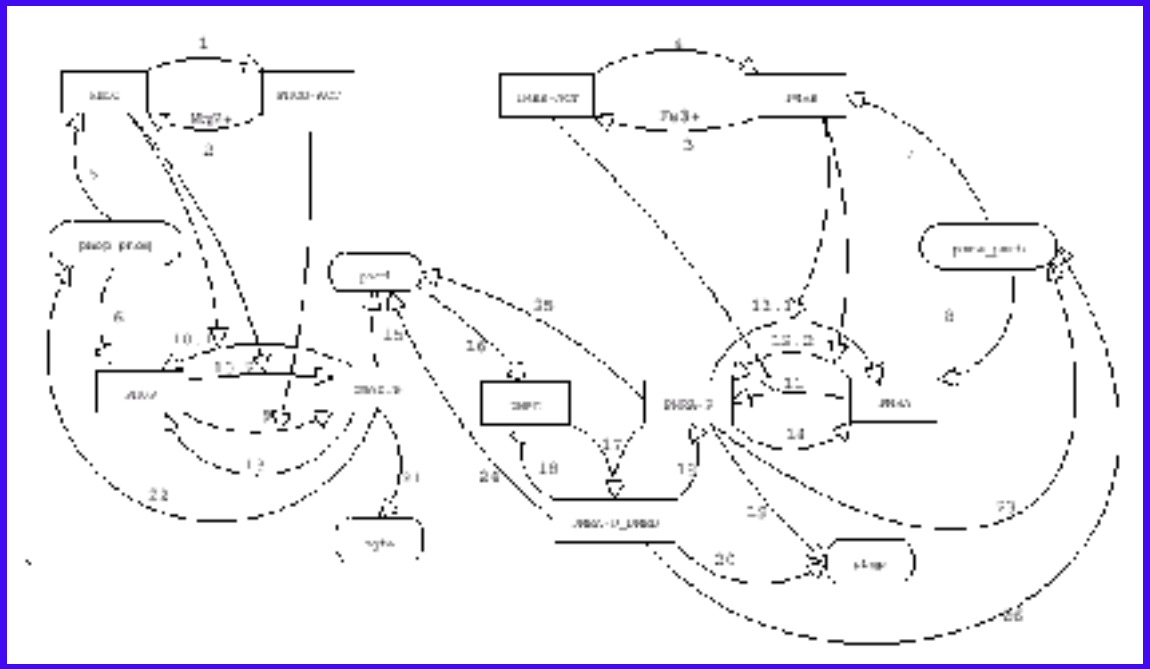
MATEMATHICAL INTERPRETATION OF BUILDING BLOCKS

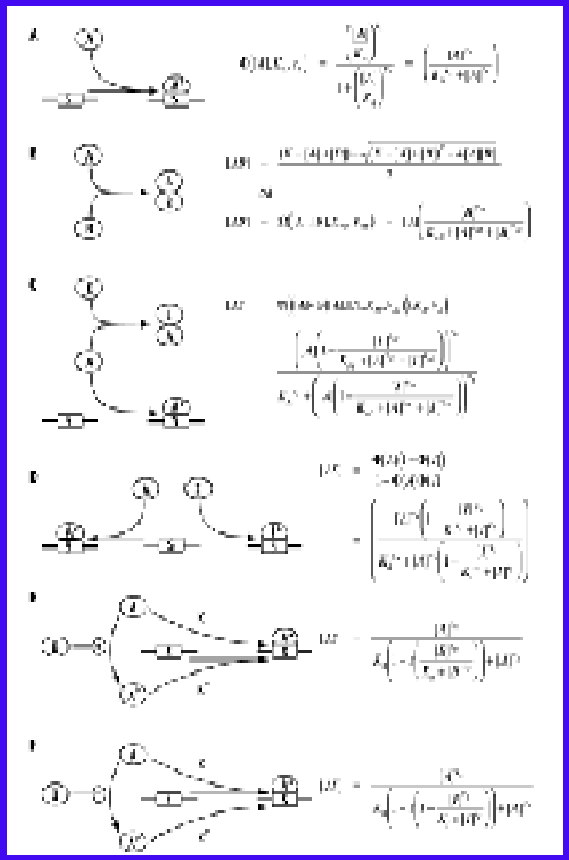


$$\frac{d[phop]}{dt} = \frac{T_0}{H_{phop}} \frac{[PHOP]^{\nu_{PHOP\_phop}}}{K_{PHOP\_phop}^{\nu_{PHOP\_phop}} + [PHOP]^{\nu_{PHOP\_phop}}}^{(1)} - \frac{T_0[phop]^{(4)}}{H_{phop}}$$

$$\frac{d[PHOP]}{dt} = \frac{T_0[phop]^{(2)}}{H_{PHOP}} - \frac{T_0[PHOP]^{(3)}}{H_{PHOP}}$$

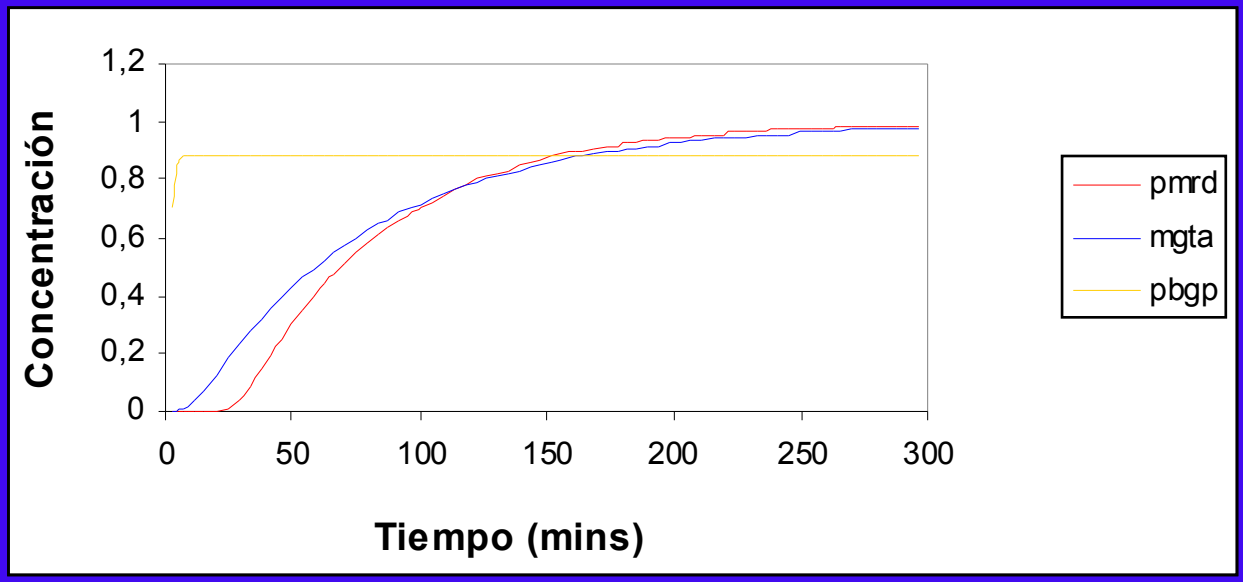
COMPOSING BUILDING BLOCKS INTO GENE NETWORKS



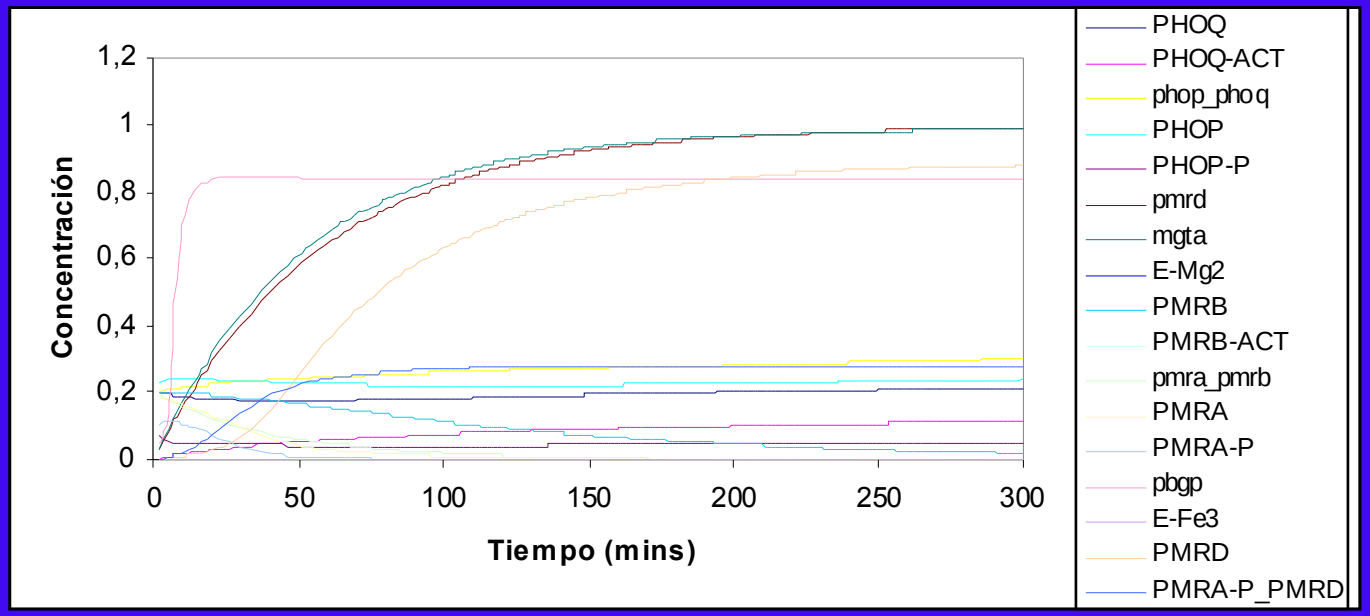


CONSTRAINT-BASED LEARNING ALGORITHM

	Entrada		Salida		
Funcionalidad	Mg2+	Fe3+	mgta	pmrd	pbgp
1	1	1	1	0	1
2	1	0	1	1	1
3	0	1	0	0	1
4	0	0	0	0	0

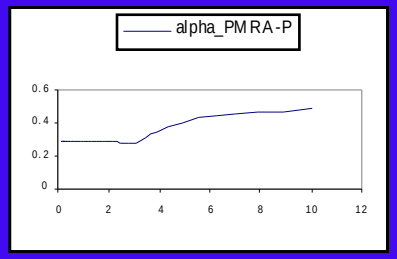


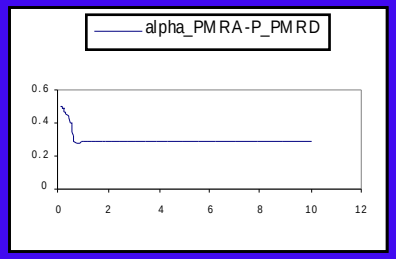
SIMULATIONS AND EVALUATION

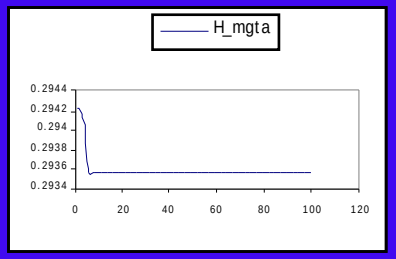


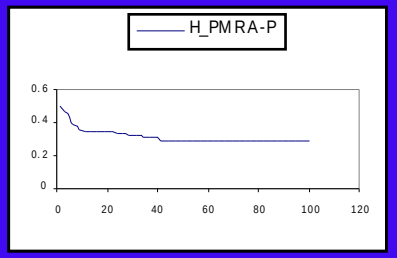
	Entrada		Salida					
Funcionalidad	Mg2+	Fe3+	mgta	pmrd	pbgp	# Parámetros	Proporción	Probabilidad
1	1	1	1	0	1	68	0,000186349	0,881357012
2	1	0	1	1	1	68	0,001092771	0,904584113
3	0	1	0	0	1	68	0,000735495	0,899332507
4	0	0	0	0	0	68	0,469708768	0,988949126

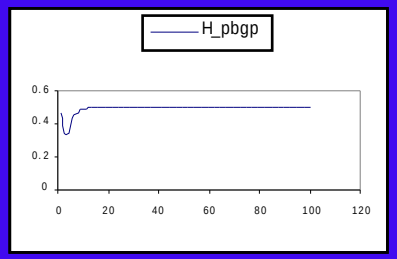
PARAMETER SENSITIVITY

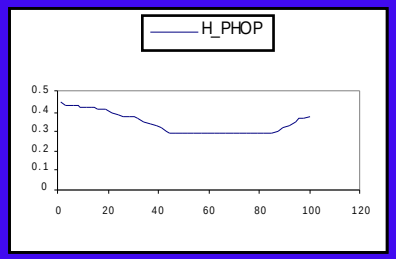


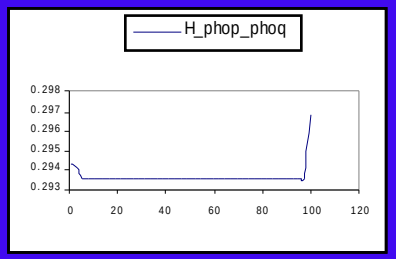


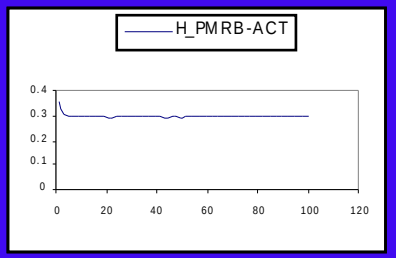


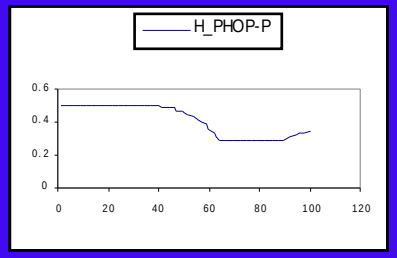


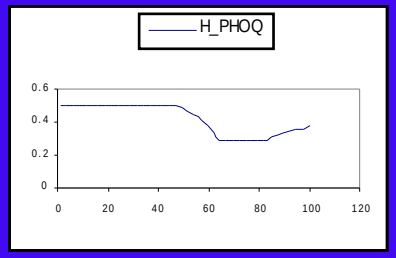


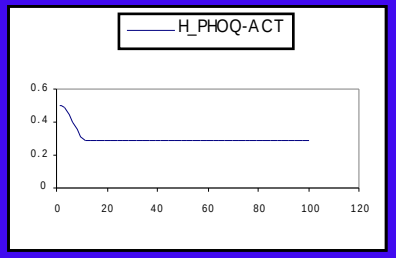


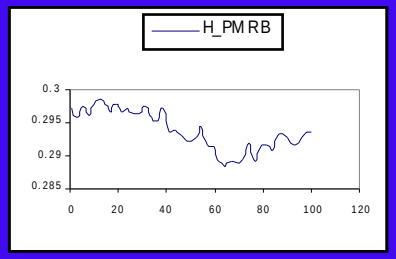


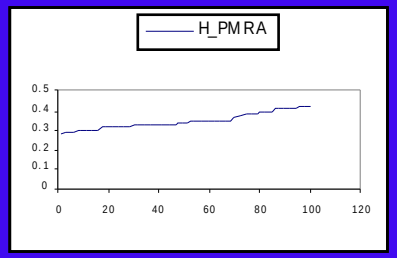


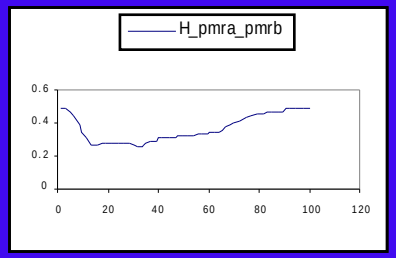


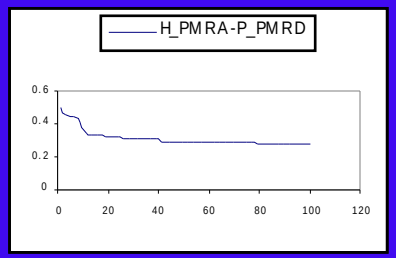






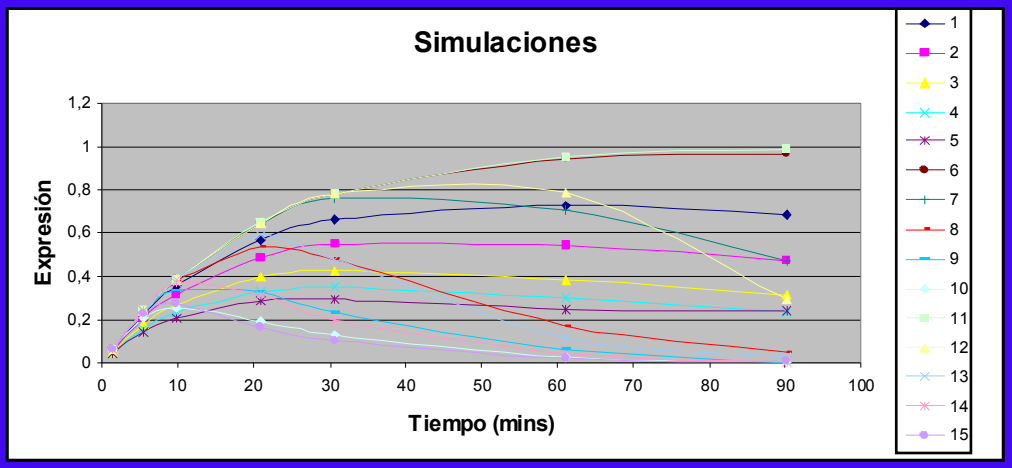


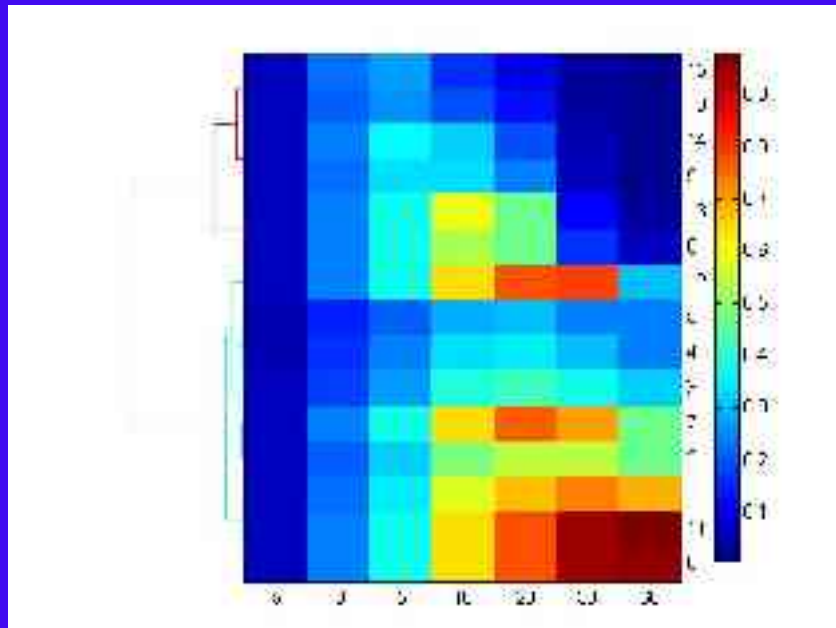




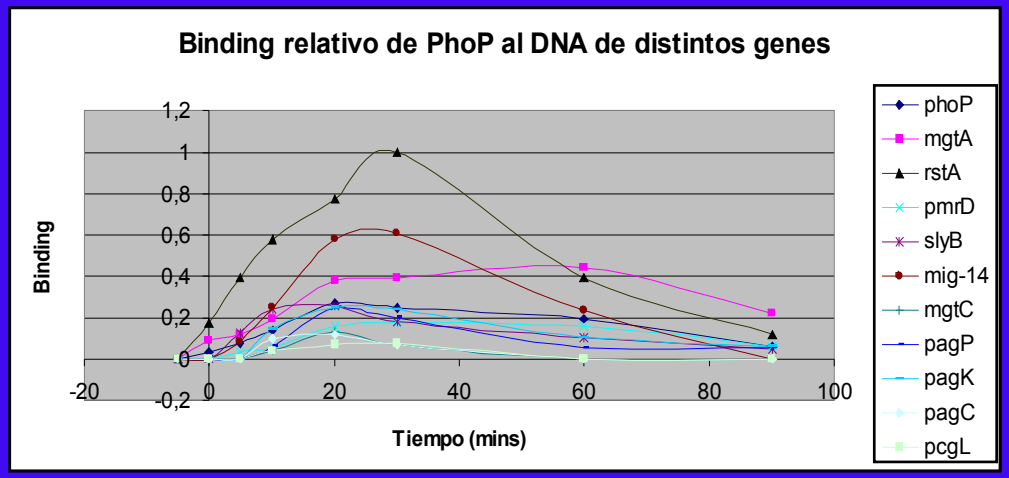
PREDICTIONS

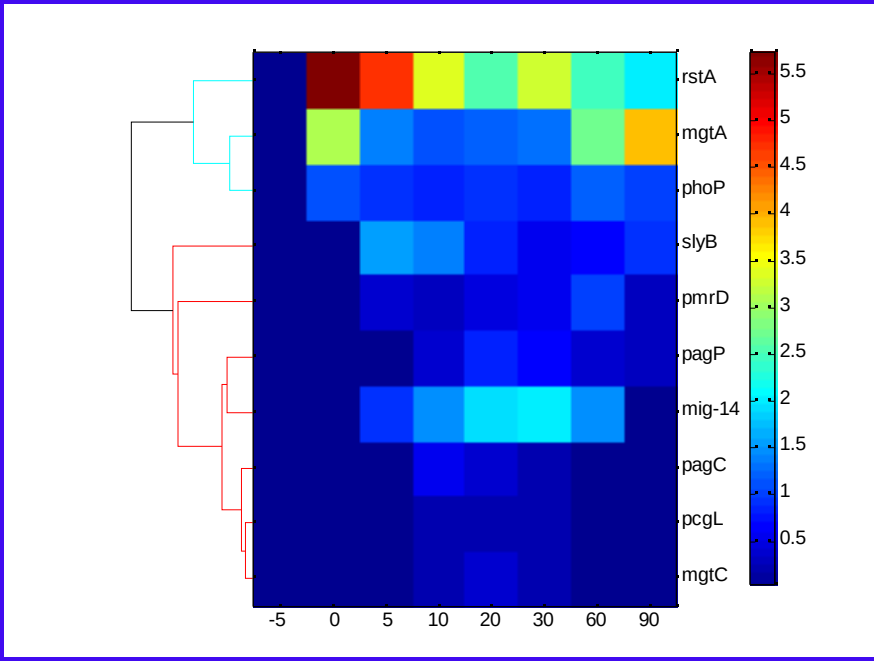
Parámetros	1	2	3	4	5	6	7	8
H_PHOP-P	10	10	10	10	10	10	10	10
H_mgta	20	20	20	20	20	20	20	20
nu_mgta	1	1	1	1	1	5	5	5
K_mgta	0,01	0,025	0,05	0,075	0,1	0,01	0,025	0,05
	Simulación							
Parámetros	9	10	11	12	13	14	15	
H_PHOP-P	10	10	10	10	10	10	10	
H_mgta	20	20	20	20	20	20	20	
nu_mgta	5	5	10	10	10	10	10	
K_mgta	0,075	0,1	0,01	0,025	0,05	0,075	0,1	





EXPERIMENTAL VALIDATION

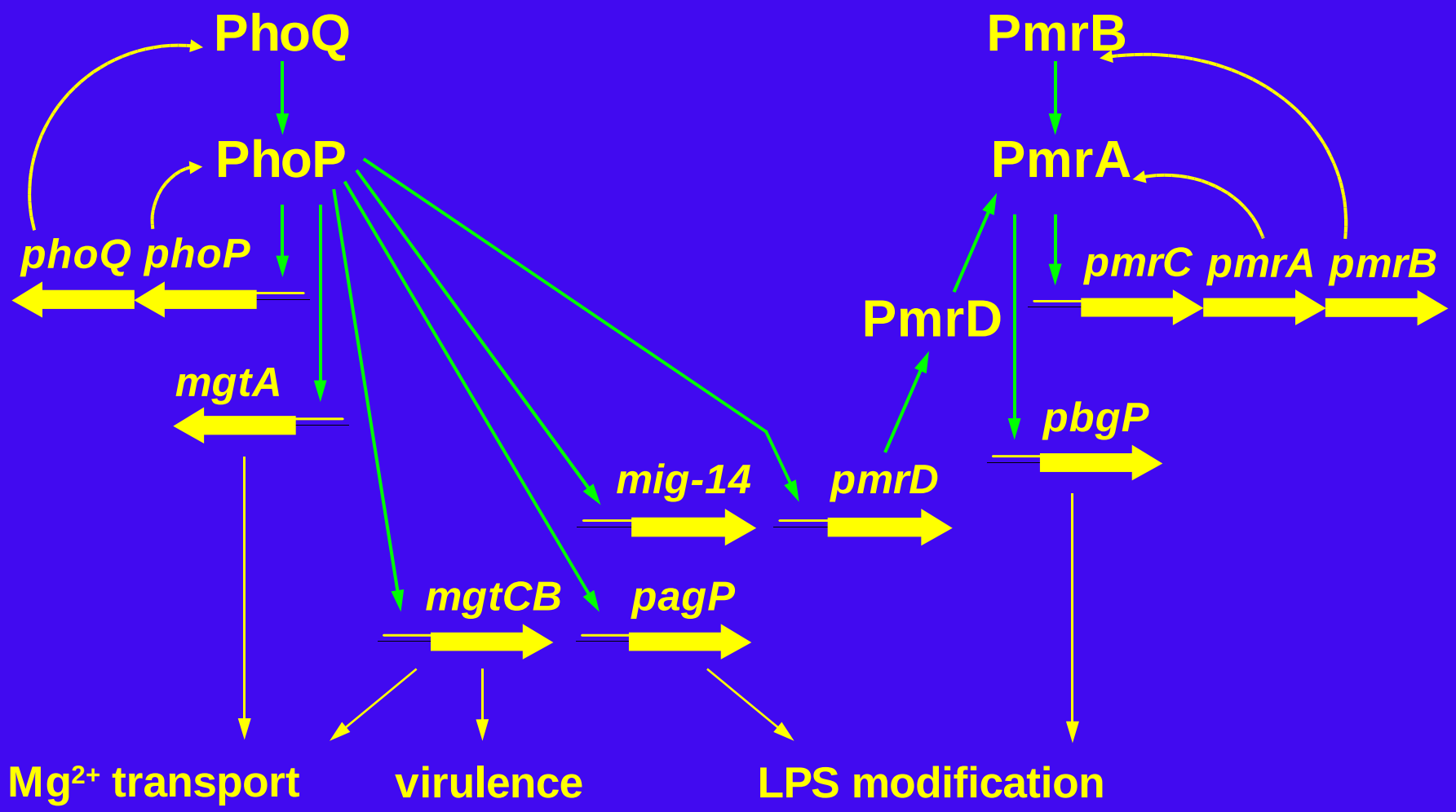




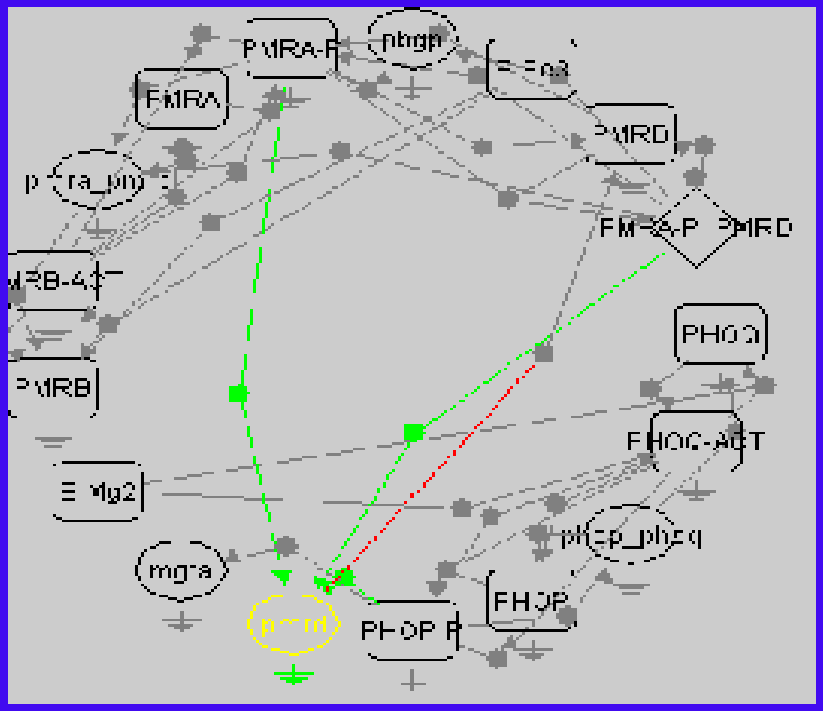
CORRELATION BETWEEN EXPERIMENTAL VALIDATION AND PREDICTIONS

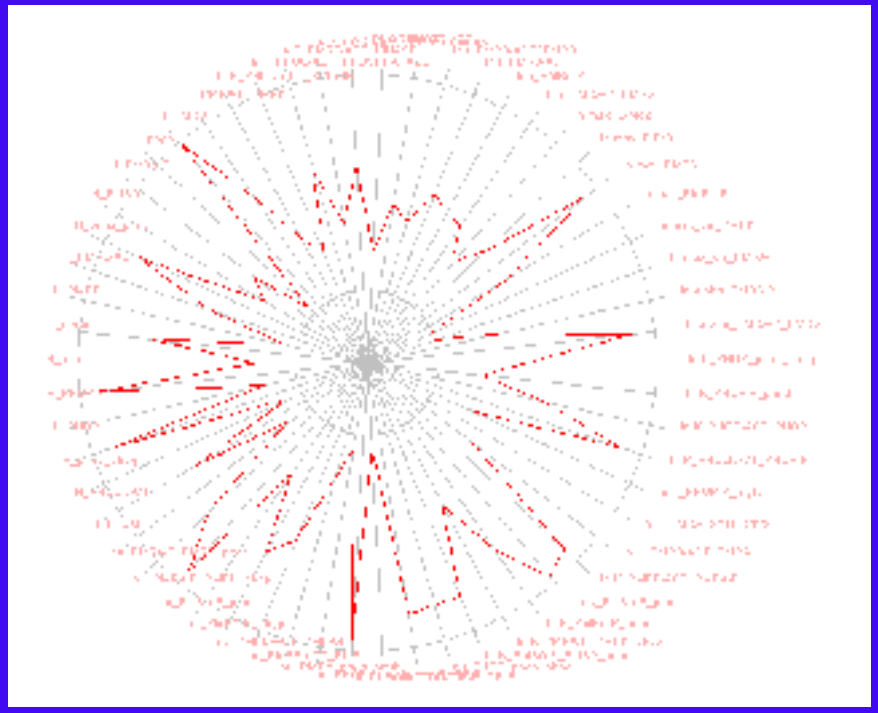
Correlación de Pearson											
Simulación	phoP	mgtA	rstA	pmrD	slyB	mig-14	mgtC	pagP	pagK	pagC	pcgL
1	0,619287092	0,857949781	0,339818737	0,829000283	0,264923353	0,482460587	0,277169074	0,556744029	0,5991426	0,180180349	0,320979193
2	0,745769371	0,906526323	0,502007318	0,901528946	0,424242626	0,628335893	0,419452365	0,671094889	0,722414443	0,332324322	0,47099036
3	0,843852314	0,915138454	0,648375986	0,936323412	0,581156672	0,751659807	0,552810422	0,762873798	0,822053179	0,482872977	0,608678758
4	0,886153114	0,898741294	0,725794354	0,936364865	0,671873682	0,811663401	0,625347639	0,803557401	0,867101681	0,569665865	0,682078812
5	0,813971572	0,842706965	0,657161056	0,882303929	0,641101063	0,746960693	0,600671959	0,782555306	0,834827405	0,54888883	0,649970279
6	0,484578386	0,784076325	0,182657971	0,735239947	0,119355869	0,334879341	0,142821565	0,436909899	0,470369998	0,042201854	0,177108088
7	0,858977561	0,961379585	0,648559205	0,971032198	0,499422627	0,75927577	0,509215514	0,747911953	0,80498073	0,413593112	0,583025581
8	0,843524871	0,499762374	**0,946114321**	0,626705145	**0,922203925**	**0,92202348**	0,895723654	**0,835517001**	**0,868634095**	0,900230693	**0,932480899**
9	0,550703753	0,117924047	0,749545426	0,247775955	0,917814195	0,650777545	0,732942357	0,530287422	0,591071935	0,850412566	0,733320591
10	0,286592176	-0,150739857	0,545795459	-0,027344699	0,799582889	0,394582616	0,512804486	0,253380729	0,325408172	0,687623039	0,509318988
11	0,468974213	0,772242027	0,167076499	0,722511526	0,107396064	0,319675582	0,132281945	0,426383442	0,458211028	0,031994059	0,165225816
12	**0,913010303**	**0,975698548**	0,713560888	**0,981413357**	0,533762236	0,804706746	0,509096392	0,715888847	0,79386547	0,42468576	0,586933512
13	0,801054535	0,429836349	0,915624862	0,563254584	0,913774241	0,894472406	**0,919408323**	0,83067834	0,842771907	**0,917072159**	0,930356106
14	0,425722417	-0,007911581	0,639678643	0,113528727	0,886942645	0,524659984	0,647662009	0,409550397	0,485339495	0,810487256	0,636511775
15	0,122722329	-0,291314081	0,403220158	-0,17977406	0,698132005	0,225771766	0,349213006	0,071990352	0,155319882	0,556525278	0,349257421
Max Val	0,913010303	0,975698548	0,946114321	0,981413357	0,922203925	0,92202348	0,919408323	0,835517001	0,868634095	0,917072159	0,932480899
Simulación	12	12	8	12	8	8	13	8	8	13	8
2 Mejor	4	7	13	7	9	13	8	13	4	8	13

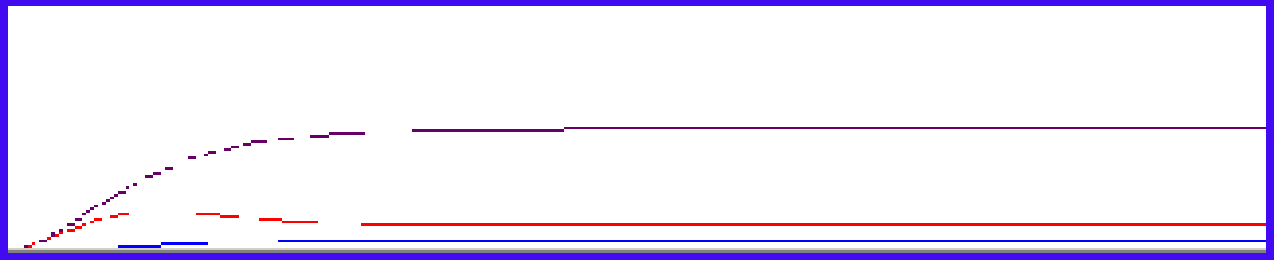
SEQUENTIAL ACTIVATION OF THE PHOP REGULON



INGENEUE ENVIRONMENT







Microarray Integrated Analysis (MIA)

The Raw Material Available for the Analysis Consist of …

- Gene expression (GeneChips) derived from longitudinal blood expression profiles of human volunteers treated with intravenous endotoxin compared to placebo
- 8 patients, four treated, patients 1-4, and four control, patients 5-8
- From each patient there are 6 samples taken at different time points: at hour 0, hour 2, 4, 6, 9, 24. For patient 6, hours 4 and 6 are missing

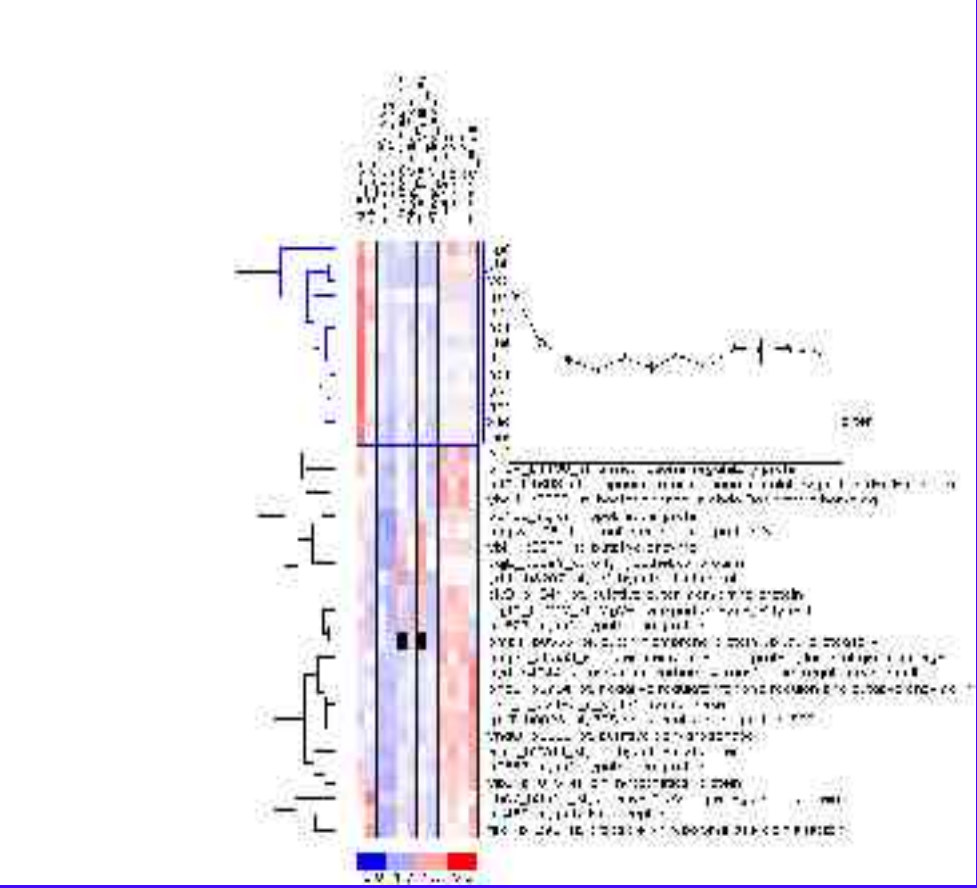
Our Ultimate Goal is to …

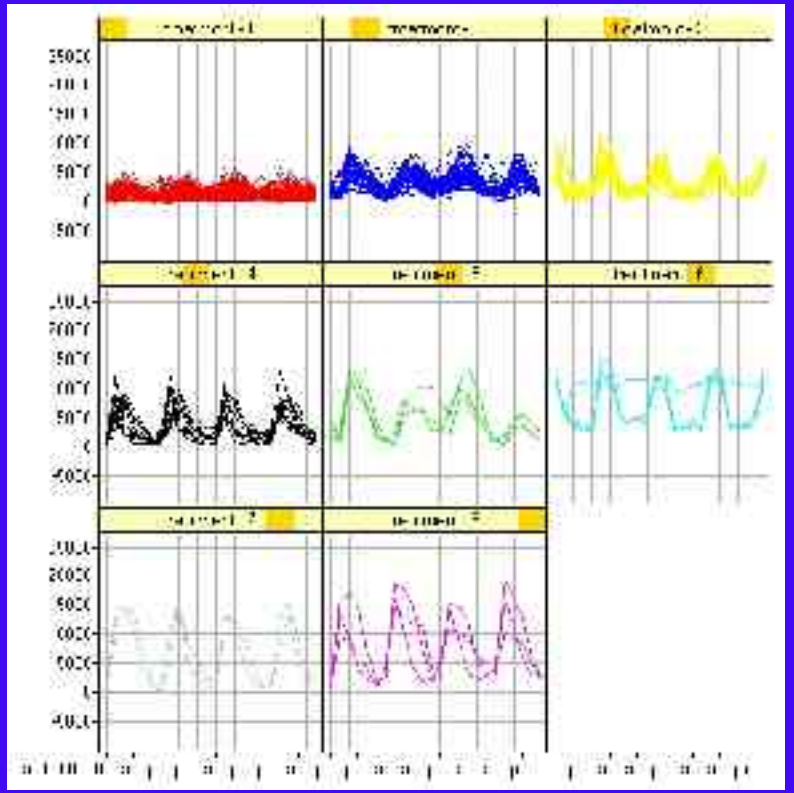
… identify molecular pathways that provide insight into the host response over time to systemic inflammatory insults, as part of a Large-scale Collaborative Research Project sponsored by the National Institute of General Medical Sciences (www.gluegrant.org).

… by Building Genetic Networks Using Reverse Engineering Techniques

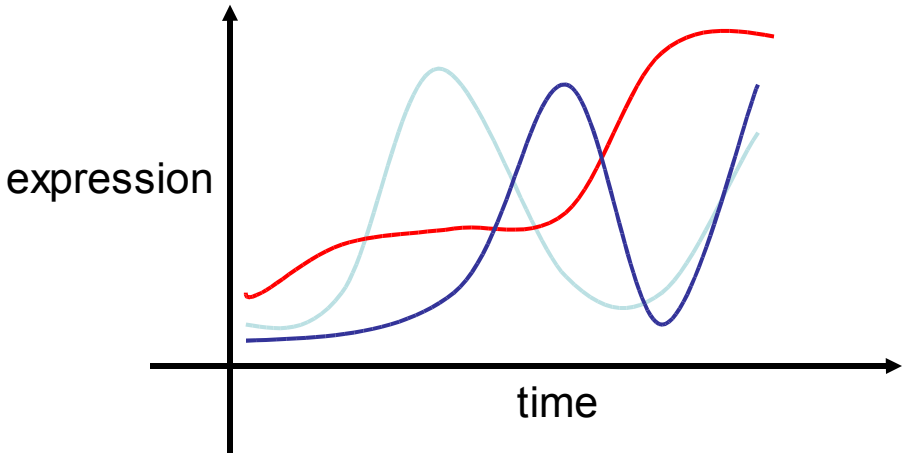
- Hierarchical approach to model the dynamics of genetic networks by doing reverse engineering
- Need of gene expression preprocessing by clustering genes into profiles
- Difficulties of typical statistical methods in detecting significant changes between treated and control gene profiles
- Revisit the reverse engineering approach by using data mining tools

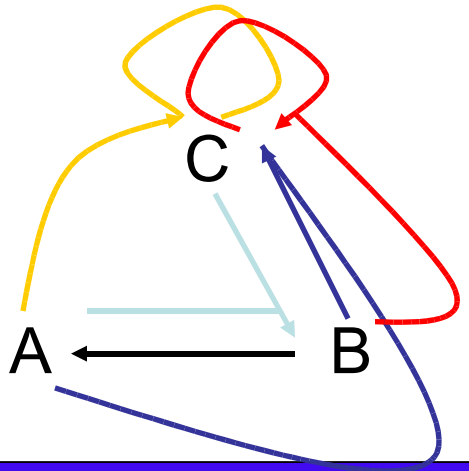
Gene Expression Patterns Detected by Clustering Methods





First Approach to Obtain Dynamic Patterns: Boolean Networks



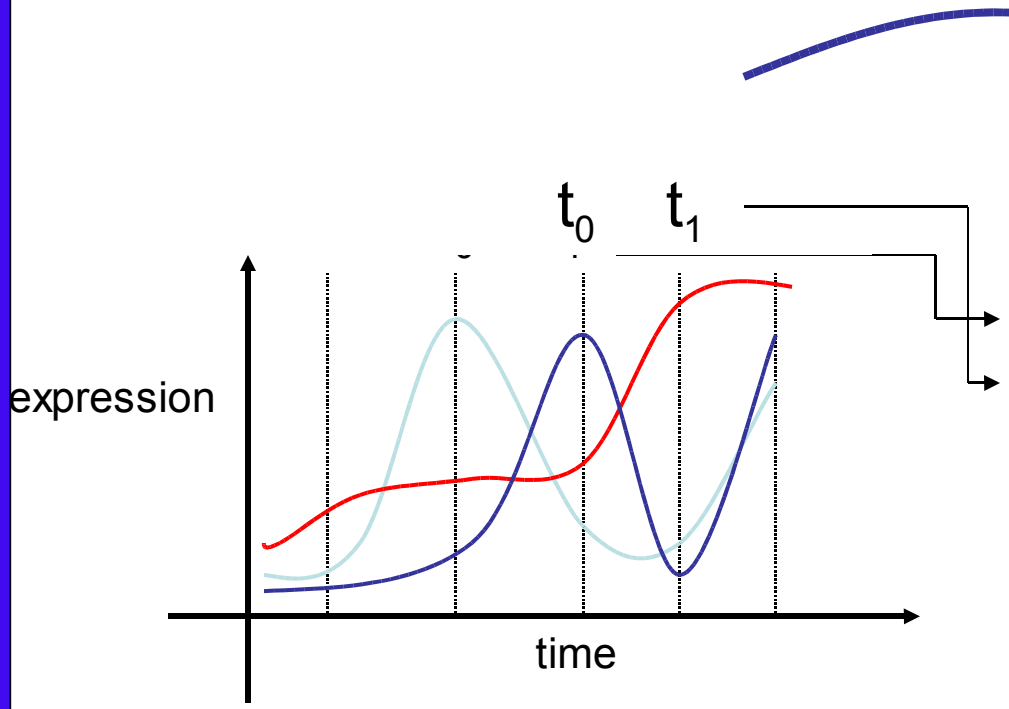


A ← B

B ← A or C

C ← (A and B) or (B and C) (A and C)

First Approach to Obtain Dynamic Patterns: Boolean Networks



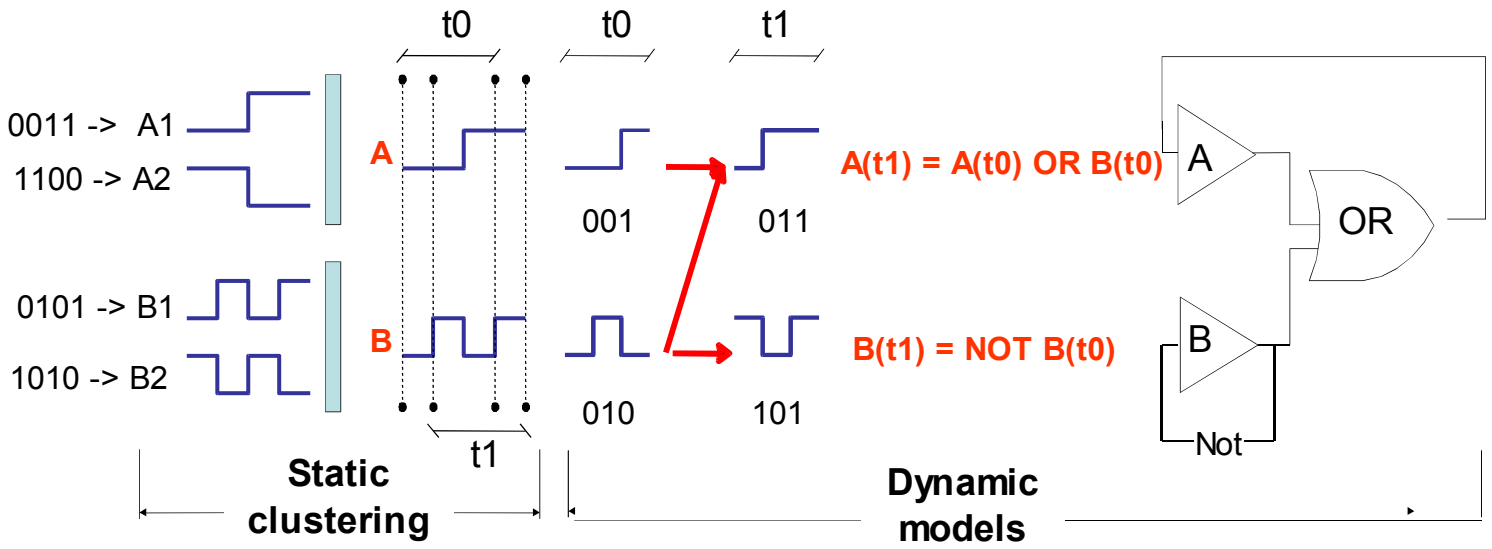
Time t			*Time t+1*		
A	B	C	A	B	C
0	0	0	0	0	1
0	0	1	0	1	0
0	1	0	1	0	0
1	0	0	1	1	1
1	0	0	0	1	0
1	0	1	0	1	1
1	1	0	1	1	1
1	1	1	1	1	1

A ← B

B ← A or C

C ← (A and B) or (B and C) (A and C)

Static Vs. Dynamic Clustering



Not B(t0) = B(t1)

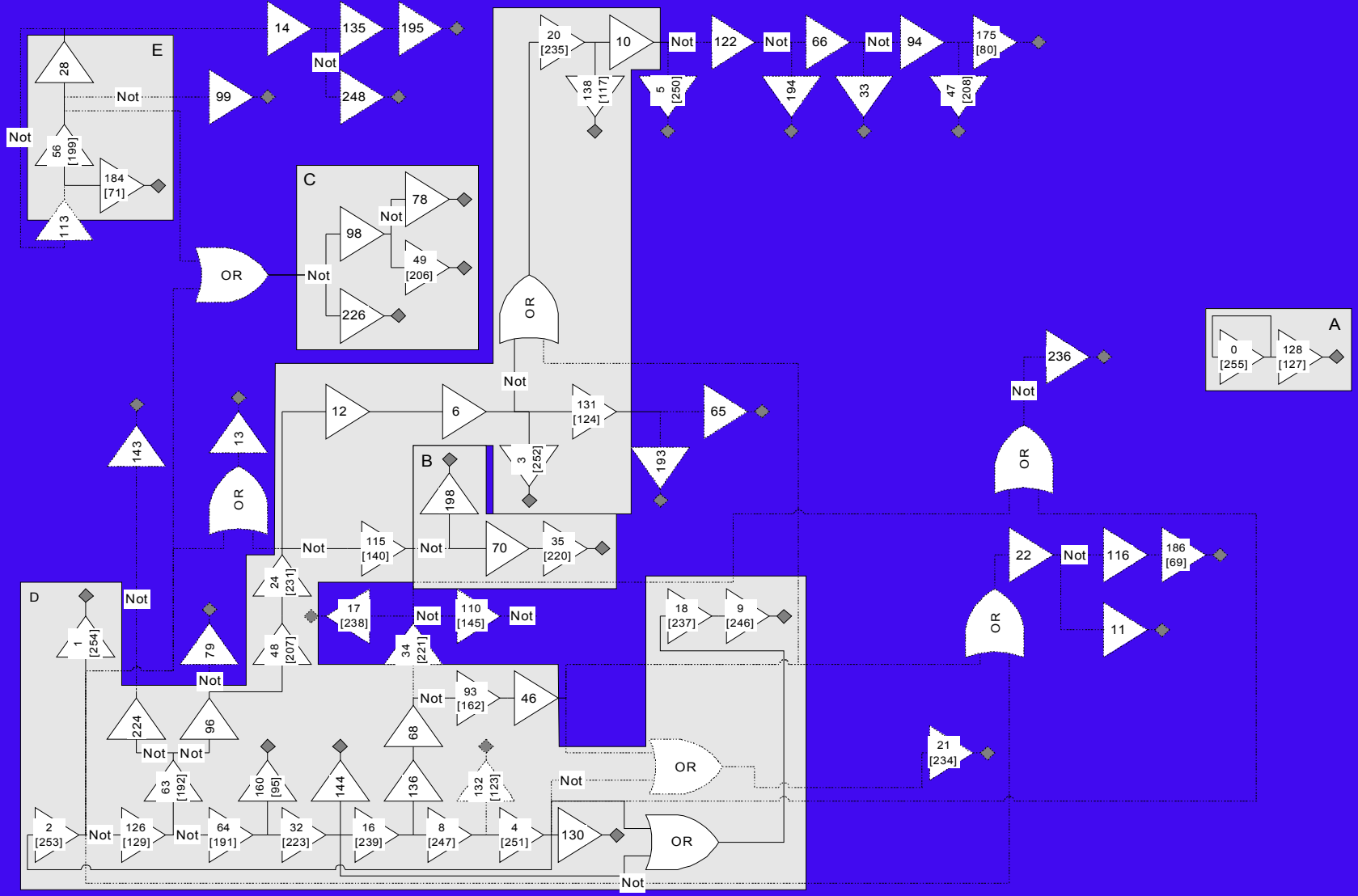
0	1
1	0
0	1

A(t0) Or B(t0) = A(t1)

0	0	0
0	1	1
1	0	1

The relationship between B(t0) and B(t1) can be captured by grouping B1 and B2 as a negative correlation. However, relationships among A(t0), B(t0) and A(t1) could not be determined in the same way because they are derived from a temporal behavior.

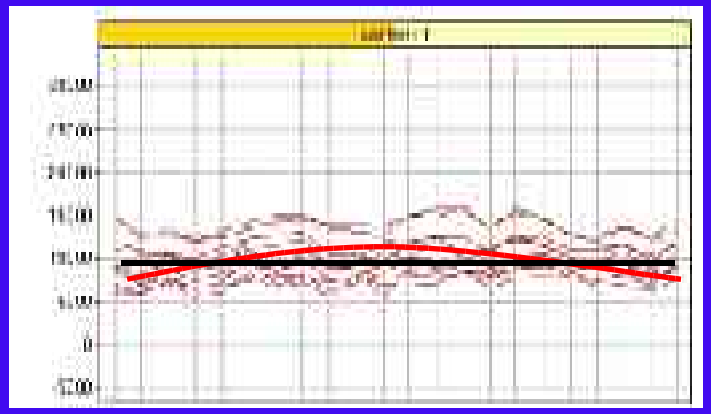
Dynamic Patterns Learned by Doing Reverse Engineering

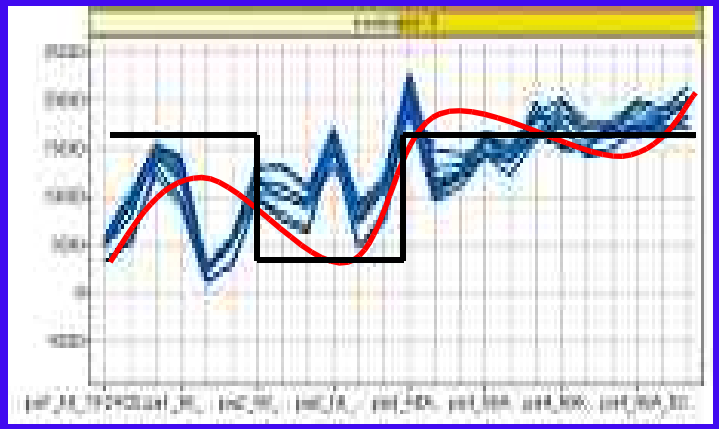


A More General Approach: Hierarchical Modeling

- Preprocess gene expression by clustering genes into prototypes
- Build Boolean circuits based on prototypes
- Recognize differential patterns between experiments and control
- Interpret differences
- Select interesting genes
- Go deep into continuous modeling

Boolean Modeling: Pattern Preprocessing

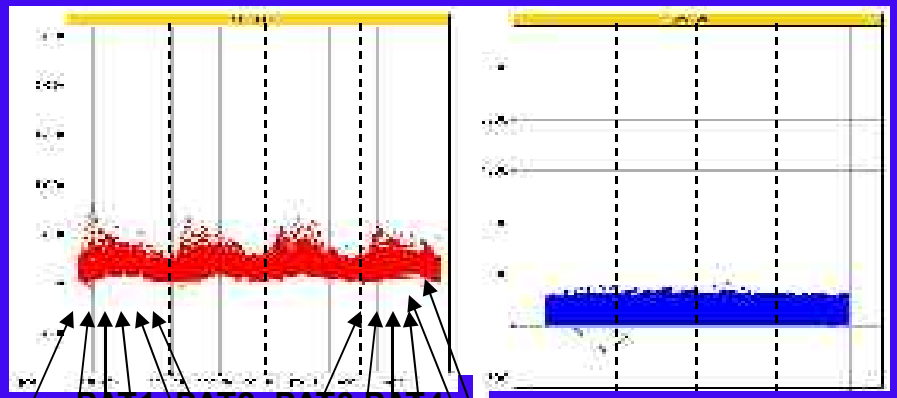




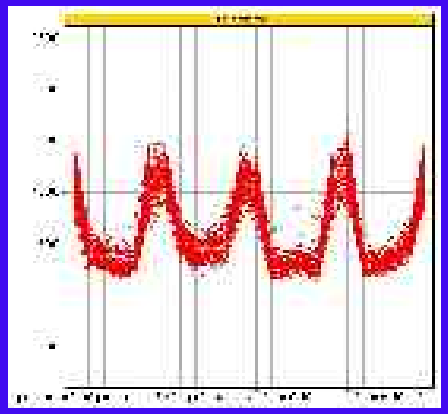
OBJECTIVE: Find Genes with Different Treatment-Control Profiles, e.g.

TREATMENT

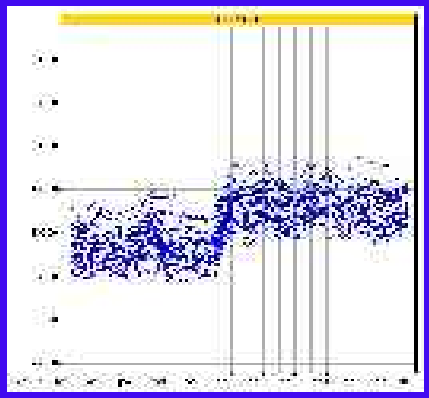
CONTROL



TREATMENT

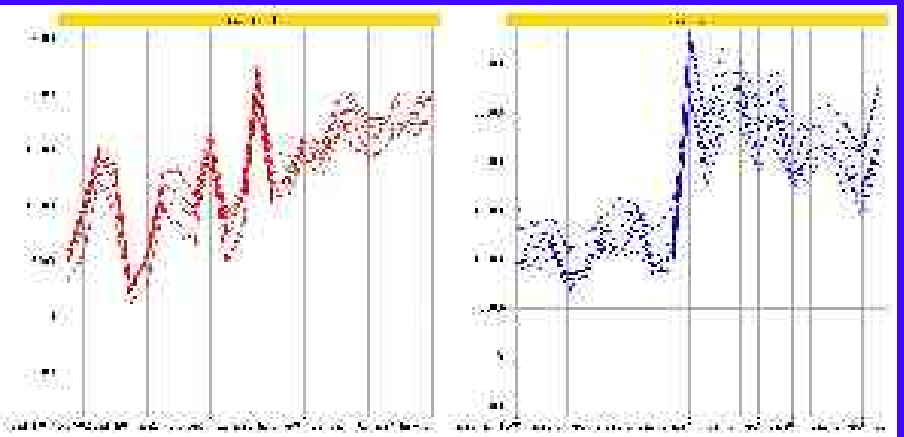


TREATMENT

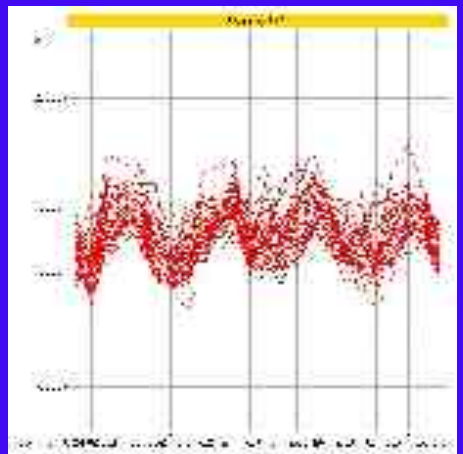


TREATMENT

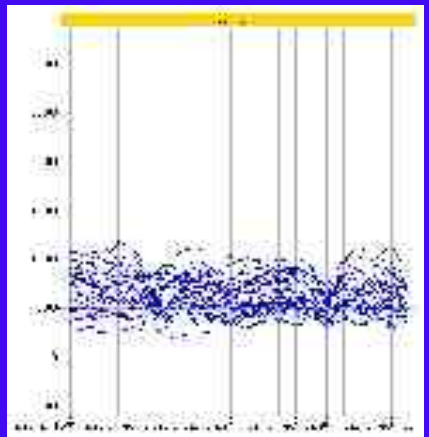
CONTROL



TREATMENT

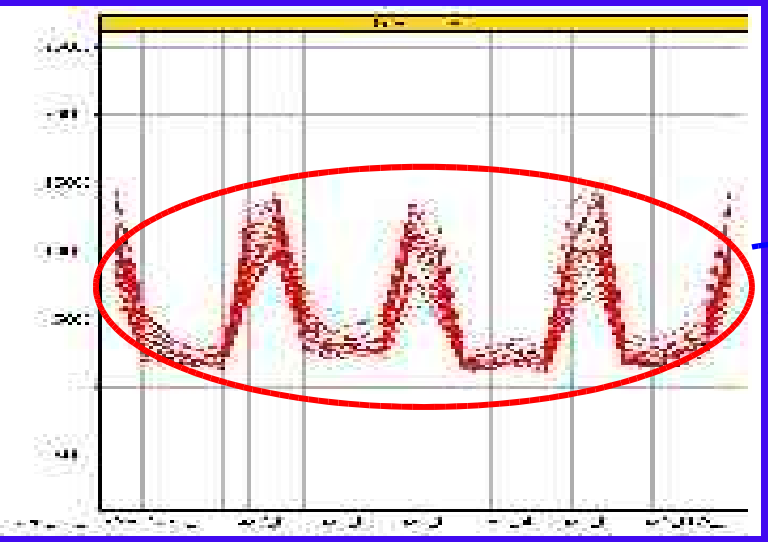


CONTROL

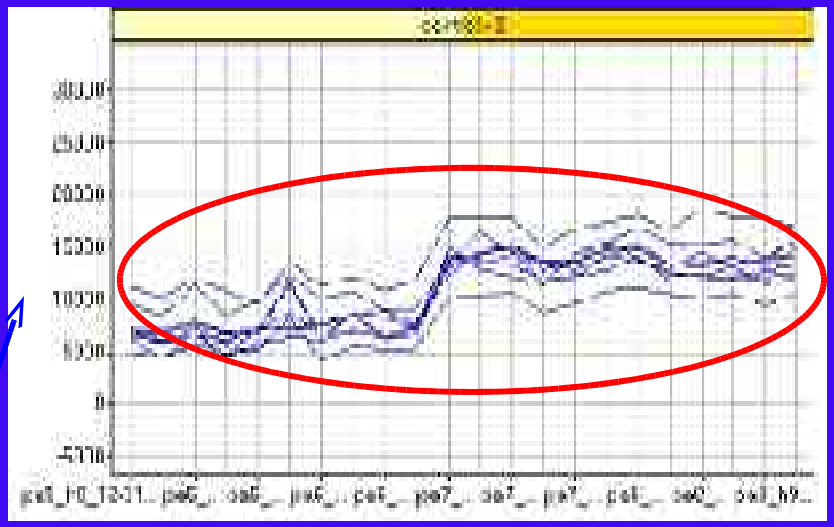


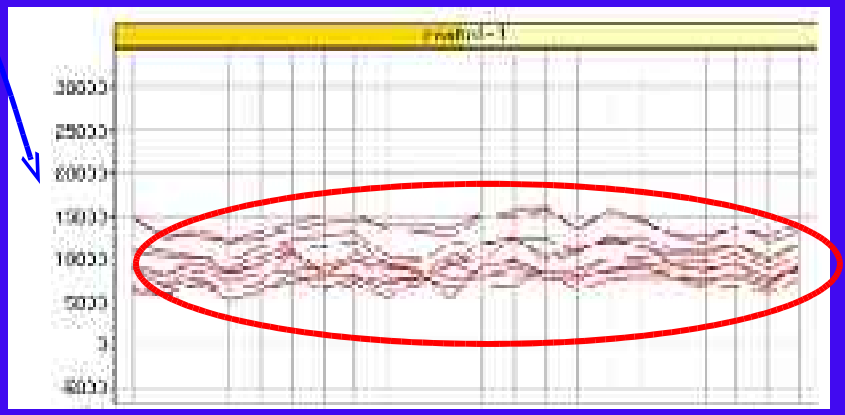
Representation Issues…

TREATMENT



CONTROL





A cluster profile from treatment might result in several different profiles in control and vice versa.

Variability in Microarray Data...

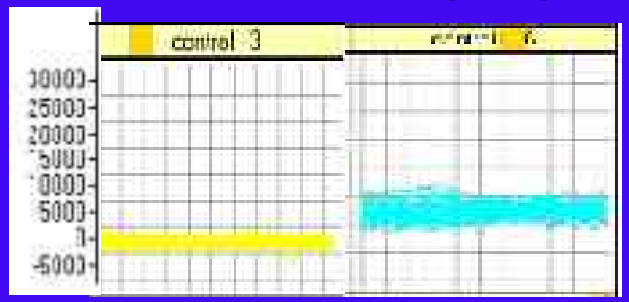
- **TOTAL VARIABILITY**: Biological Variability (main interest) + Systematic Variability (technical) + Random Variability

- The statistical methods for microarray data analysis try to extract the biological variability and set aside the systematic and random variability.

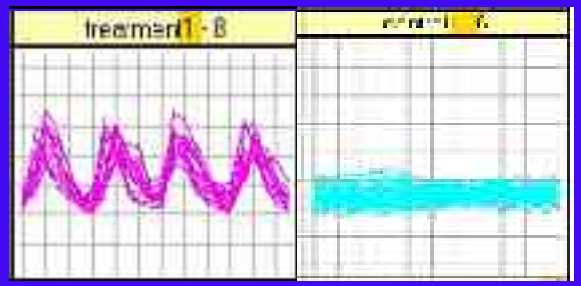
Steps in the Statistical Analysis

- **SCALING:** conversion of fluorescence values to numeric values in an established rank. We use Li-Wong scaling model in the whole analysis process.
- **NORMALIZATION:** The process of removing systematic variability (detection of outliers).
- **FILTERING:** preliminary elimination of genes that show no expression change at all between treatment and control.
- **DIFFERENTIAL ANALYSIS:** Select a set of genes that are differentially expressed from treatment to control (show biological variability). This analysis will be performed over treatment / control and also over **time**. Since we work with data over time series, it is interesting to notice the behaviour of a gene over time. Accordingly we will analyse the data over both treatment and time.

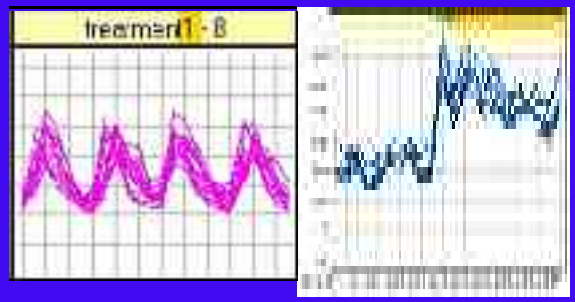
TREATMENT CONTROL



TREATMENT CONTROL



TREATMENT CONTROL



Summarizing...

METHOD	SCALING	NORMALIZING	FILTERING	DIFFERENTIAL ANALYSIS
LI-WONG	Li-Wong	Li-Wong	Yes	t-test
SAM	Li-Wong	Li-Wong	Yes/No	Multiple test, permutation false discovery rate
ANOVA	Li-Wong	Anova	No	Anova
CLUSTERING	Li-Wong	Li-Wong	Yes	Profile Differentiation

Number of Genes Selected by each Method

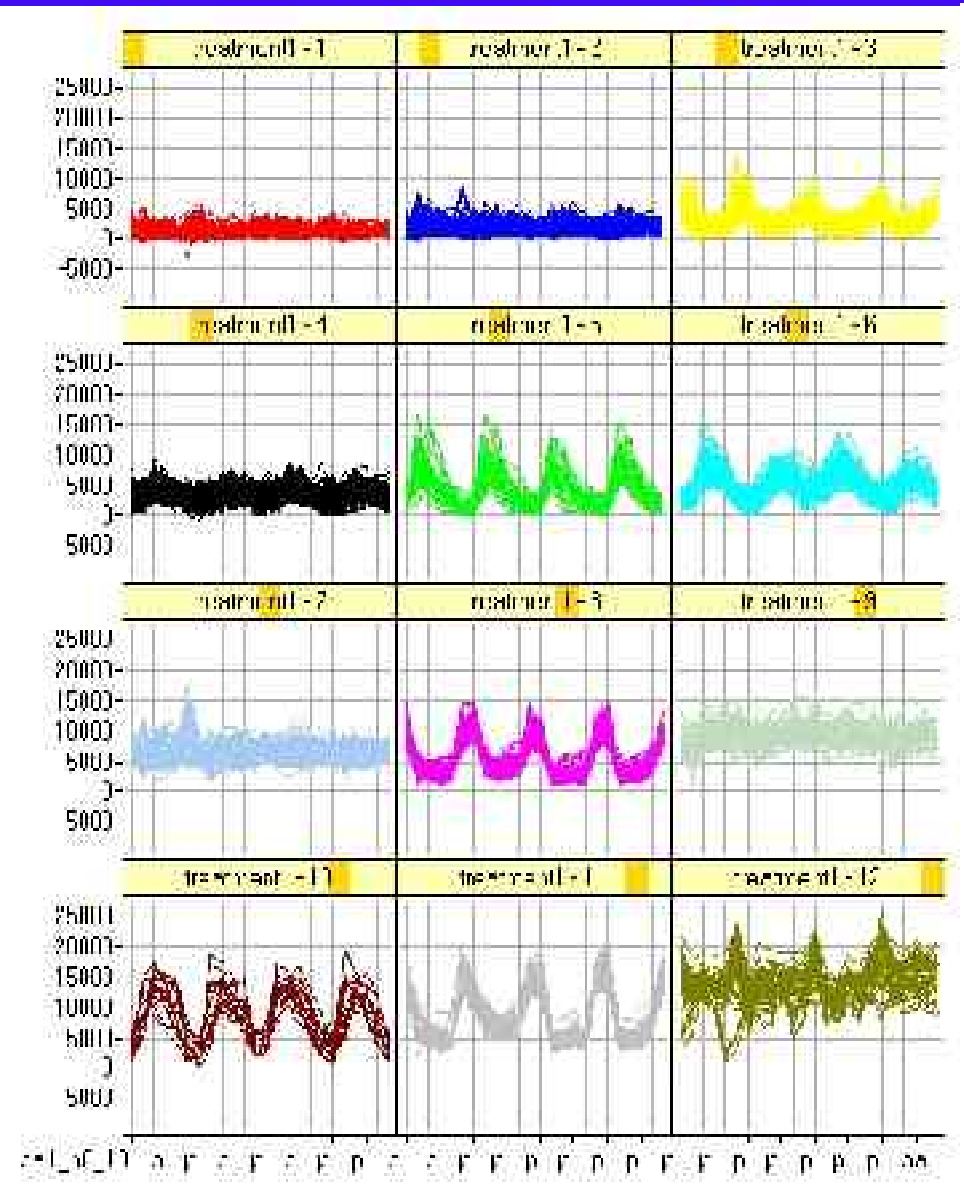
METHOD	Number of Genes selected
Li-Wong filter	1463
Li-Wong t-test	875
Li-Wong t-test with time-blocks	943
SAM over all genes FDR 0.798%	2764
SAM over Li-Wong filter default values FDR 0,608%	463
SAM over a broad Li-Wong filter FDR 1.80%	2650
ANOVA over treatment (p = 0.05)	13151, 10256 eliminated by clustering, 2895 left
ANOVA over time (p = 0.05)	4588, 3362 eliminated by clustering, 1226 left
ANOVA over time / treatment (p = 0.05)	6070, 3355 elimlnated by clustering, 2715 left

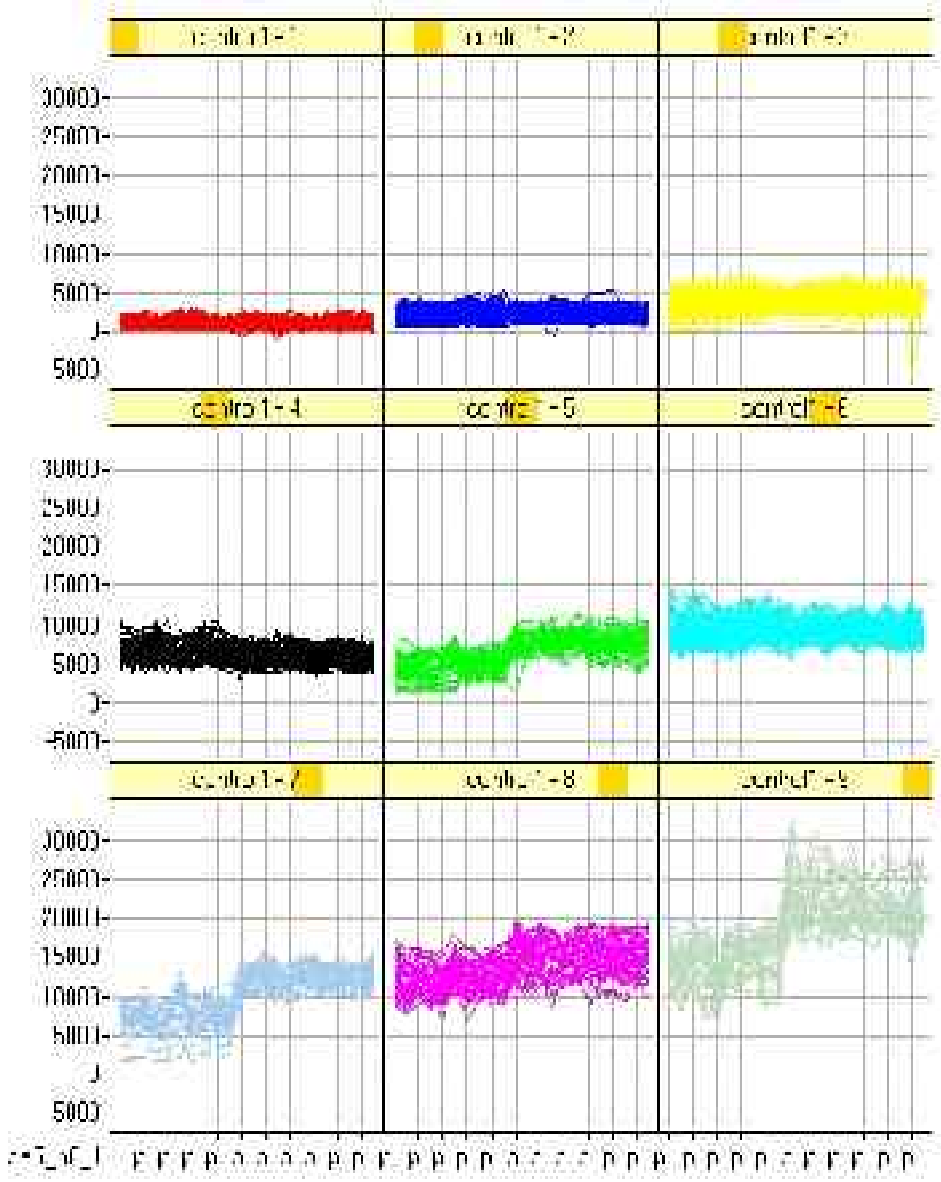
table 1

Results: ANOVA over Treatment, 2895 Genes

TREATMENT

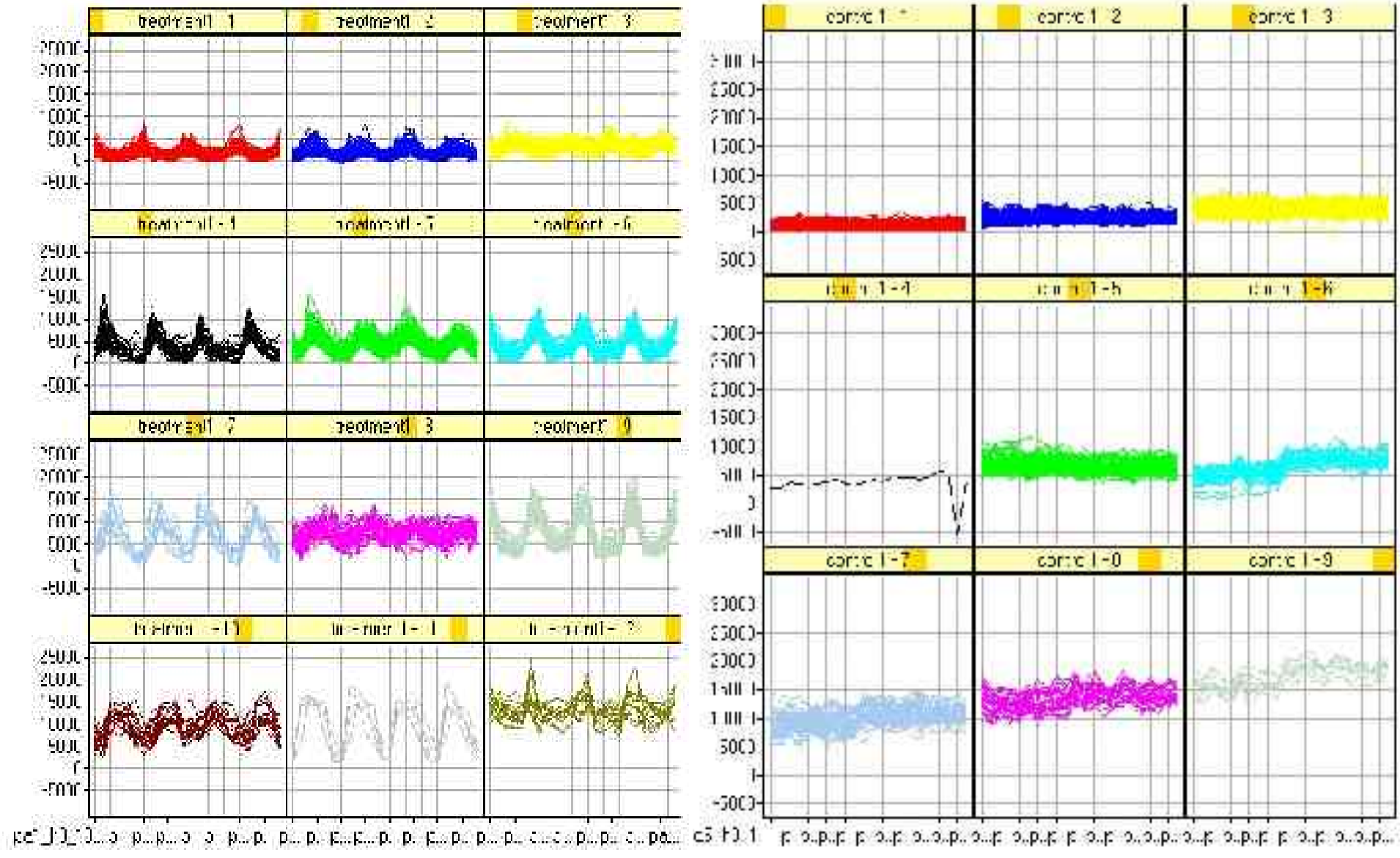
CONTROL





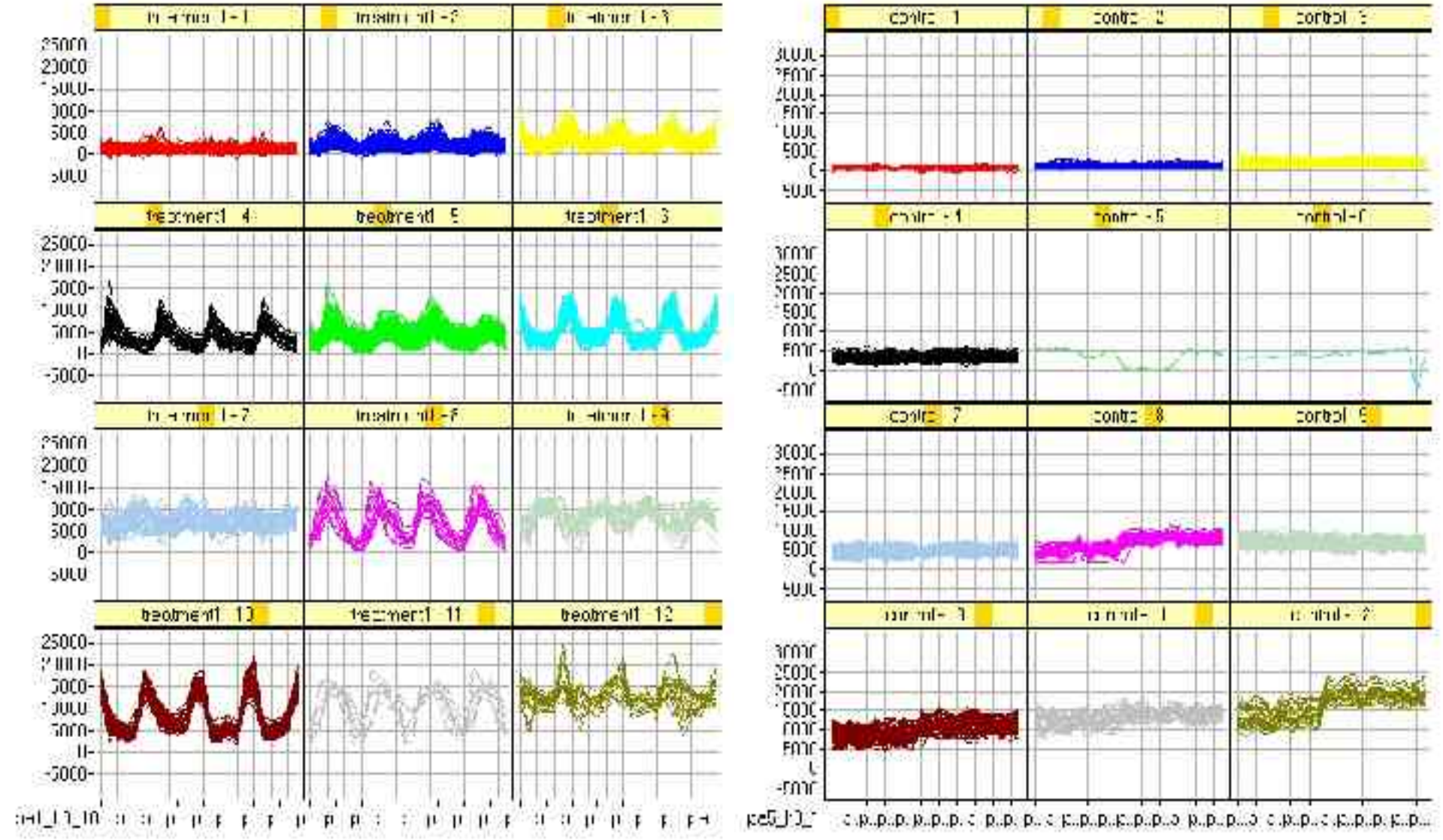
Results: ANOVA over Time, 1226 Genes

TREATMENT **CONTROL**



Results: ANOVA over Time/Treatment, 2715 Genes

TREATMENT **CONTROL**

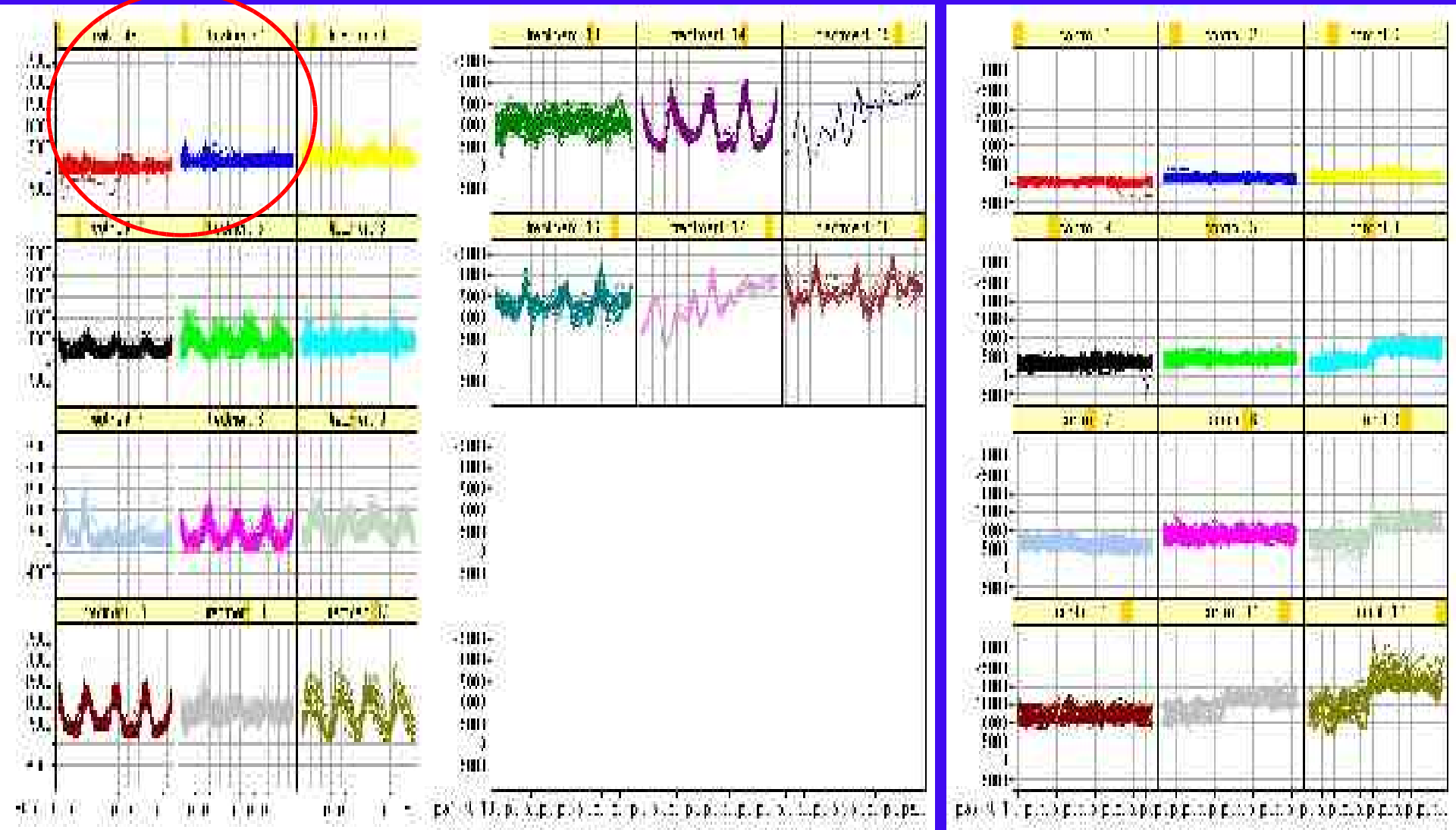


Results: 7217 Genes Selected as the Union set of All Previous Methods

TREATMENT

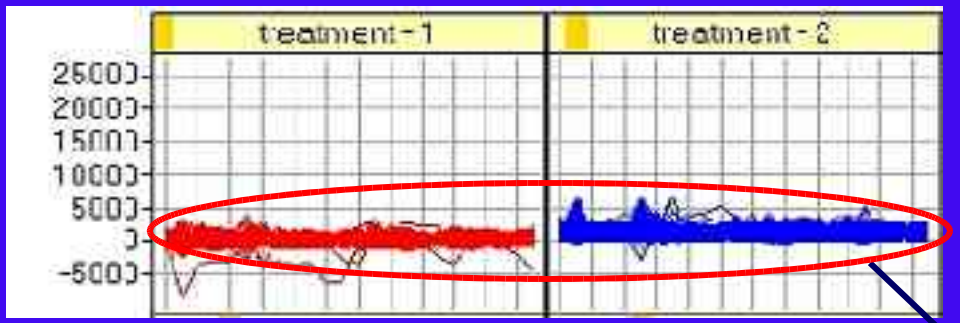
CONTROL

CONTROL



Results: Clustering

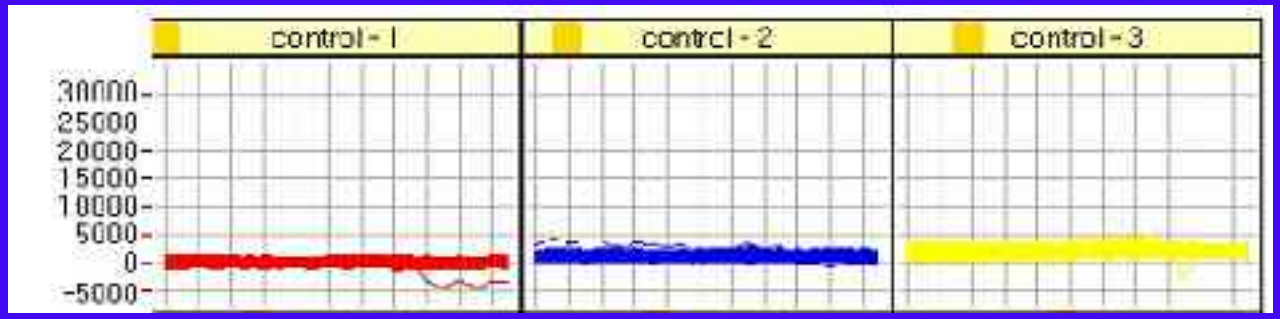
TREATMENT



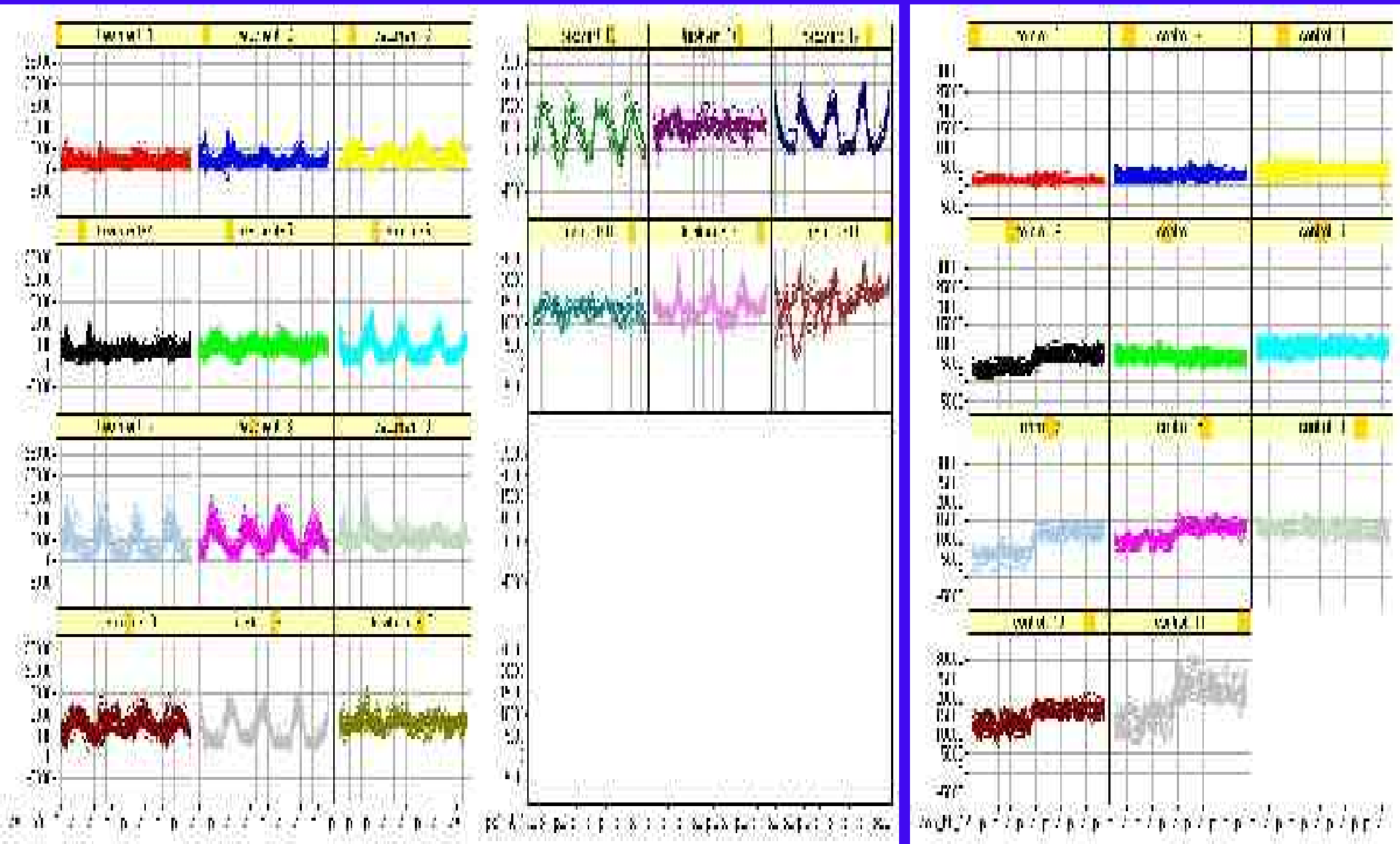
From the 7217 genes 5367 do not show an explicit differential profile expressed from control to treatment. Clustering has been applied and these clusters are not further studied, at least at this point.

5367 Genes

CONTROL



Summarizing Results, 1850 Genes Selected as Differentially Expressed from All Methods



Method Contribution to the Final Gene Selection

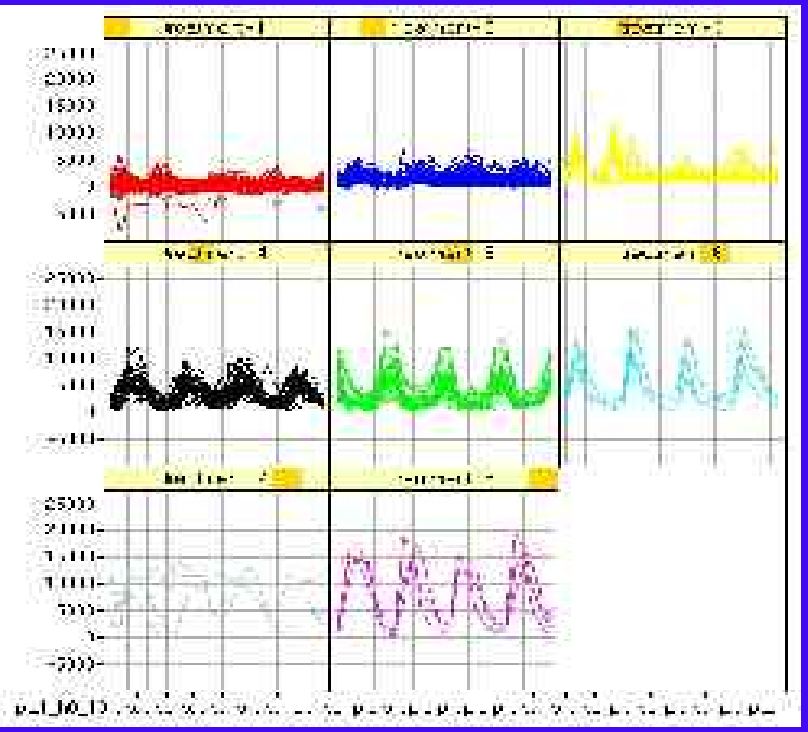
METHOD	Number of Genes selected	Percentage over the initial amount of genes of each method
LI-Wong + filter	366 (19,78 % of 1850)	25,01%
Li-Wong t-test	298 (16,10% ")	34,05%
Li-Wong t-test n with time-blocks	322 (17,40% ")	34,14%
SAM over all genes FDR 0,798%	366 (19,78% ")	13,24%
SAM over Li-Wong filter default values FDR 0,608%	159 (8,60% ")	34,34%
SAM over a broad Li-Wong filter FDR 1,80%	392 (21,19% ")	14,79%
ANOVA over treatment (p = 0.05)	1509 (81,57% ")	11,47%
ANOVA over time (p = 0.05)	1005 (54,32% ")	21,90%
ANOVA over time / treatment (p = 0.05)	1195 (64,60 ")	19,68%

table 2

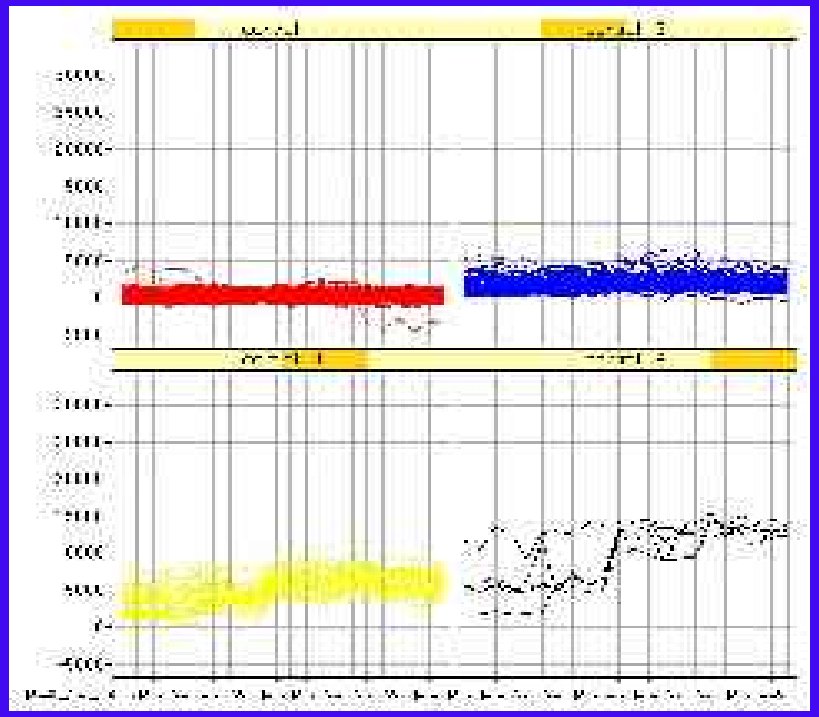
The percentage shown between parenthesis is the proportion of genes from the original method selection prior to applying any clustering (table 1) that stay in this final selection.

Genes Found by Li-Wong Filtered + t- test Comparisons and not by SAM

TREATMENT



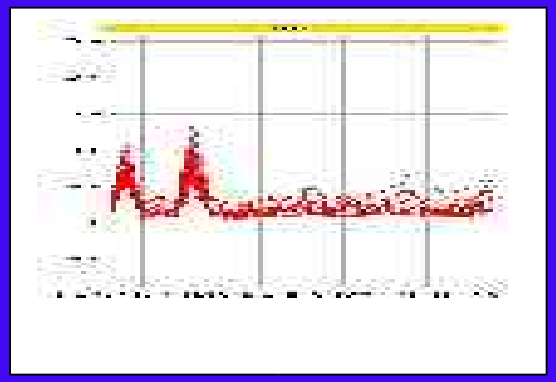
CONTROL



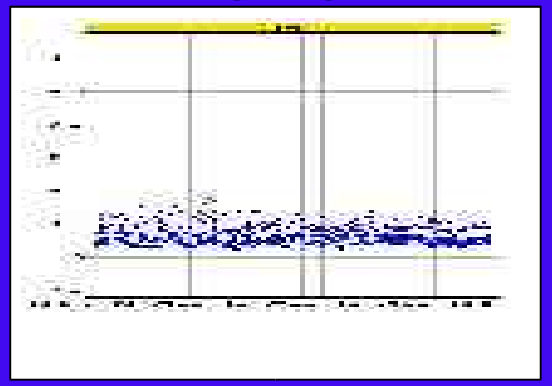
All the gene patterns found by any of the Li-Wong variants have been also found by ANOVA over treatment, some particular genes do not coincide though, but a small number.

Genes Found by Li-Wong with Time-Blocks

TREATMENT



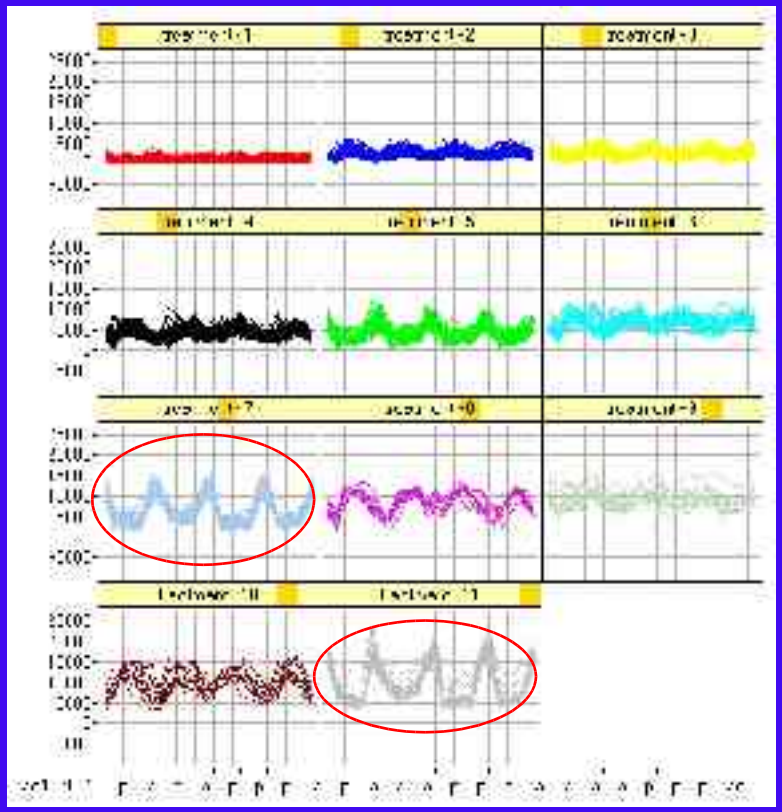
CONTROL



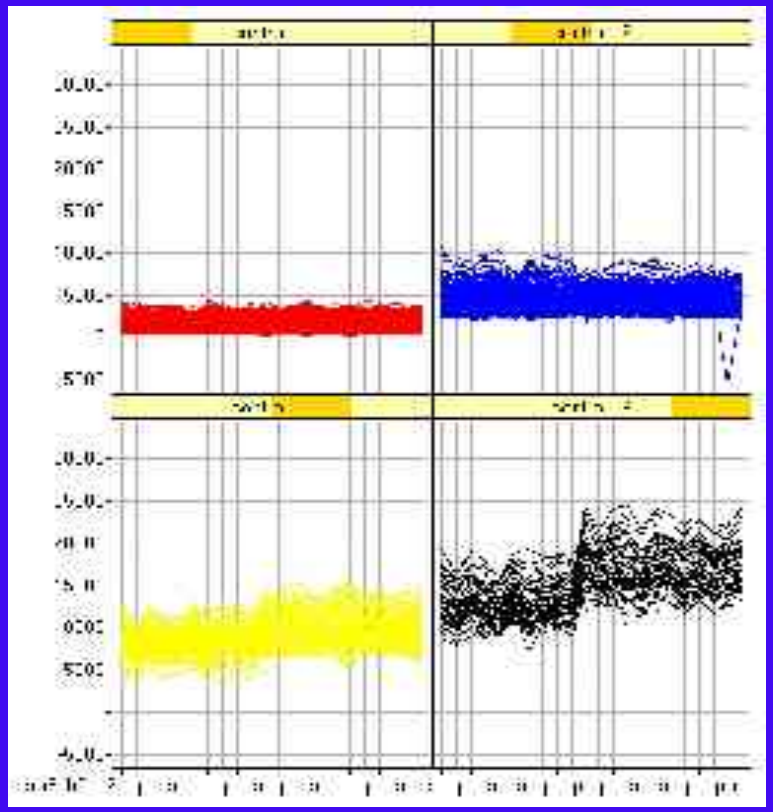
This type of profile shows very differentiated gene behavior in a patient with respect to all the others. These differences are between patients in the same experimental group, which can be caused by gender, age, conditions of the experiment... or other more interesting causes. Finding these patterns gives extra information about the experiment and should be taken into account.

Genes Found by SAM and ANOVA over Time/Treatment but not by Li-Wong

TREATMENT

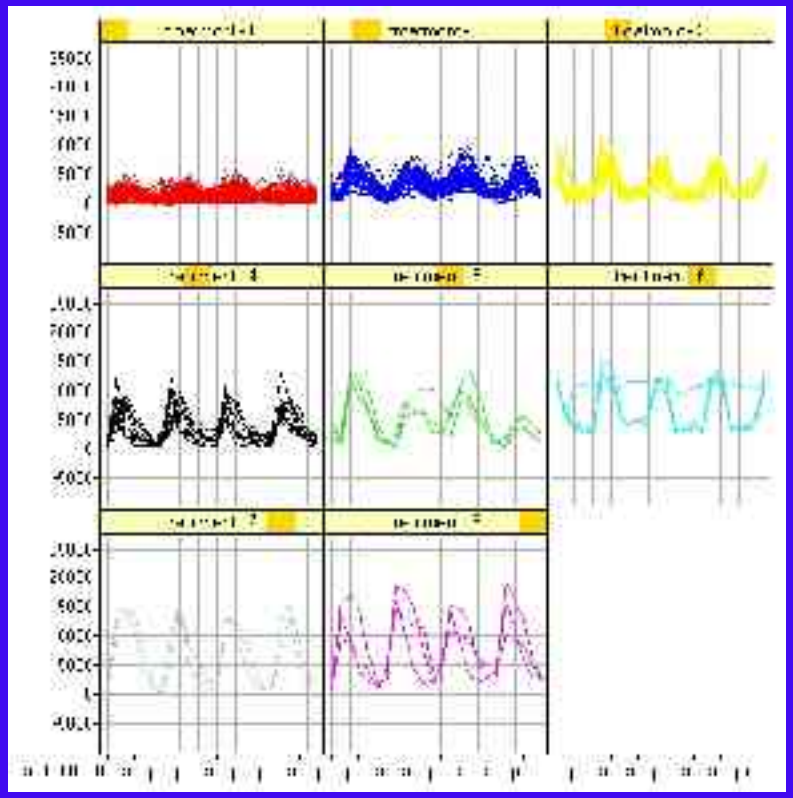


CONTROL

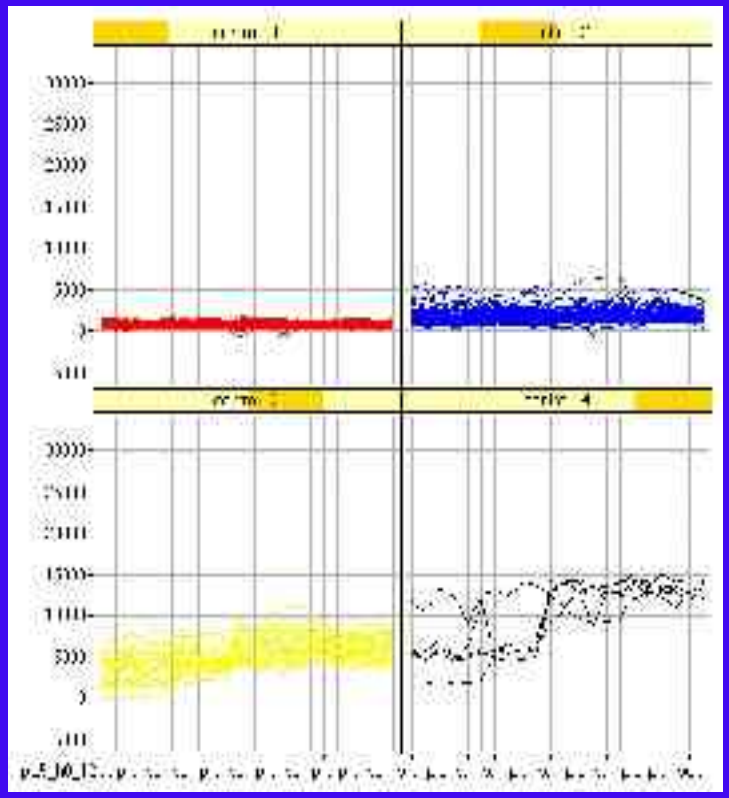


Genes Found by ANOVA over Treatment and by Li-Wong but Not by SAM

TREATMENT

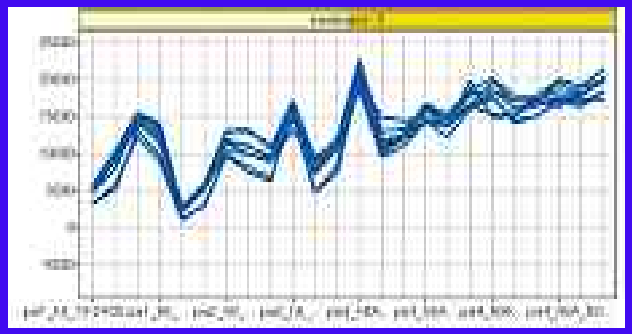


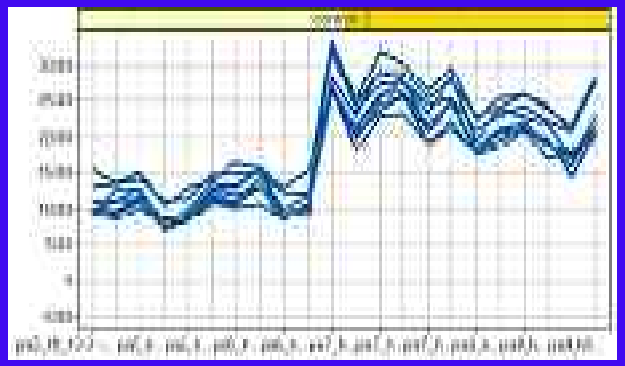
CONTROL



Genes Not Found by ANOVA over Time

TREATMENT

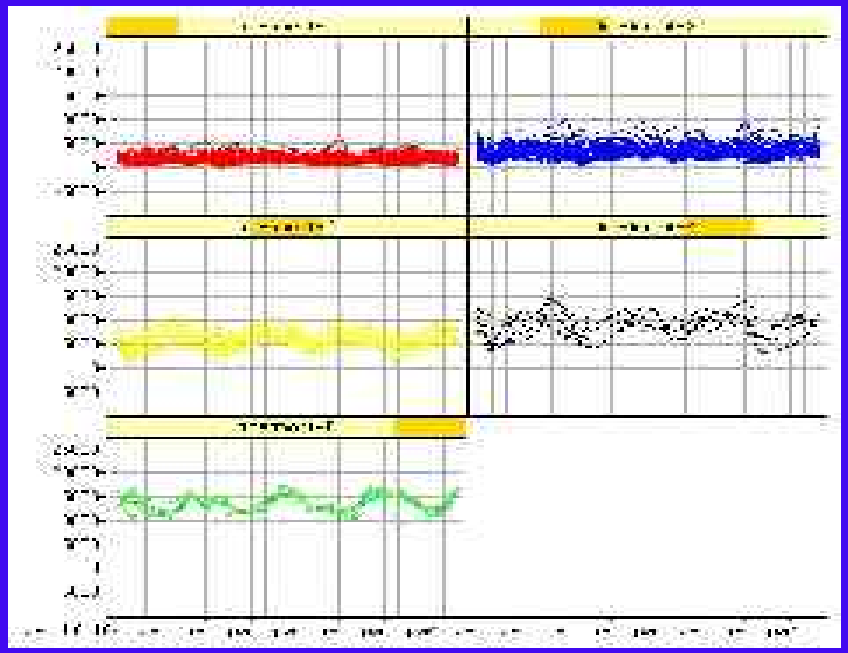




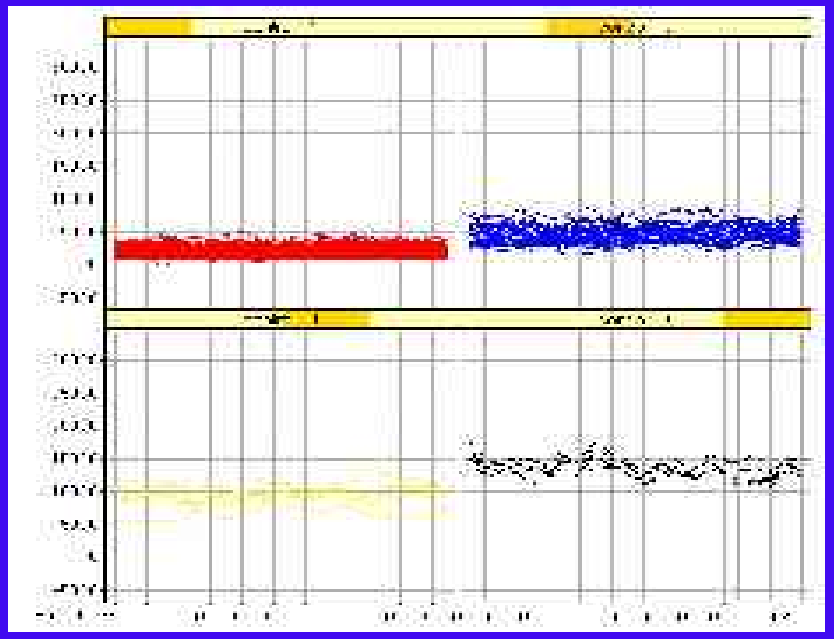
CONTROL

Genes Found by ANOVA over Time and over Time/Treatment but Not by ANOVA over Treatment

TREATMENT

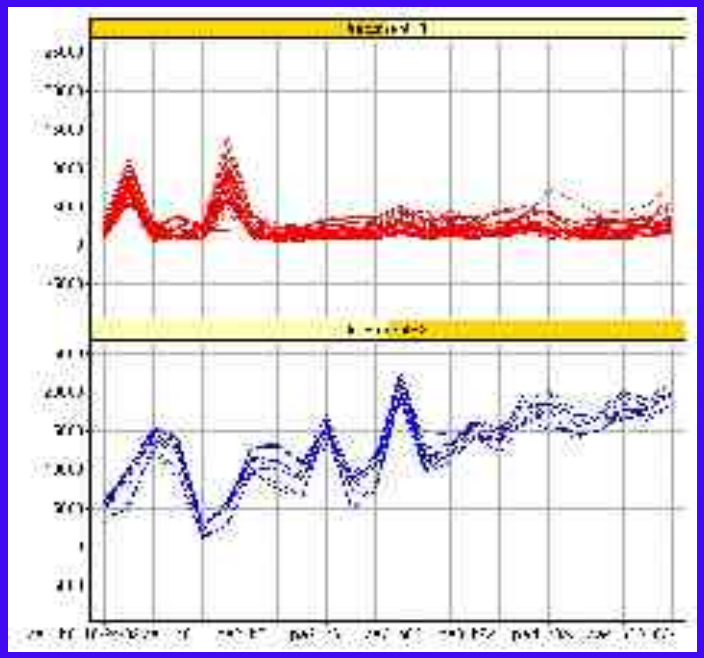


CONTROL

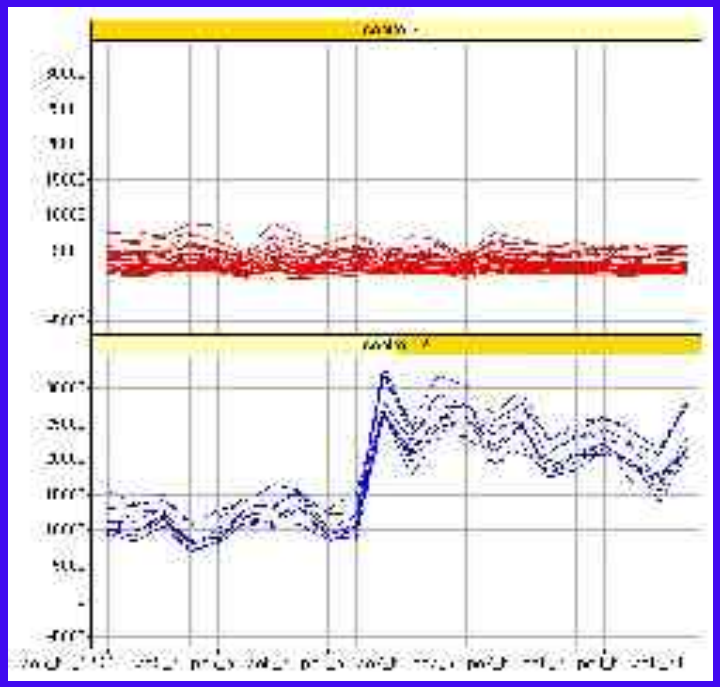


Genes Found by ANOVA over Treatment but Not by ANOVA over Time

TREATMENT

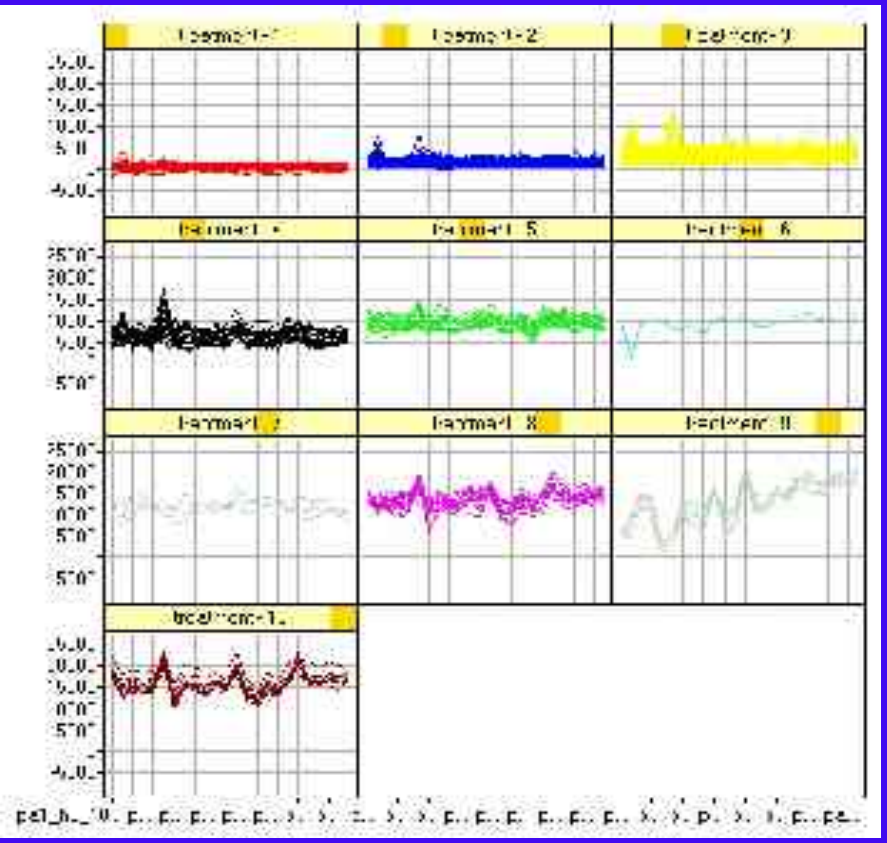


CONTROL

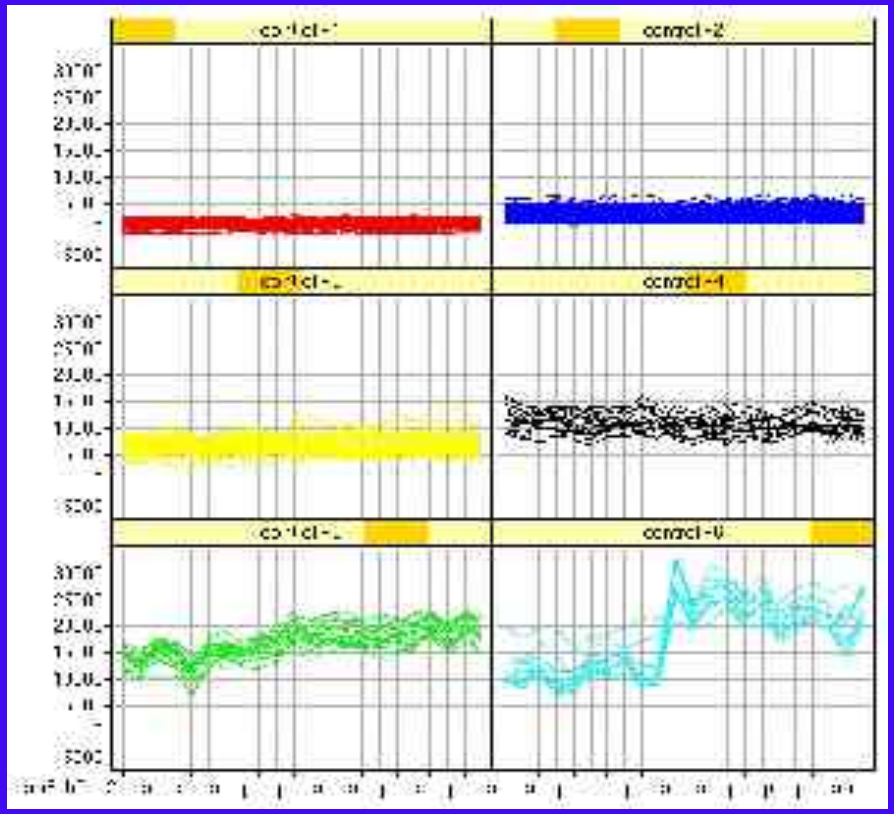


Genes Found by ANOVA over Treatment but Not by any Other Combination of Methods

TREATMENT



CONTROL



Coincidence Statistics: Pairwise Coverage of Examples

%	Li-Wong filter	Li-Wong filter + t-test	Li Wong filter + t-test w/ time-blocks	SAM over all genes	SAM over genes from Li-Wong filter	SAM over genes from broader Li-Wong filter	ANOVA over treatment	ANOVA over time	ANOVA over time/treatment
Li-Wong filter				36,61		15,027	90,16	73,31	84,15
Li-Wong filter + t-test			96,30	42,28	51,34	51,67	100	77,85	94,63
Li-Wong filter + t-test w/ time-blocks		89,13		39,44	48,14	48,14	93,17	78,57	91,61
SAM over all genes	36,33	34,42	34,69	--	27,87	58,20	87,16	56,01	73,22
SAM over genes from Li-Wong filter	--	96,77	97,48	64,15	--	97,45	97,48	75,47	91,87
SAM over genes from broader Li-Wong filter	14,03	39,28	39,54	54,33	37,5	--	72,96	57,14	82,65
ANOVA over treatment	21,86	19,74	19,88	21,13	10,27	25,58	--	49,63	63,61
ANOVA over time	25,97	23,08	25,17	20,40	11,94	22,29	74,53	--	88,65
ANOVA over time/treatment	25,77	25,60	24,69	22,43	12,22	27,11	80,53	74,556	--

table 3

Cell values correspond to $\frac{\#(Row \cap Column)}{\#Column}$ *percentage*

High values -> most genes from row method are included in those from column method

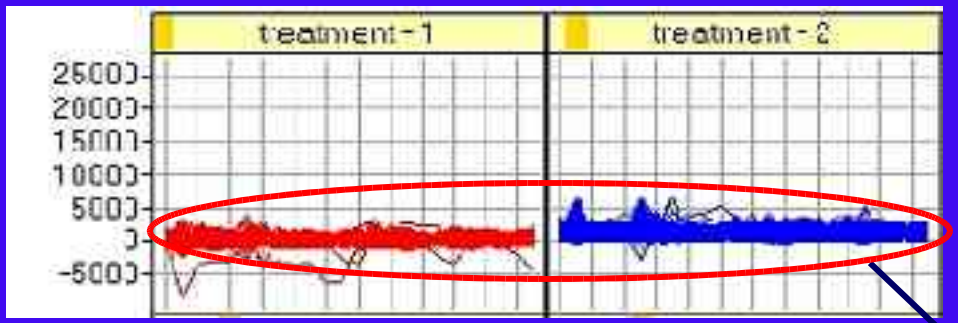
Low values -> few genes from row method are included in those from column method

Which ones are the Best Fitting Methods? The Problem of Longitudinal Data

- Clustering Visualization and Decision Making
- Dependent Longitudinal Data: Repeated Measures ANOVA
- Data Interactions: Relationship Between Gene Expression Changes and Other Features
- Profile Search by Using Similarity to Prototypes

Clustering Visualization and Decision Making

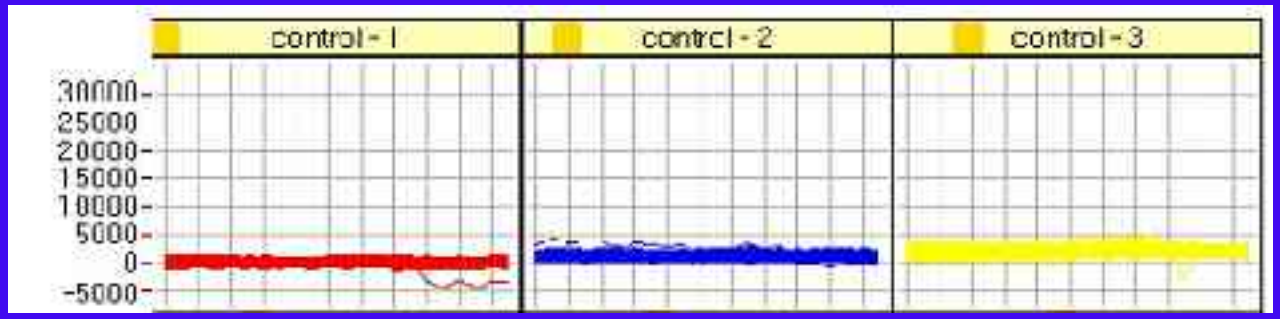
TREATMENT



From the 7217 genes 5367 do not show an explicit differential profile expressed from control to treatment. Clustering has been applied and these clusters are not further studied, at least at this point.

5367 Genes

CONTROL



Dependent Longitudinal Data: Repeated Measures ANOVA

In repeated measures design ...

- **sample members are measured on several occasions or trials**
- **each trial represents the measurement of the same characteristic under a different condition**

... but in multivariate design ...

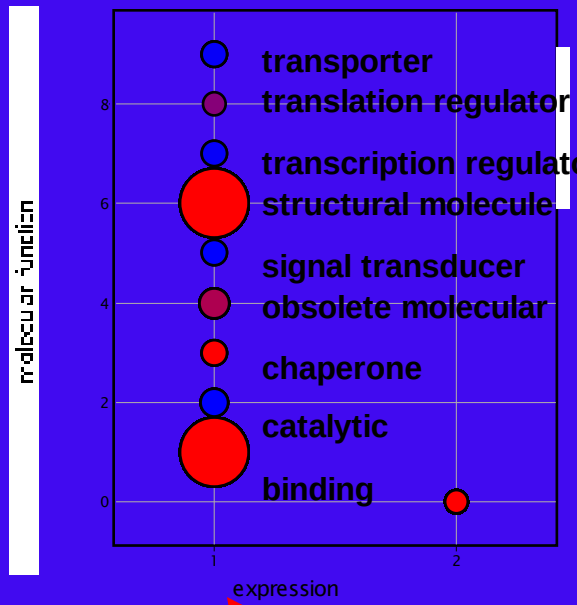
- **sample members are measured on several occasions, or trials**
- **each trial represents the measurement of a different characteristic**

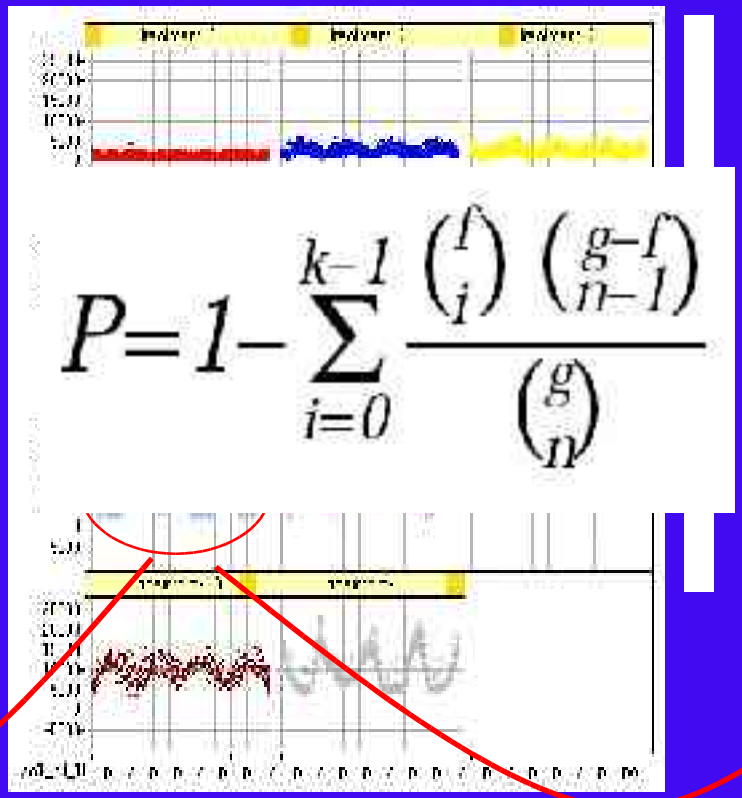
Data Interactions: Relationship Between Gene Expression Changes and …

- Gene Ontology (Molecular functions and biological processes)
- Binding sites and CIS-features
- Physical locations
- Maps of biological pathways

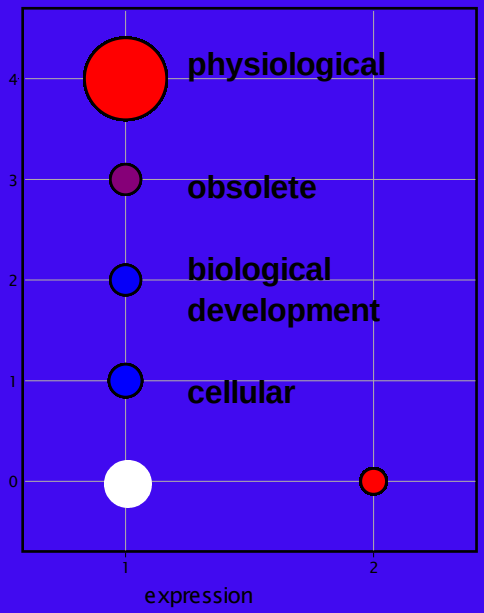
Data Interactions: Gene Ontology

Molecular Function **Expression** **Biological Process**

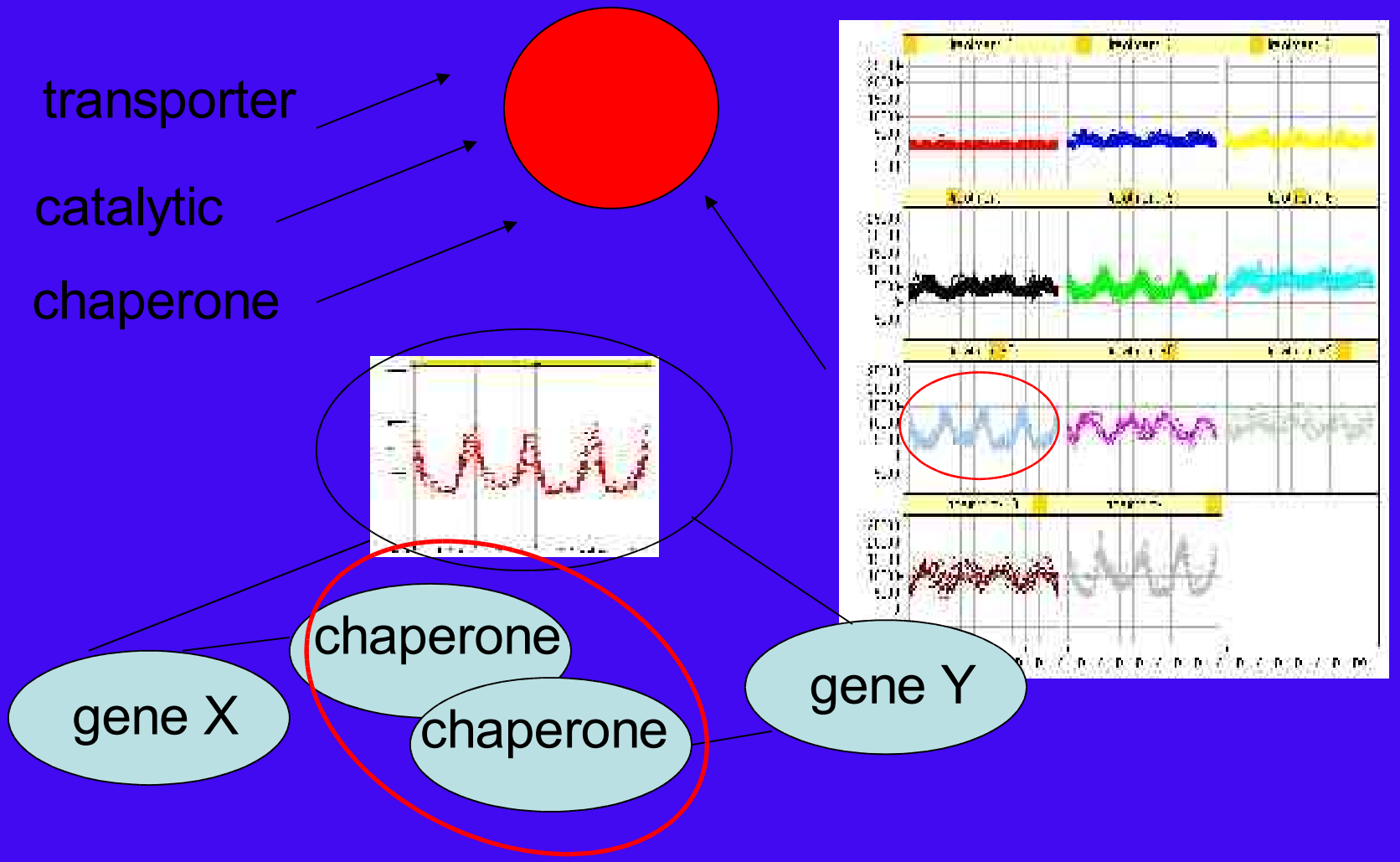




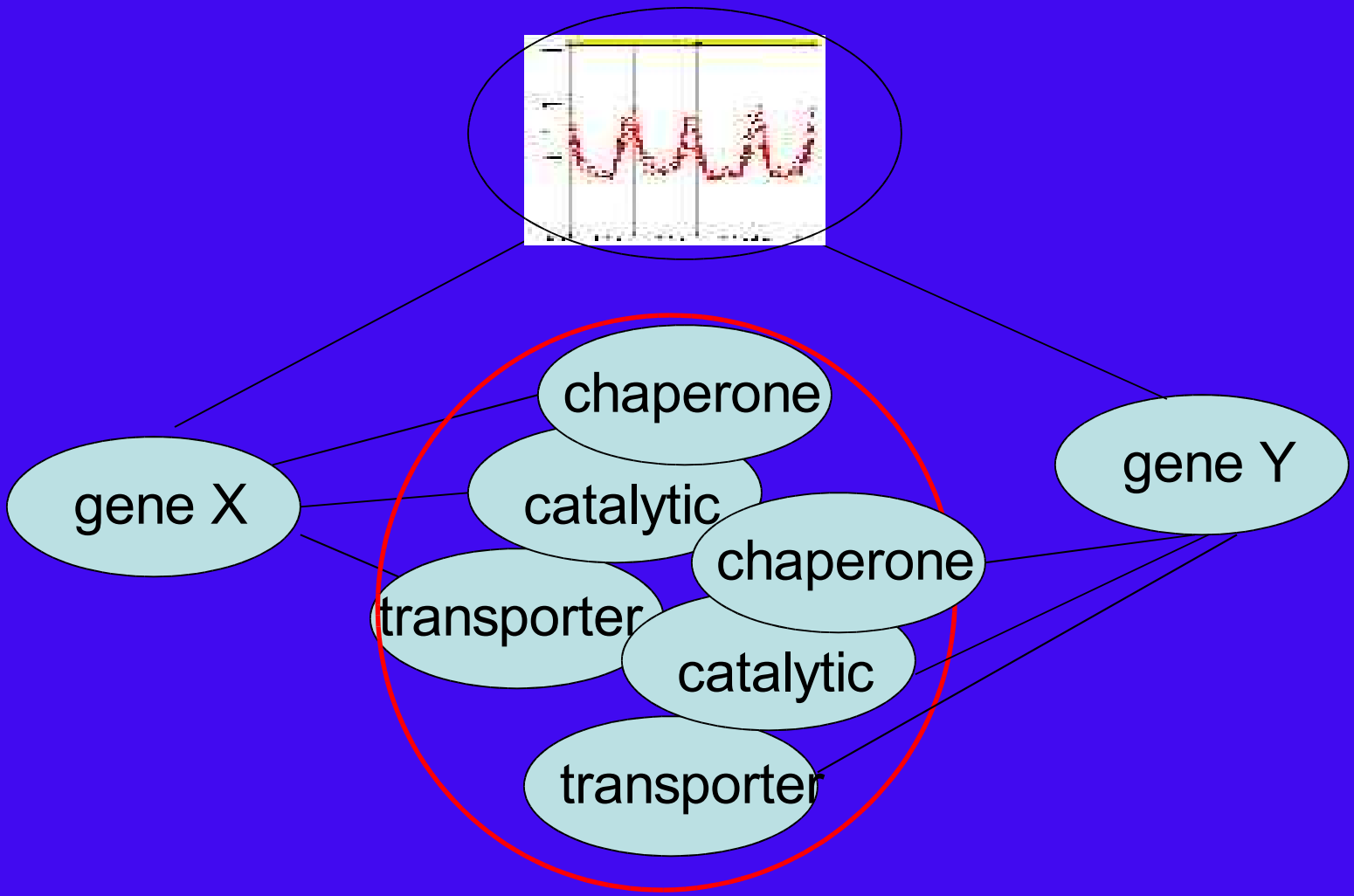
$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$$



Associations of Functions and Processes



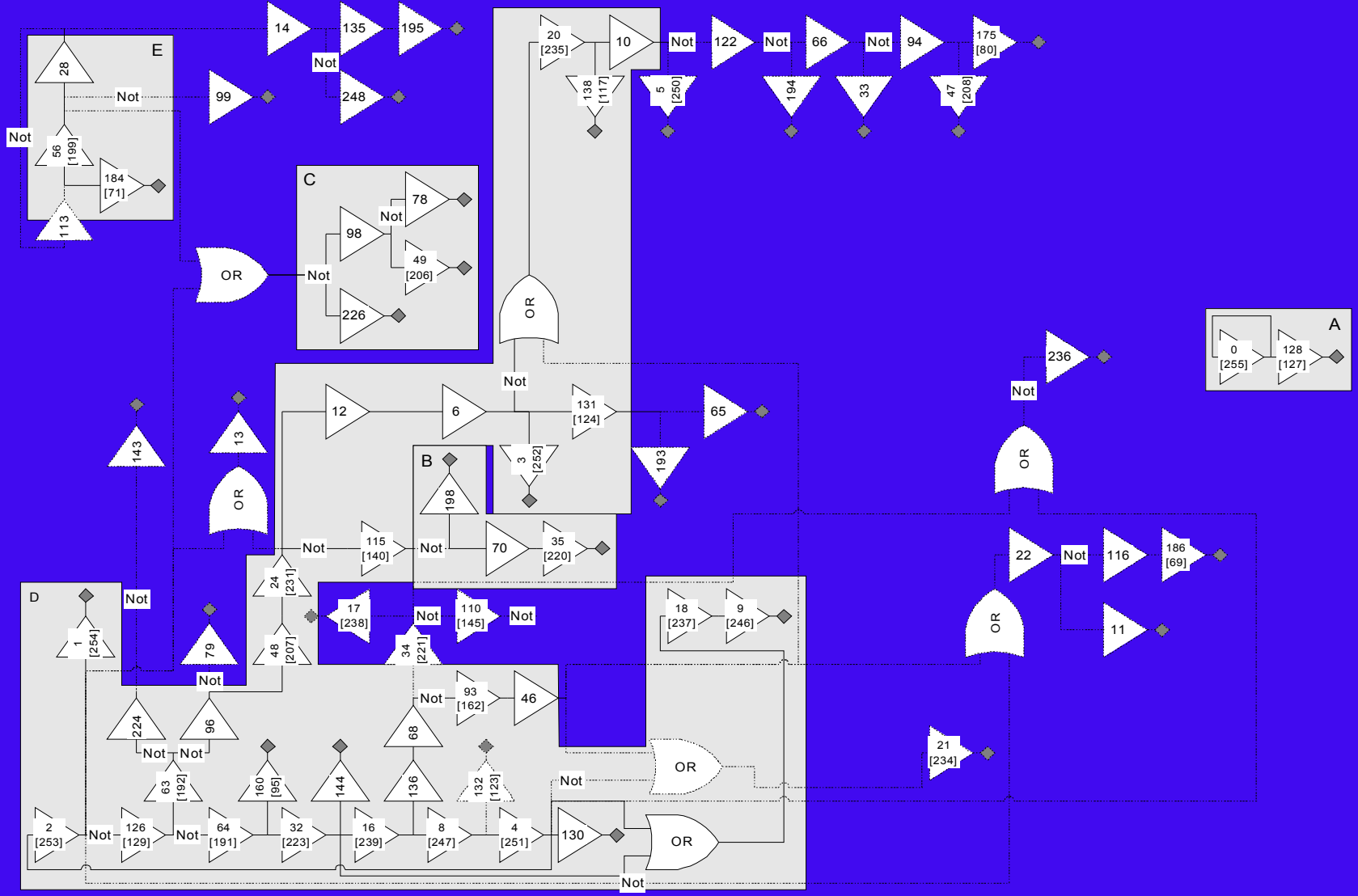
Combined Associations of Functions and Processes



Going Back to the Dynamic Modeling Problem

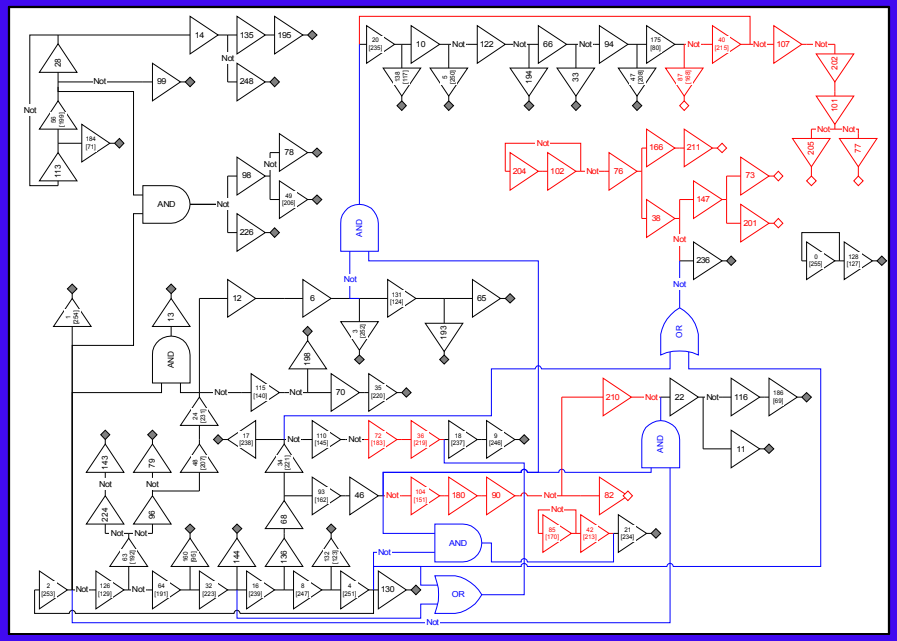
- Preprocess gene expression by clustering genes into prototypes
- Build Boolean circuits based on prototypes
- Recognize differential patterns between experiments and control
- Interpret differences
- Select interesting genes
- Go deep into continuous modeling

Recognize Differential Patterns between Experiment and Control Representations

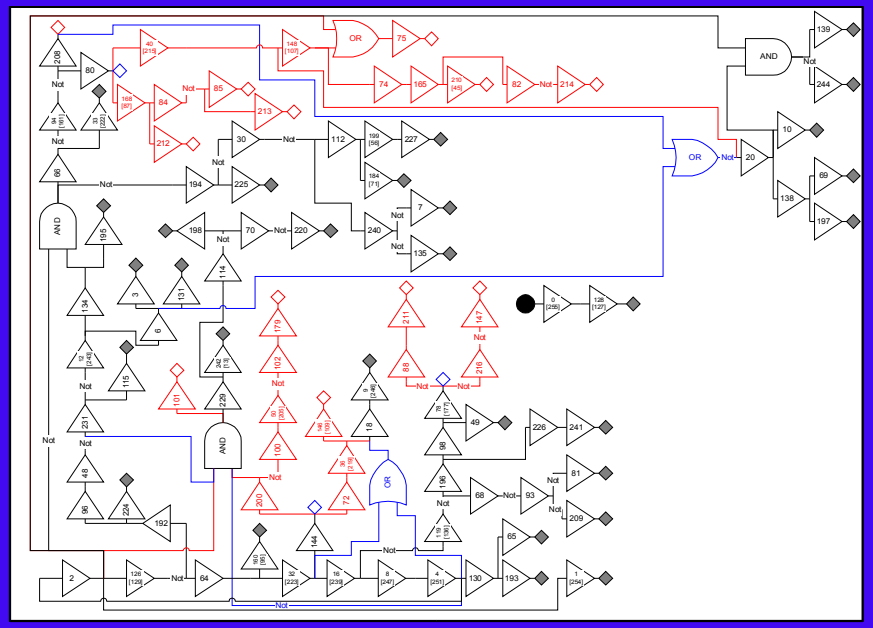


Recognize Differential Patterns between Experiment and Control Representations

TREATMENT

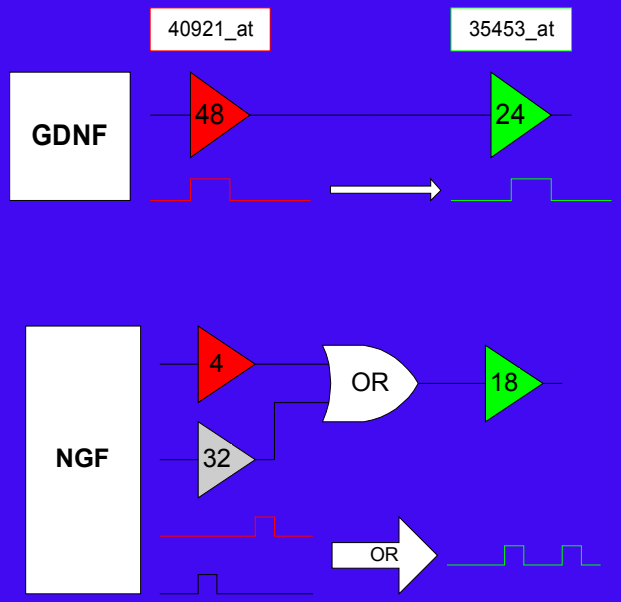


CONTROL

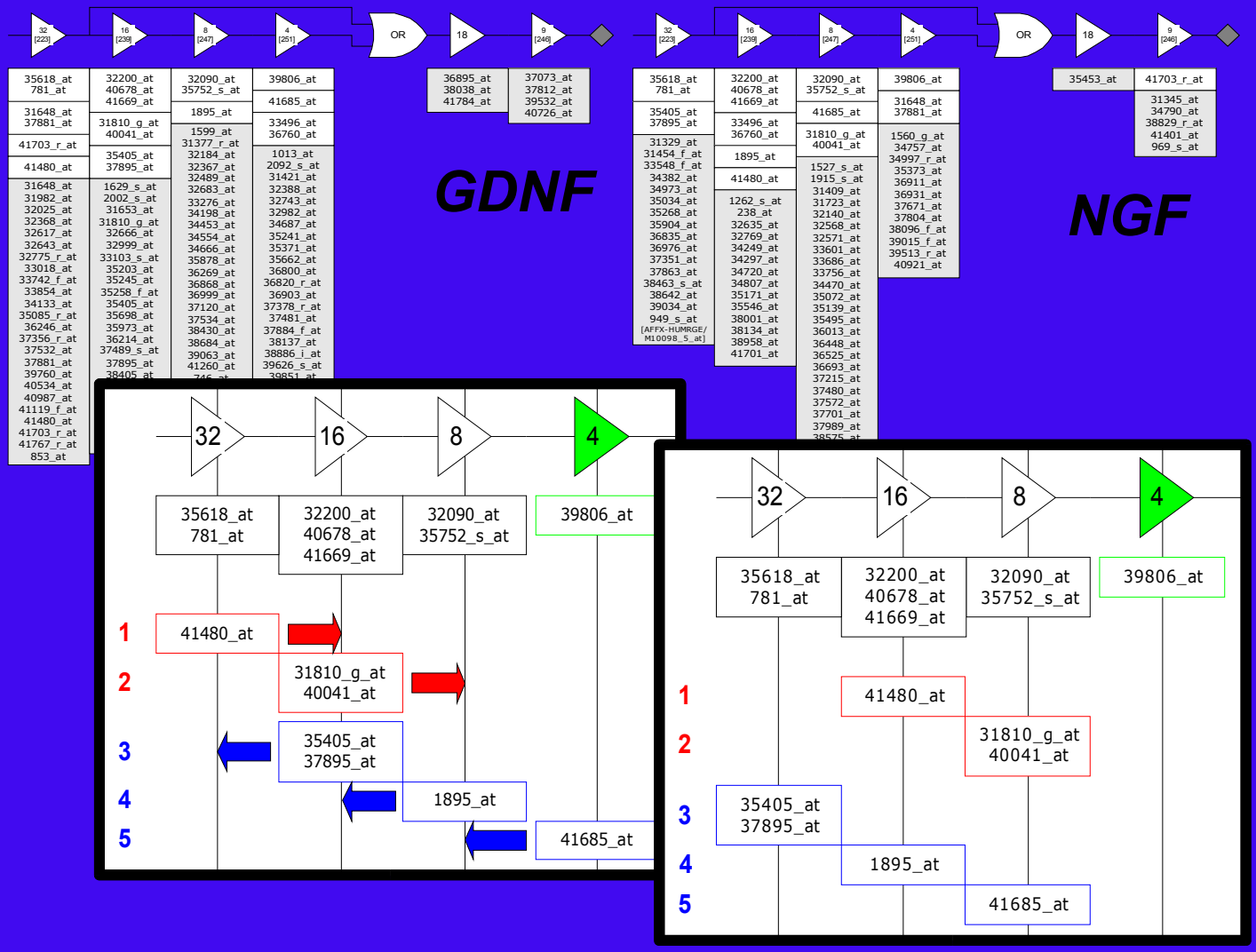


- **Find architectural differential sub-circuits**
- **Reverse gene mapping**

Find Architectural Differences as Sub-circuits, e.g., members of the nerve growth factor (NGF) and neurotrophic factor (GDNF) families



Preserve Relationships Among Genes by Using Temporal Clustering



SUMMARIZING …

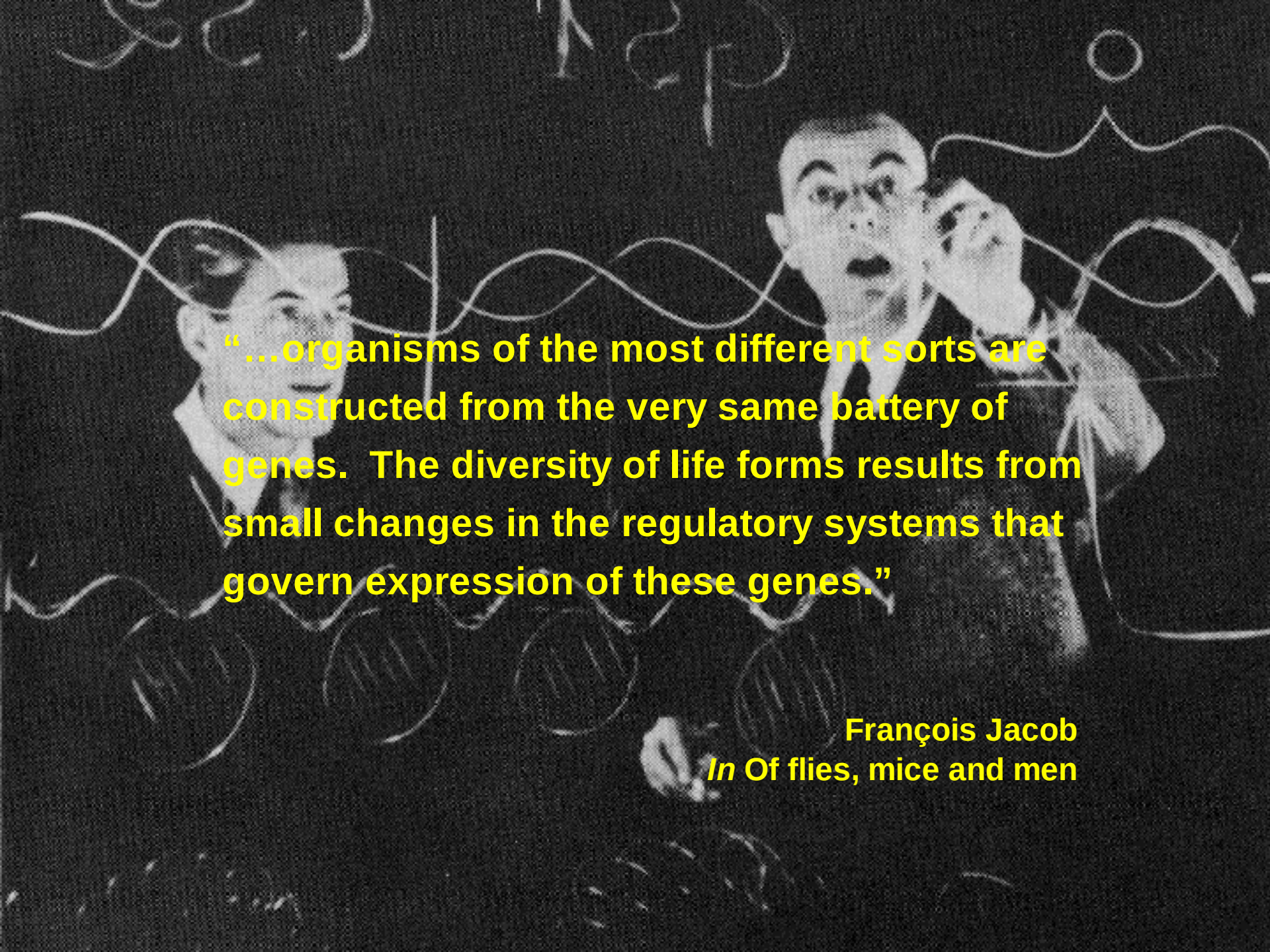
GPS, for Gene Promoter Scan

HPAM, for Hybrid Promoter Analysis Methodology

GENIE, for Gene Expression Networks Iterative Explorer

MIA, for Microarray Integrated Analysis

http://gps-tools.wustl.edu.



"…organisms of the most different sorts are constructed from the very same battery of genes. The diversity of life forms results from small changes in the regulatory systems that govern expression of these genes."

François Jacob
***In* Of flies, mice and men**