Design of Probabilistic Genetic Networks

Applications to Malaria and Cell Cycle.

Junior Barrera

DCC – IME/USP

Universidade de São Paulo

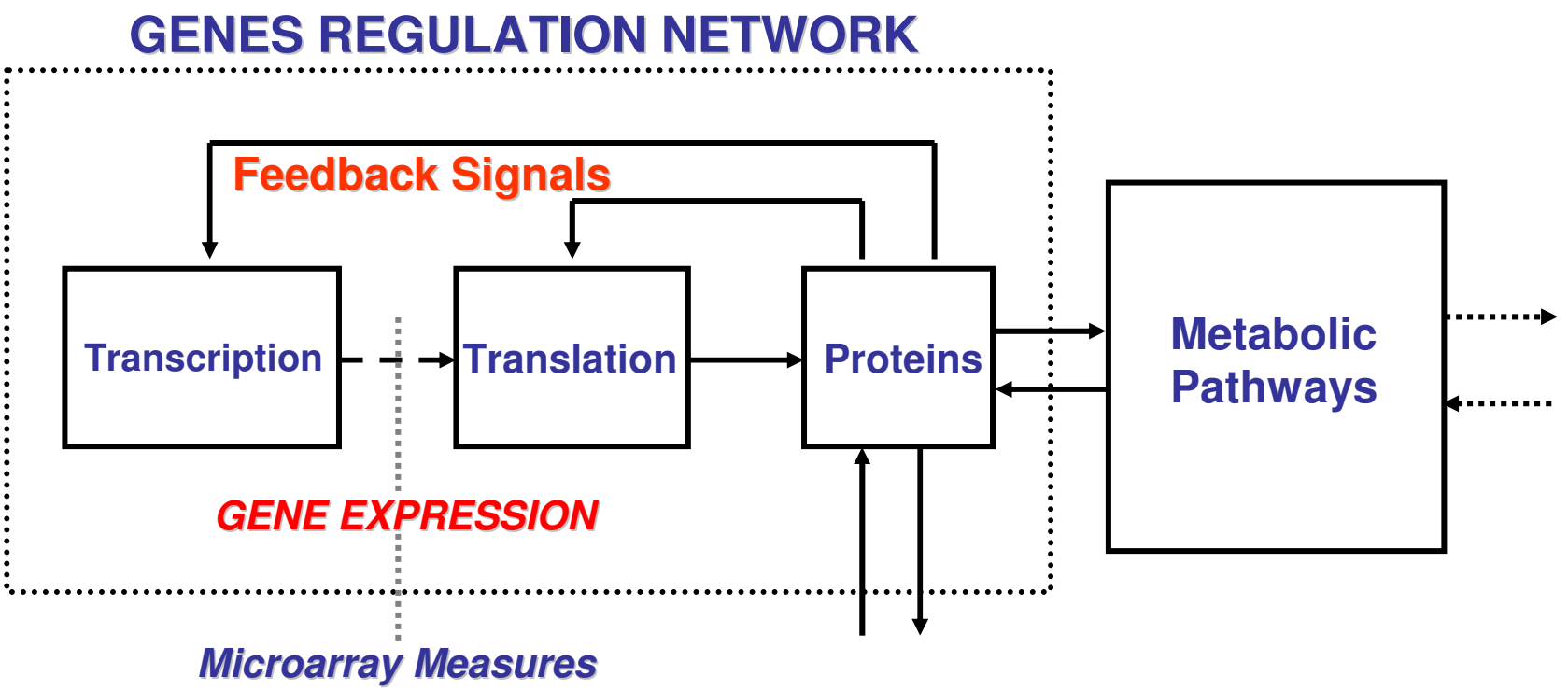
Layout

- Introduction
- Probabilistic Genetic Networks (PGN)
- Estimation of PGNs
- Architecture estimation
- Malaria
- Cell Cycle
- Future works

Layout

- Introduction
- Probabilistic Genetic Networks (PGN)
- Estimation of PGNs
- Architecture estimation
- Malaria
- Cell Cycle
- Future works

Genetic Control System



Layout

- Introduction
- Probabilistic Genetic Networks (PGN)
- Estimation of PGNs
- Architecture estimation
- Malaria
- Cell Cycle
- Future works

Model characteristics and Biological motivations

- Describes genes expression dynamics

  Gene expression is the only signal measured

- Discrete time and range

  At the time-space resolution level considered, molecular synthesis and interaction are sequential discrete phenomena

- Genes expression dynamics are stochastic signals

  There is noise in molecular synthesis and interaction

- Genes are non linear causal stochastic gates

  Transcription is regulated by the interaction between proteins and DNA

Model characteristics and Biological motivations

- PGN is built by the connection of non linear causal stochastic gates

  Genes network is the interaction of genes

- PGN dynamics is a vector of discrete stochastic signals

  Genes network dynamics is a vector of stochastic signals

Model formalization

Number of genes in the PGN: N

Expression of gene i at time t : $x_i(t) \in R \subset Z, |R|$ *finite*

State of the PGN at time $t \in \{0,1,2,...\}$: $x = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \\ \\ \\ x_N(t) \end{bmatrix}$

Model formalization

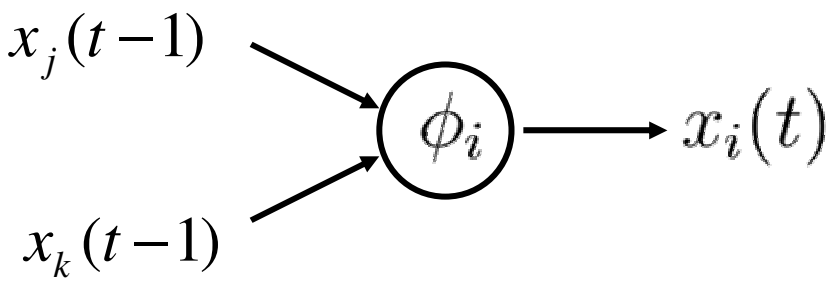
Dynamics:

$$x_i(t) = \phi_i(x(t-1)),$$
$$t > 1 \quad and \quad i \in \{1,2,...,N\}$$

ϕ_i is a stochastic function :

$$x_i(t) = \begin{cases} r_1 & p(r_1 \mid (x(t-1)) \\ r_2 & p(r_2 \mid (x(t-1)) \\ & \\ r_{|R|} & p(r_{|R|} \mid (x(t-1)) \end{cases}$$

Nomenclature



Predictors

Target

Model formalization

The sequence of random vectors $X_0, X_1, ..., X_t$,
with observations in R^N is a stochastic process

Markov chain is a stochastic process such that

$$P(X_t = x(t) \mid P(X_0 = x(0), X_1 = x(1), ..., X_{t-1} = x(t-1)) =$$

$$P(X_t = x(t) \mid X_{t-1} = x(t-1))$$

A Markov chain is homogenous when

$P(X_t = x(t) \mid X_{t-1} = x(t-1))$ Is the same for every t

An homogeneous Markov chain is characterized by the Distribution of X_0 and the state transition probability matrix:

$$(\pi_0, \pi_{Y/X})$$

Model formalization

A PGN is an homogenous Markov chain $(\pi_0, \pi_{Y/X})$ which is

- conditionally independent,

$$P(x[t+1] \mid x[t]) = \prod_{i=1}^{n} p(x_i[t+1] \mid x[t])$$

- almost deterministic and

$$\forall t, \forall x(t) \in R^N, \forall i \in \{1,..., N\}, \exists r \in R : p(r \mid x(t)) \approx 1$$

- has limited dependence.

$$\forall i \in \{1,.., N\}, \exists j,\ j << N : \forall t, \forall x(t) \in R^N, \forall r \in R,\ p(r \mid x(t)) = p(r \mid x^r_i(t))$$

Characterization

- Markov chain

$$P(x(t) \mid x(t-1))$$

Matrix dimension: $|R|^N$ x $|R|^N$

- PGN,

$$p(x_i(t) \mid x(t-1))$$

Matrix dimension: $|R|^j$ x $|R|$ x N

Layout

- Introduction
- Probabilistic Genetic Networks (PGN)
- Estimation of PGNs
- Architecture estimation
- Malaria
- Cell Cycle
- Future works

Distribution estimation

Training data for the target gene i

$$(x_1^j, y_1), (x_2^j, y_2), \ldots, (x_m^j, y_m)$$

Estimator

$$n = \left| \left\{ (x_k^j, y_k) : x_k^j = x^j \right\} \right|$$

$$p = \left| \left\{ (x_k^j, y_k) : x_k^j = x^j \wedge y_k = r \right\} \right|$$

$$\hat{P}_i(X^j = x^j) = \frac{n}{m} \qquad \hat{P}_k(Y = r \mid X^j = x^j) = \frac{p}{n}$$

Distribution estimation

Problem

Luck of data, usually there are non observed x^j

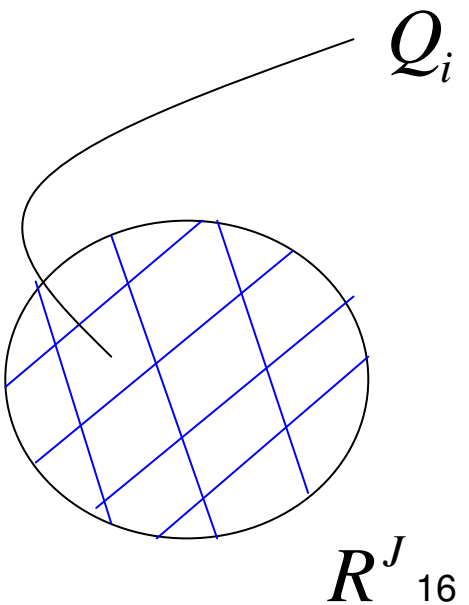
Solution

Constraint the probability families considered partitioning R^j

$Q = \{Q_1, Q_2, ..., Q_n\}$ is a partition of R^J

$$\forall x_k^j \in Q_i : P(y \mid x_k^j) = P(y \mid Q_i)$$

$$P(Q_i) = \sum_{x^j \in Q_i} P(x^j)$$



Distribution estimation

Training data for the target gene i

$$(x_1^j, y_1), (x_2^j, y_2), \ldots, (x_m^j, y_m)$$

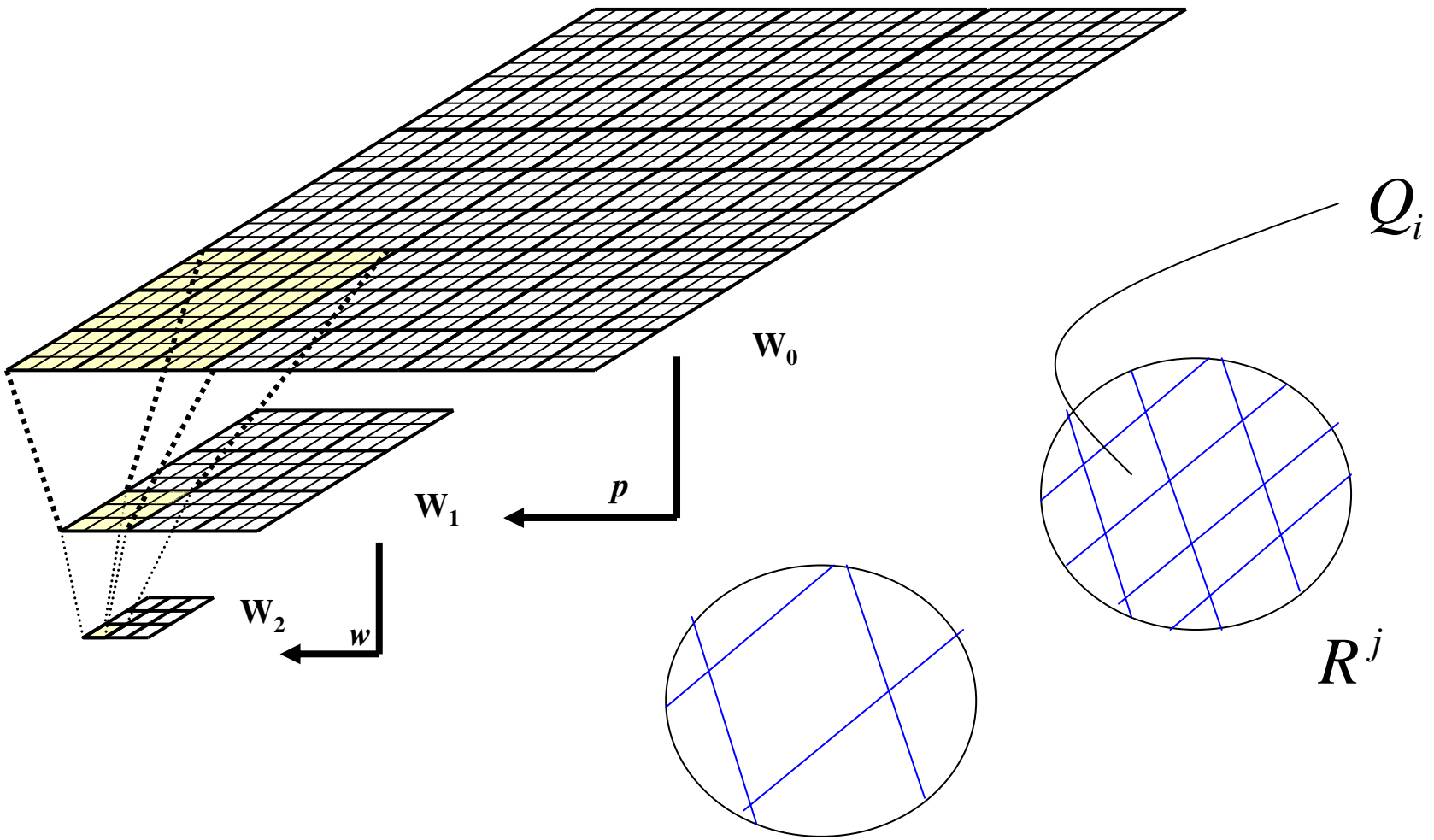
Distributions constraint estimator

$$N_{Q_i} = \sum_{k=1}^{m} c_{Q_i}(x_k^j) \qquad c_{Q_i}(x^j) = 1 \Leftrightarrow x^j \in Q_i$$

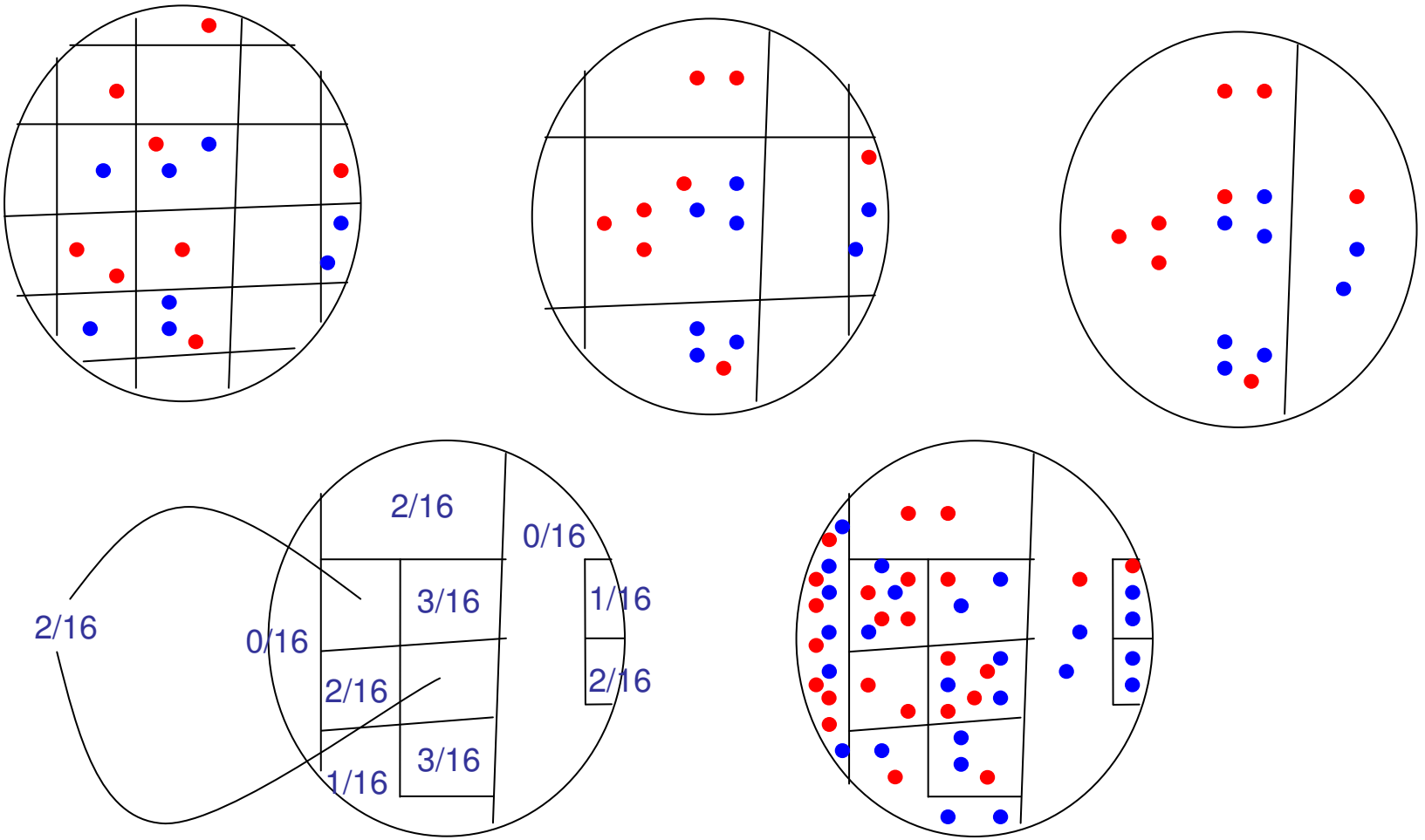
$$L_{Q_i,r} = \sum_{k=1}^{m} l_{Q_i,r}, r(x_k^j, y_k) \qquad l_{Q_i,r}(x^j, y) = 1 \Leftrightarrow x^j \in Q_i \wedge y = r$$

$$\hat{P}(Q_i) = \frac{N_{Q_i}}{m} \qquad \hat{P}(r \mid Q_i) = \frac{L_{Q_i,r}}{N_{Q_i}}$$

Multi-resolution



Example: multi-resolution estimation



P(X) **P(Y|X)**

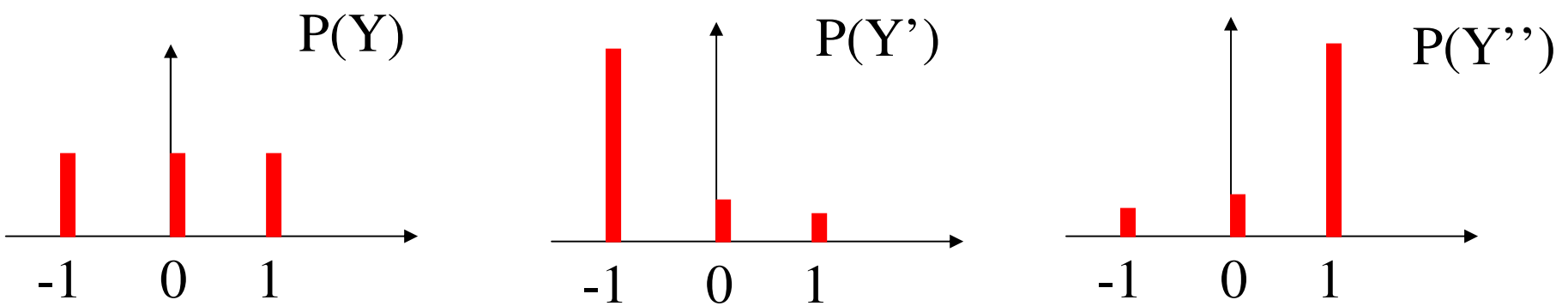
Measuring estimator quality

Entropy

$$H(Y) = -\sum_{y \in \{-1,0,1\}} P(y) \log P(y)$$

Mutual information

$$I(X,Y) = H(Y) - H(Y \mid X) \geq 0$$



Distributions of Y

$$H(Y) > H(Y') = H(Y'')$$

Measuring estimator quality

Mean conditional entropy

$$E[H(Y \mid X)] = \sum P(X) H(Y \mid X)$$

Mean mutual information

$$E[I(X,Y)] = H(Y) - E[H[Y \mid X]]$$

Mean mutual information estimation

$$\hat{E}[H(Y \mid X)] = -\sum \hat{P}(X) \sum \hat{P}(Y \mid X) \log(\hat{P}(Y \mid X)).$$

$$\hat{E}[I(X,Y)] = H(\hat{Y}) - \hat{E}[H(Y \mid X)]$$

Optimal estimator

Mean conditional entropy for a constraint distribution

$$E[H(Y \mid X)] = \sum_{i=1}^{n} P(Q_i) H(Y \mid Q_i)$$

Choose a space of partitions and choose the ones that minimize

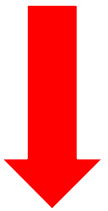
$$\hat{E}[H(Y \mid X)]$$

Example: all possible partitions generated by a family of classifiers; partitions generated by a family of pyramid's

Layout

- Introduction
- Probabilistic Genetic Networks (PGN)
- Estimation of PGNs
- Architecture estimation
- Malaria
- Cell Cycle
- Future works

$$(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$$



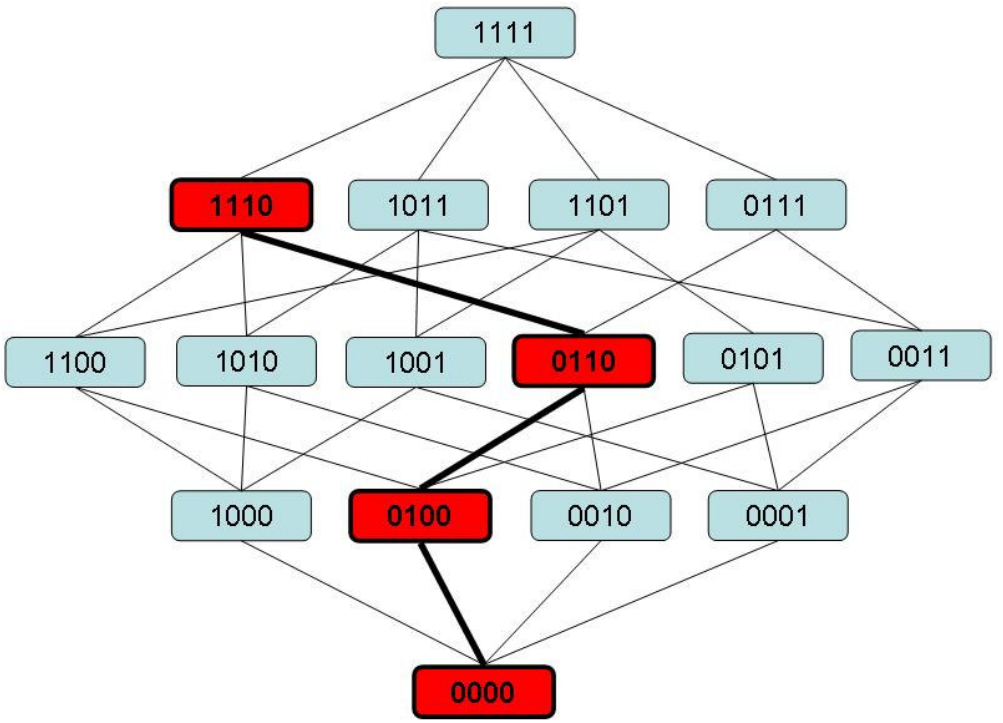
Feature selection

$$x \equiv x^j$$

$$A \subset \{1,2,\ldots,N\}$$

$$j = |A|$$

Feature selection



Boolean Lattice

Feature selection

- A particular case of distribution estimation

  Equivalence classes built eliminating features

- Processing a node

  For each feature set several partitions may be compared

- Linear partitions

  May be used for generating the search space for a feature set

- Walking through the search space

  Some heuristics: SFS, SFFS, U-curve

Inhibitory and excitatory interactions

Interaction parameter: $a_i \in K = \{-k,...,0,...k\}$

$$\sum_{i=1}^{N} a_i x_i(t) = \sum_{i=1}^{N} a_i x_i(t+\Delta) \Rightarrow$$

$$P(r \mid x(t)) = P(r \mid x(t+\Delta))$$

- A vector of parameters $a \in K^N$ defines a partition

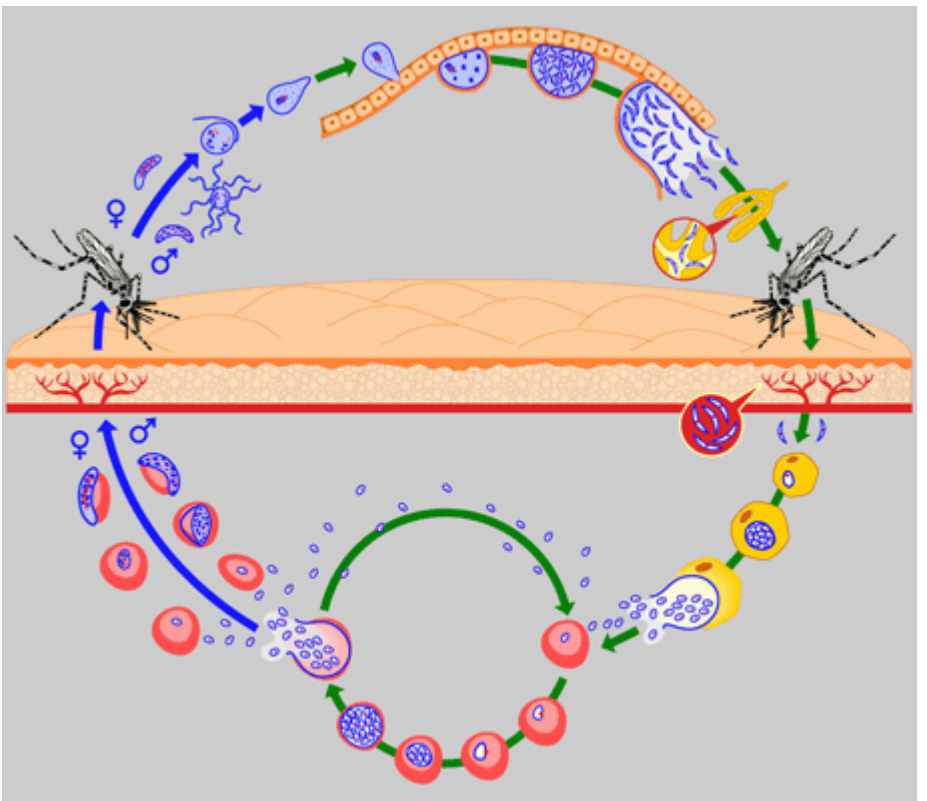
- The search space is generated by $a \in K^N$

- The best vector gives interaction weight and signal

Layout

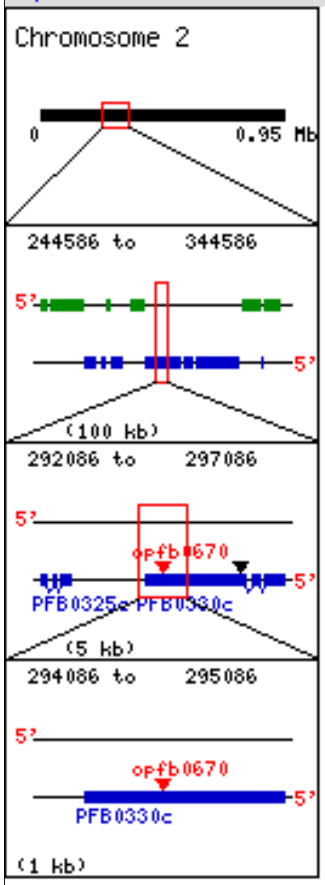
- Introduction
- Probabilistic Genetic Networks (PGN)
- Estimation of PGNs
- Architecture estimation
- Malaria
- Cell Cycle
- Future works

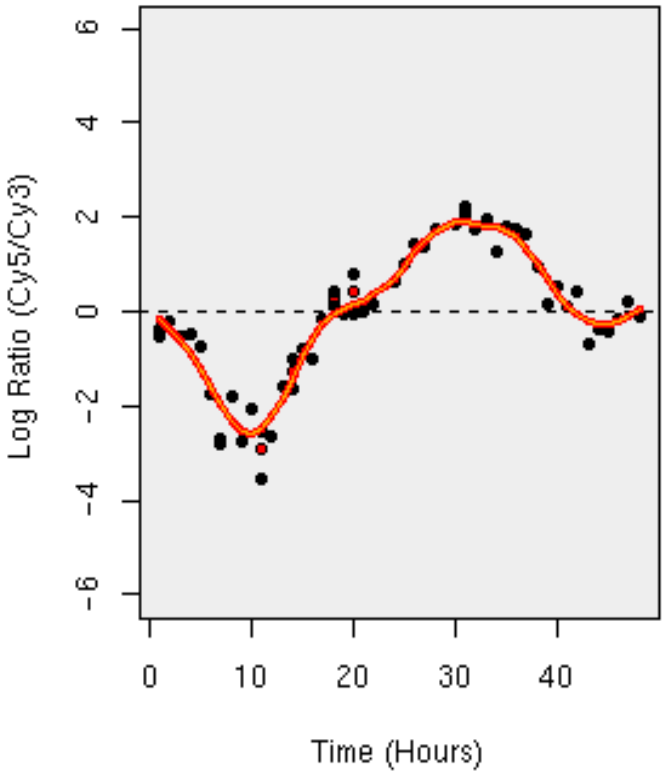
The life cycle of the malaria parasite

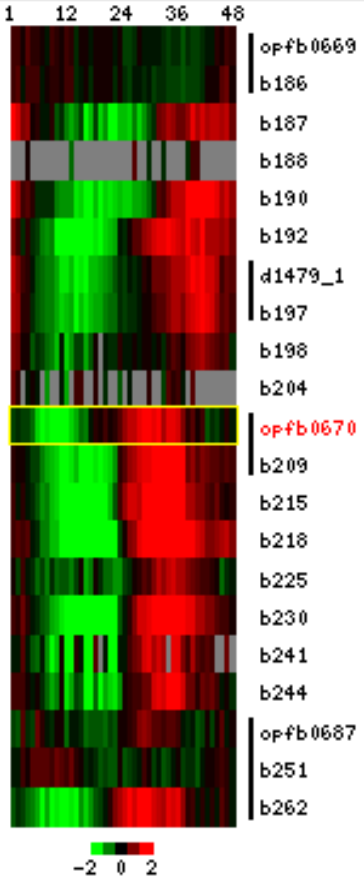


Back

OligoID	Status	Maximum Hour	Minimum Hour	Amplitude (log2)	Score (%)	Phase (-Pi to +Pi)	CGH %3D7	Avg. Med. Intensity
opfb0670	UNIQUE	30	10	4.5	87	0.06	89	3211.57







OLIGO

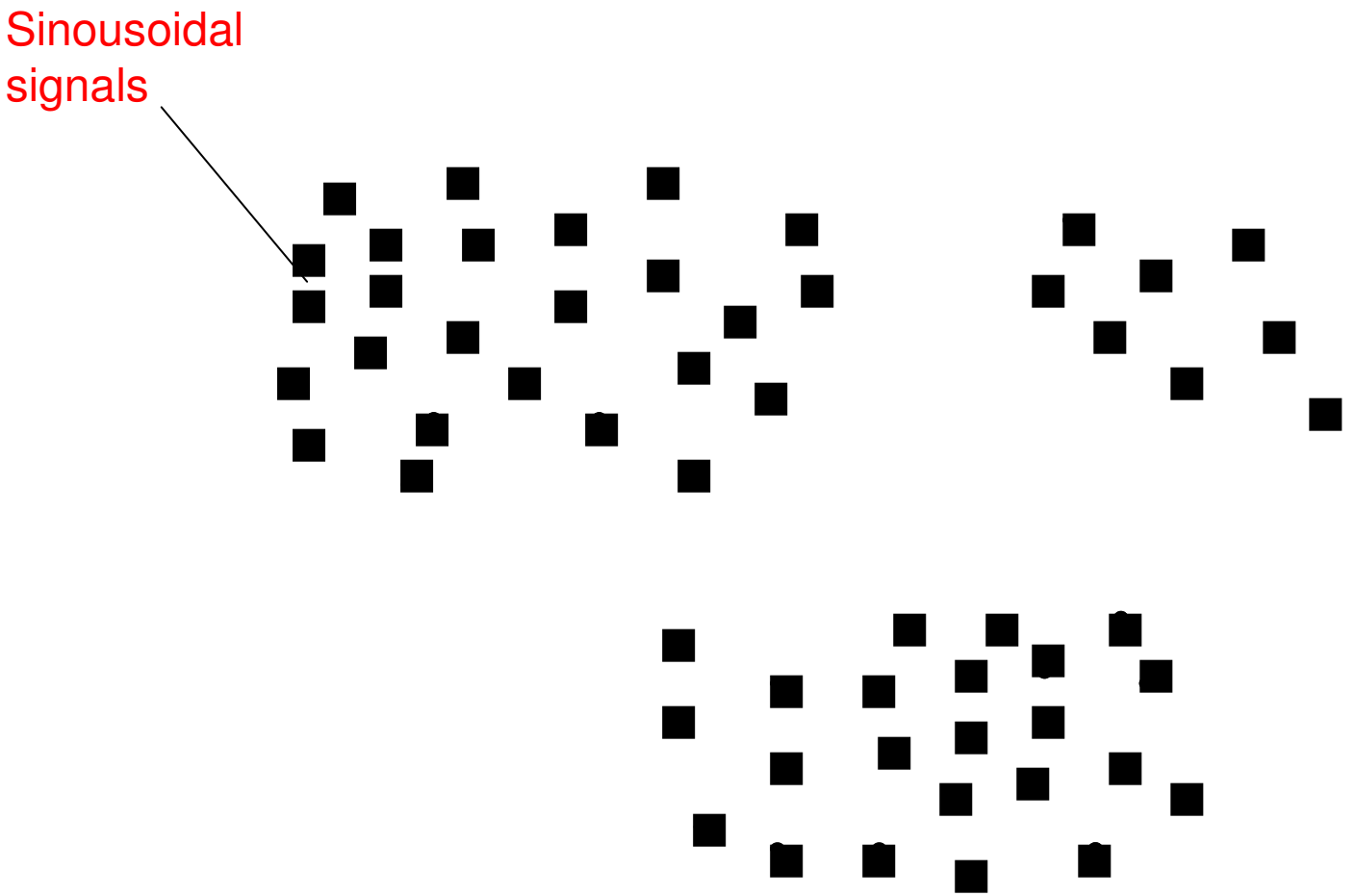
PlasmoDB ID	Description
PFB0330c	cysteine protease, putative

Oligo Sequence BLAST @ PlasmoDB

5'

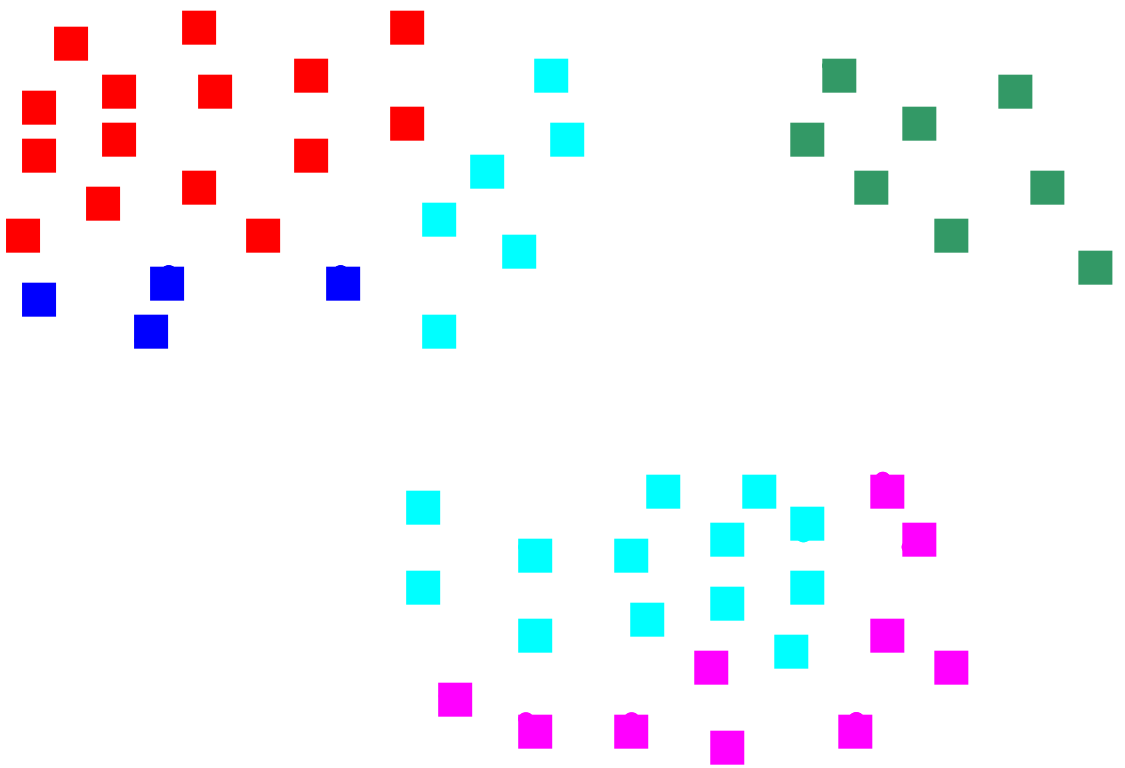
CTGCCCAAGATGAGCCACCTACTGATAATGTAGAATCACAAGCAGAAAATAACAAAAAAACAGAAATTTA

Malaria parasite genes with almost sinusoidal signals



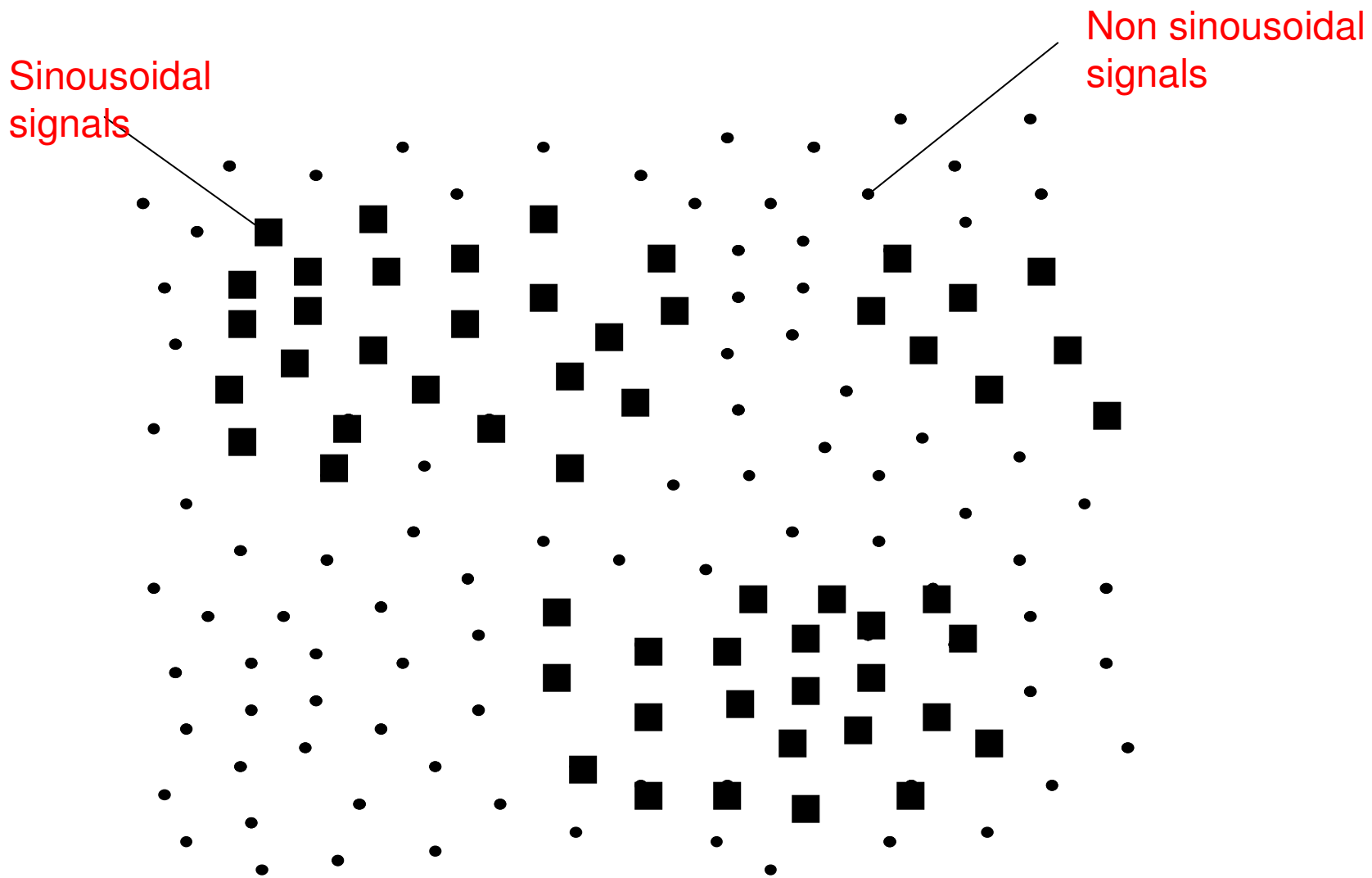
DeRisi, 2003.

Functional Classification

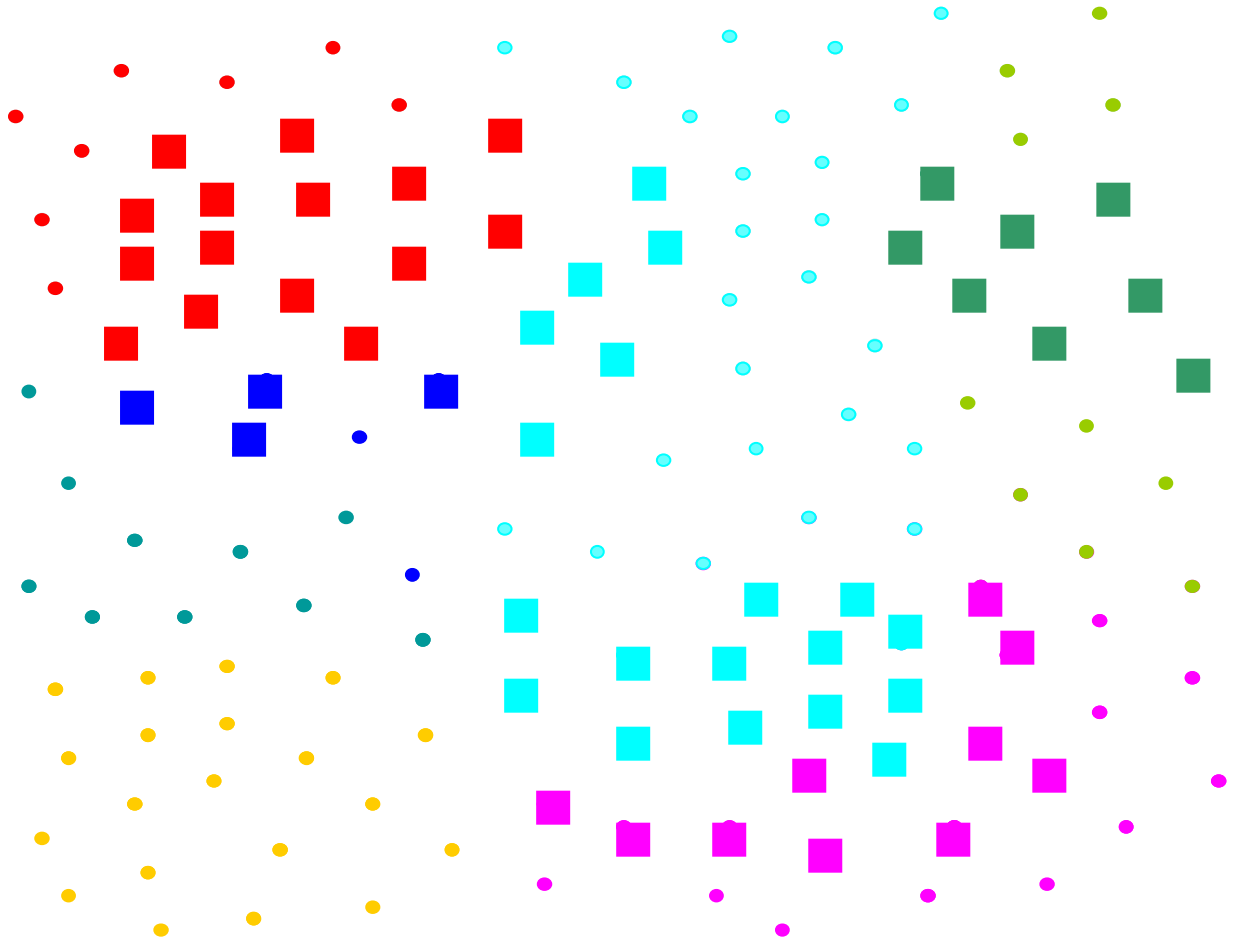


DeRisi, 2003.

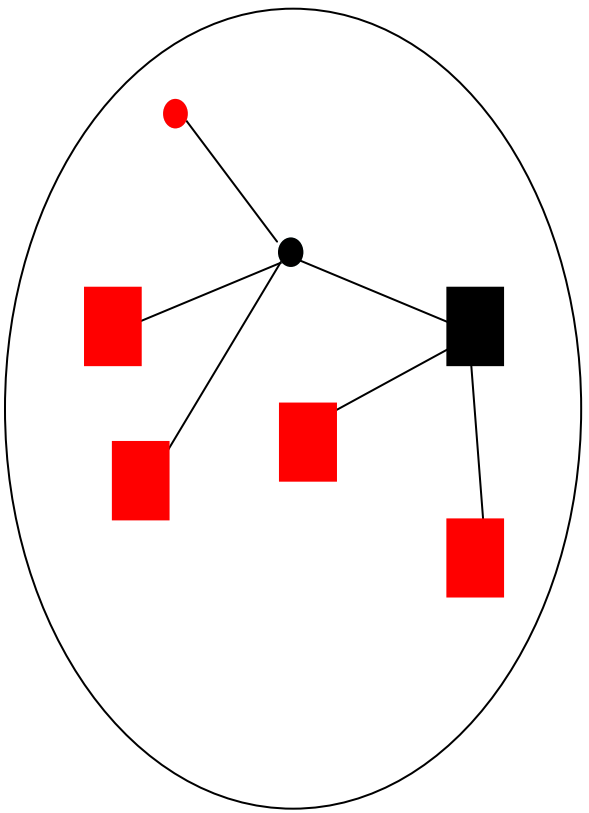
Malaria parasite genes



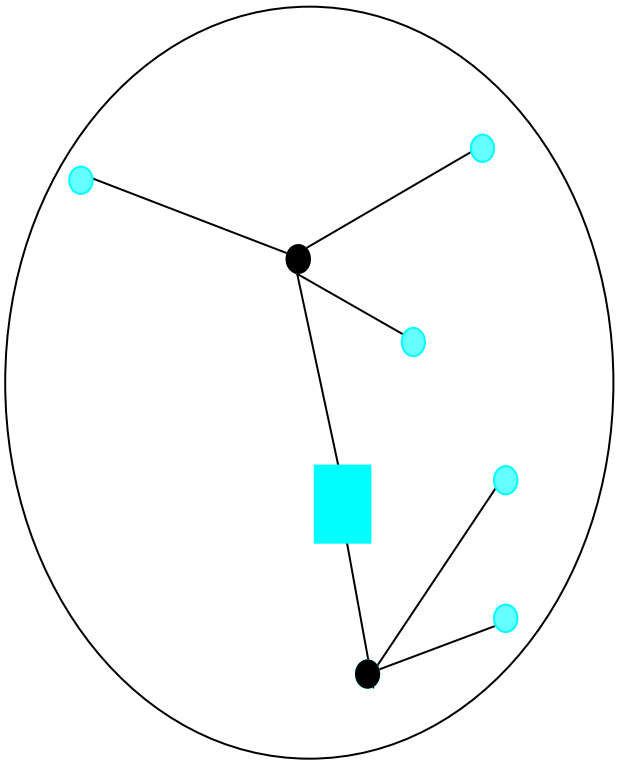
Functional Classification



Interaction Graph

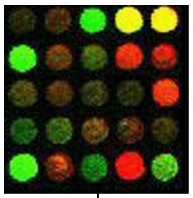


Glycolisis



Apicoplast

System architecture



USP dataset determination

USP dataset

Scaling and quantization

Quantized USP dataset

Target genes

Design of PGN

Functional groups

Overview set

GraphViz

Table of predictors

Output graph

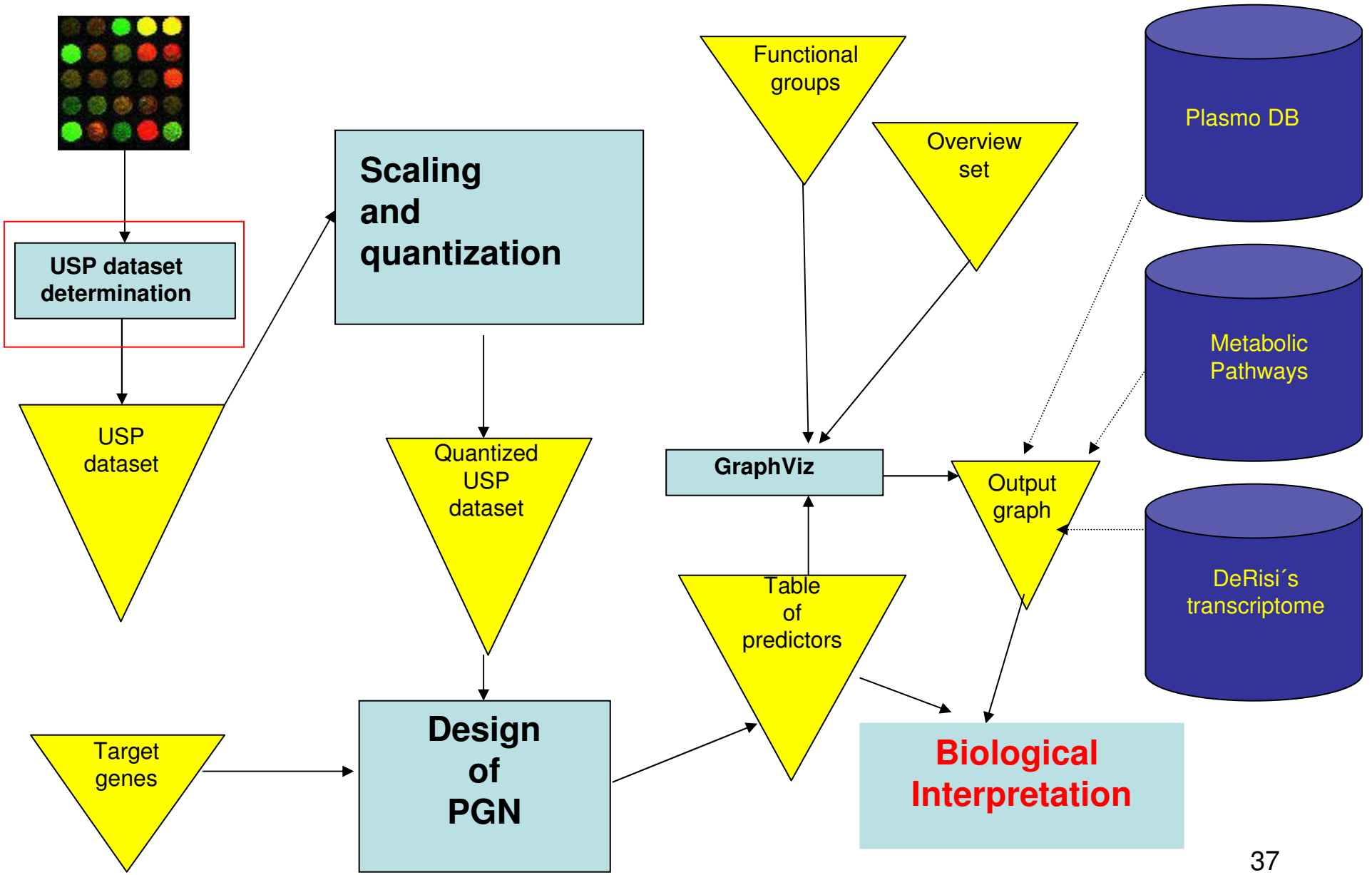
Plasmo DB

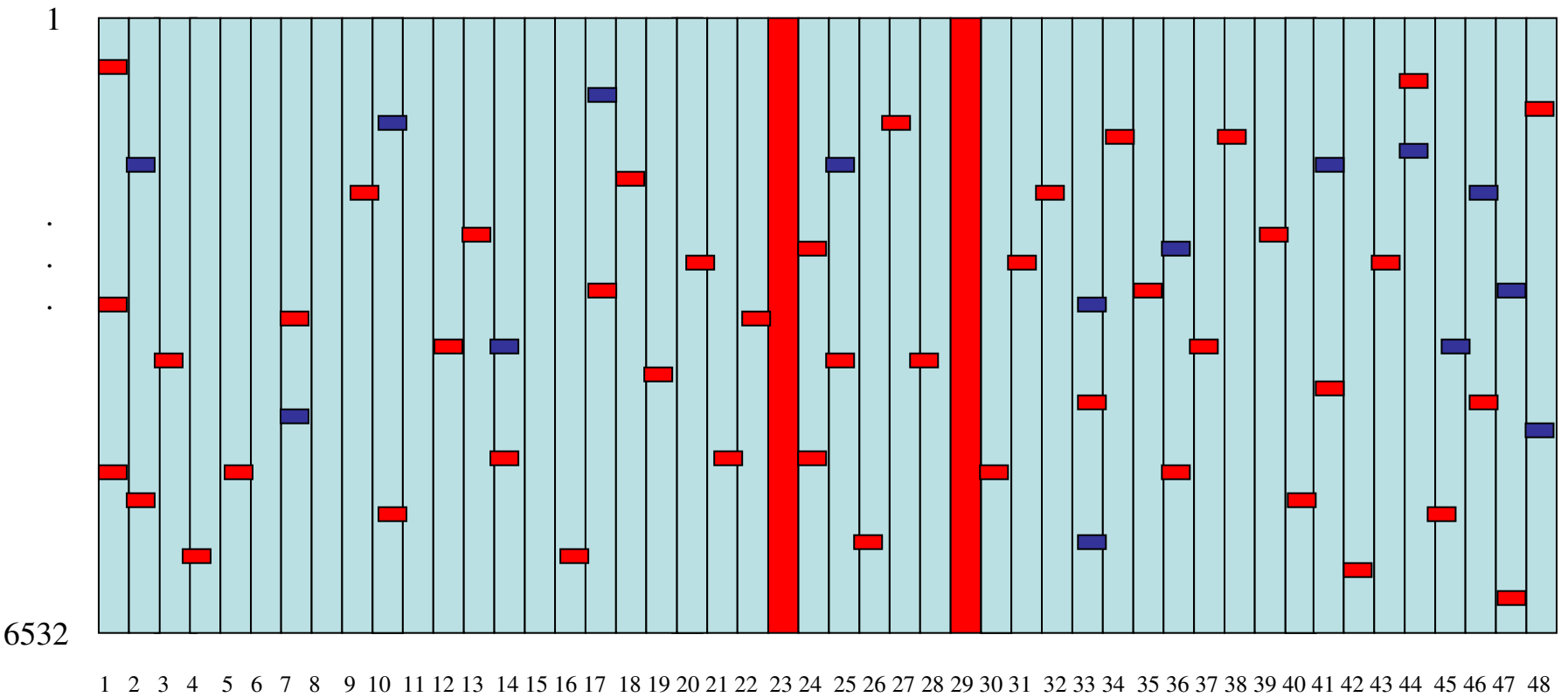
Metabolic Pathways

DeRisi´s transcriptome

Biological Interpretation

System architecture





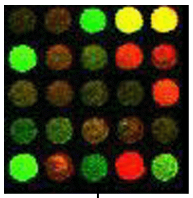
Good spots

Weak spots

Bad spots

NO INTERPOLATION

System architecture



USP dataset determination

USP dataset

Scaling and quantization

Quantized USP dataset

Functional groups

Overview set

GraphViz

Output graph

Plasmo DB

Metabolic Pathways

DeRisi´s transcriptome

Table of predictors

Target genes

Design of PGN

Biological Interpretation

Scaling

For each i, estimate the mean $\hat{E}[x_i[t]]$

and standard deviation $\hat{\sigma}[x_i[t]]$

of the spots

Scale normalization of the spots

$$n_i[t] = \frac{x_i[t] - \hat{E}[x_i[t]]}{\hat{\sigma}[x_i[t]]}$$

Quantization

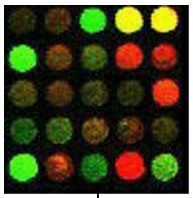
Let $n_i^+[t]$ and $n_i^-[t]$ denote, respectively, the normalized signals greater and lower than zero at t.

If $n_i^+[t] > \hat{E}[n_i^+[t]]$, then $x_i[t] = +1$

If $n_i^-[t] \geq \hat{E}[n_i^-[t]]$ and $n_i^+[t] \leq \hat{E}[n_i^+[t]]$, then $x_i[t] = 0$

If $n_i^-[t] < \hat{E}[n_i^-[t]]$, then $x_i[t] = -1$

System architecture



USP dataset determination

USP dataset

Scaling and quantization

Quantized USP dataset

Target genes

Design of PGN

Functional groups

Overview set

GraphViz

Table of predictors

Plasmo DB

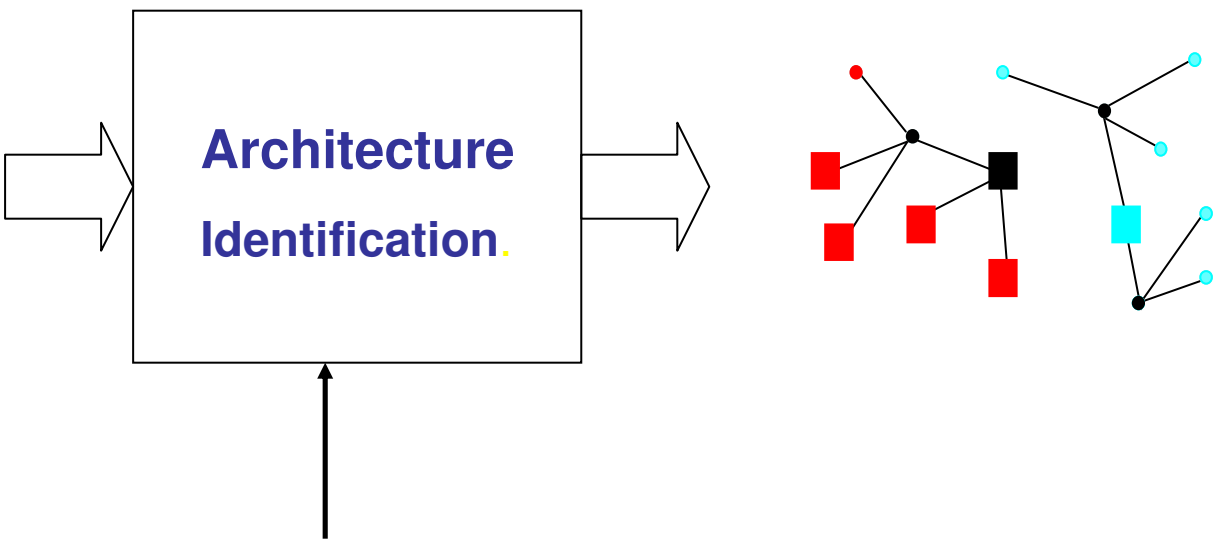
Metabolic Pathways

DeRisi´s transcriptome

Output graph

Biological Interpretation

$x[1], x[2], \ldots, x[48]$



Estimation of P(Y|X)

Y: the target gene at t+1, that is, $Y = x_i[t+1]$

X: the predictors at t, that is, $X = (x_j[t], x_k[t])$

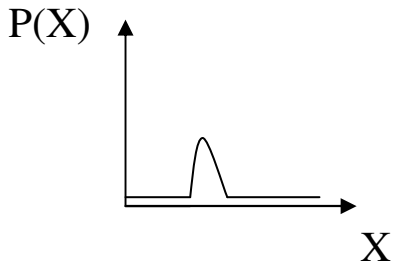
For a fixed parameter n

If #(X=(a,b)) $\geq$ n, then $\hat{P}(Y = c \mid X = (a,b)) = \frac{\#((Y = c) \wedge X = (a,b))}{\#(X = (a,b))}$

If #(X=(a,b)) $<$ n, then $\hat{P}(Y \mid X = (a,b))$ is uniform

Estimation of P(X) for a fixed parameter n

$$X = (x_j[t], x_k[t])$$



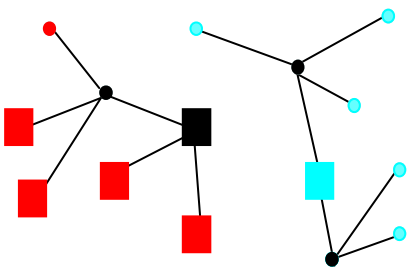
$$N^+ = \sum_{\#(X=(a,b))\geq n, \forall (a,b)} \#(X=(a,b)) \qquad N^- = \sum_{\#(X=(a,b))<n, \forall (a,b)} \#(X=(a,b))$$

If #(X=(a,b)) ≥ n, then $\hat{P}(X=(a,b)) = \frac{N^+}{N^- + N^+} \times \frac{\#(X=(a,b))}{N^+}$

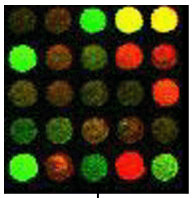
If #(X=(a,b)) < n, then $\hat{P}(X=(a,b)) = \frac{N^-}{N^- + N^+} \times \frac{1}{3^2 - |\{(a,b):\#(X=(a,b))\geq n\}|}$

Building Interaction Graphs

- For each target gene, rank the couples of all genes by their estimated mutual information and sample size;
- When two mutual information are equal, the one estimated from a larger sample comes first;
- Choose the best couples;
- Design the interaction graph



System architecture



USP dataset determination

USP dataset

Scaling and quantization

Quantized USP dataset

Target genes

Design of PGN

Functional groups

Overview set

GraphViz

Table of predictors

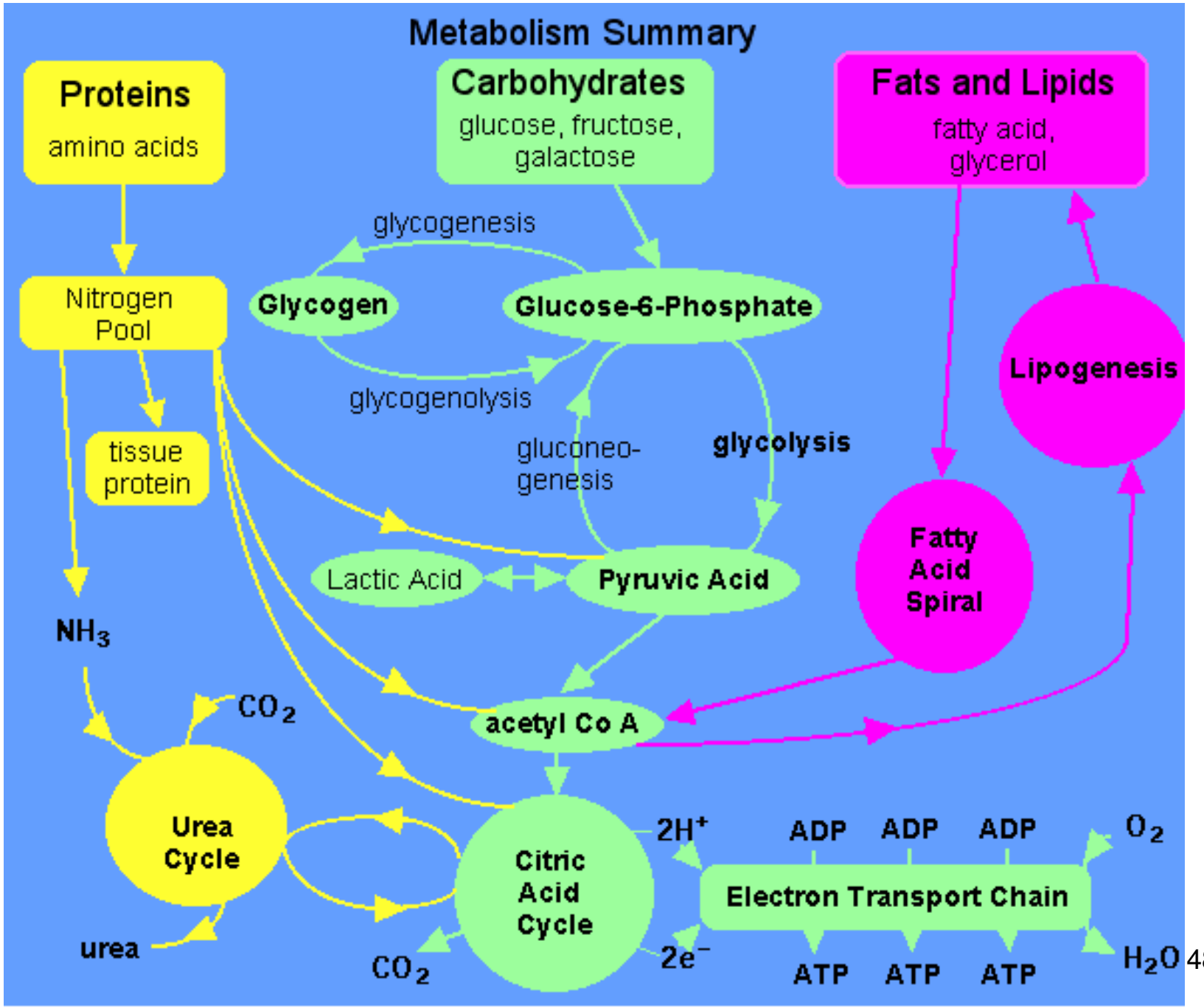
Output graph

Plasmo DB

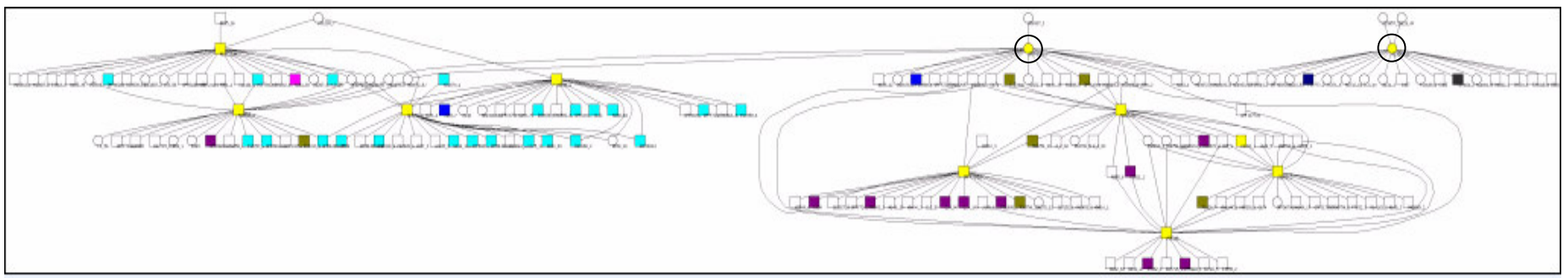
Metabolic Pathways

DeRisi´s transcriptome

Biological Interpretation


Metabolism Summary
Proteins
amino acids
Carbohydrates
glucose, fructose, galactose
Fats and Lipids
fatty acid, glycerol
glycogenesis
Nitrogen Pool
Glycogen
Glucose-6-Phosphate
Lipogenesis
glycogenolysis
tissue protein
gluconeo-genesis
glycolysis
Fatty Acid Spiral
Lactic Acid
Pyruvic Acid
NH3
CO2
acetyl Co A
Urea Cycle
Citric Acid Cycle
2H+
ADP ADP ADP
O2
Electron Transport Chain
urea
CO2
2e−
ATP ATP ATP
H2O

Glycolytic PGN network (single genes)



glycolysis

transcription machinery

cytoplasmic translation

ribonucleotide synthesis

deoxynucleotide synthesis

DNA replication

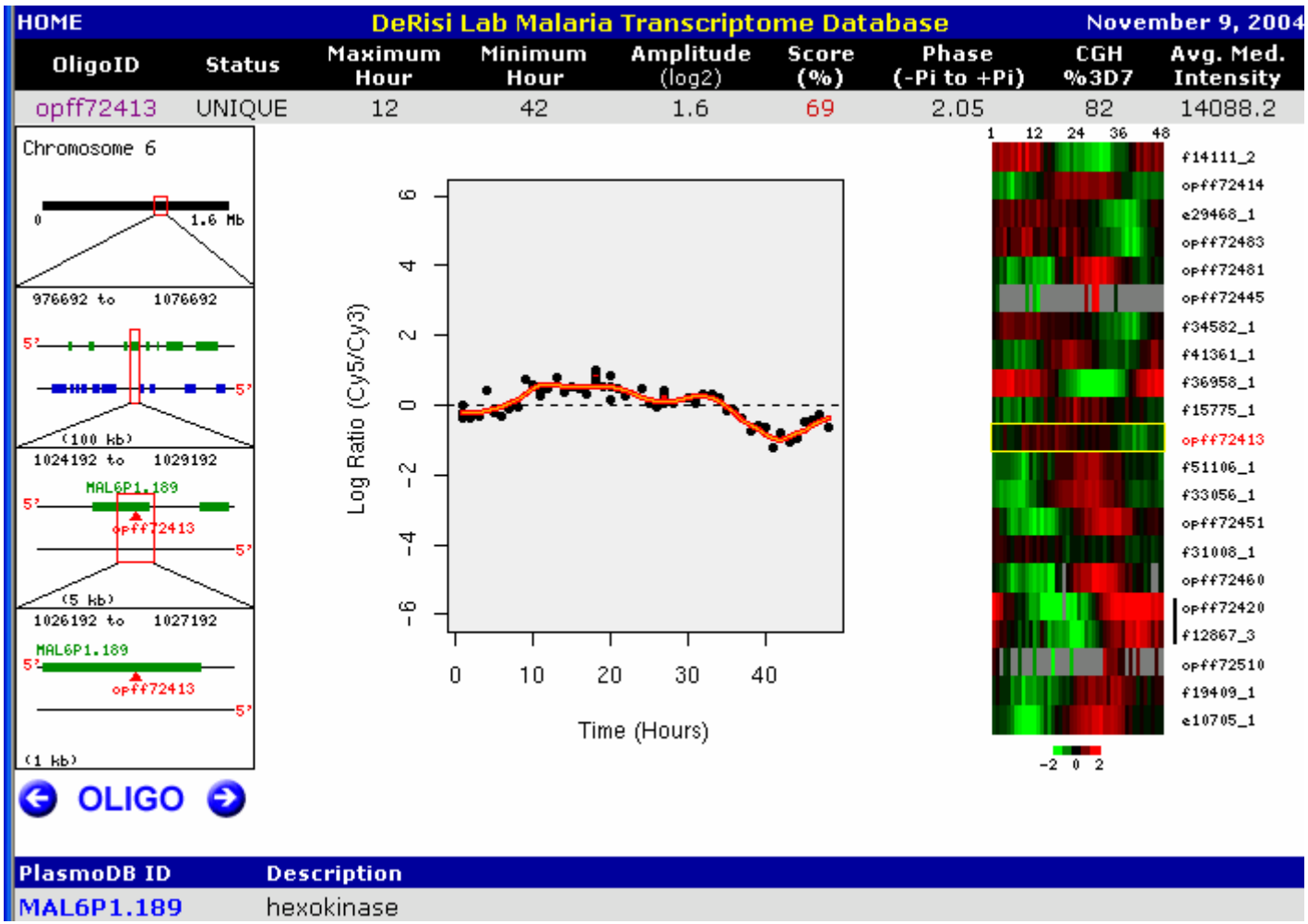
proteoasome

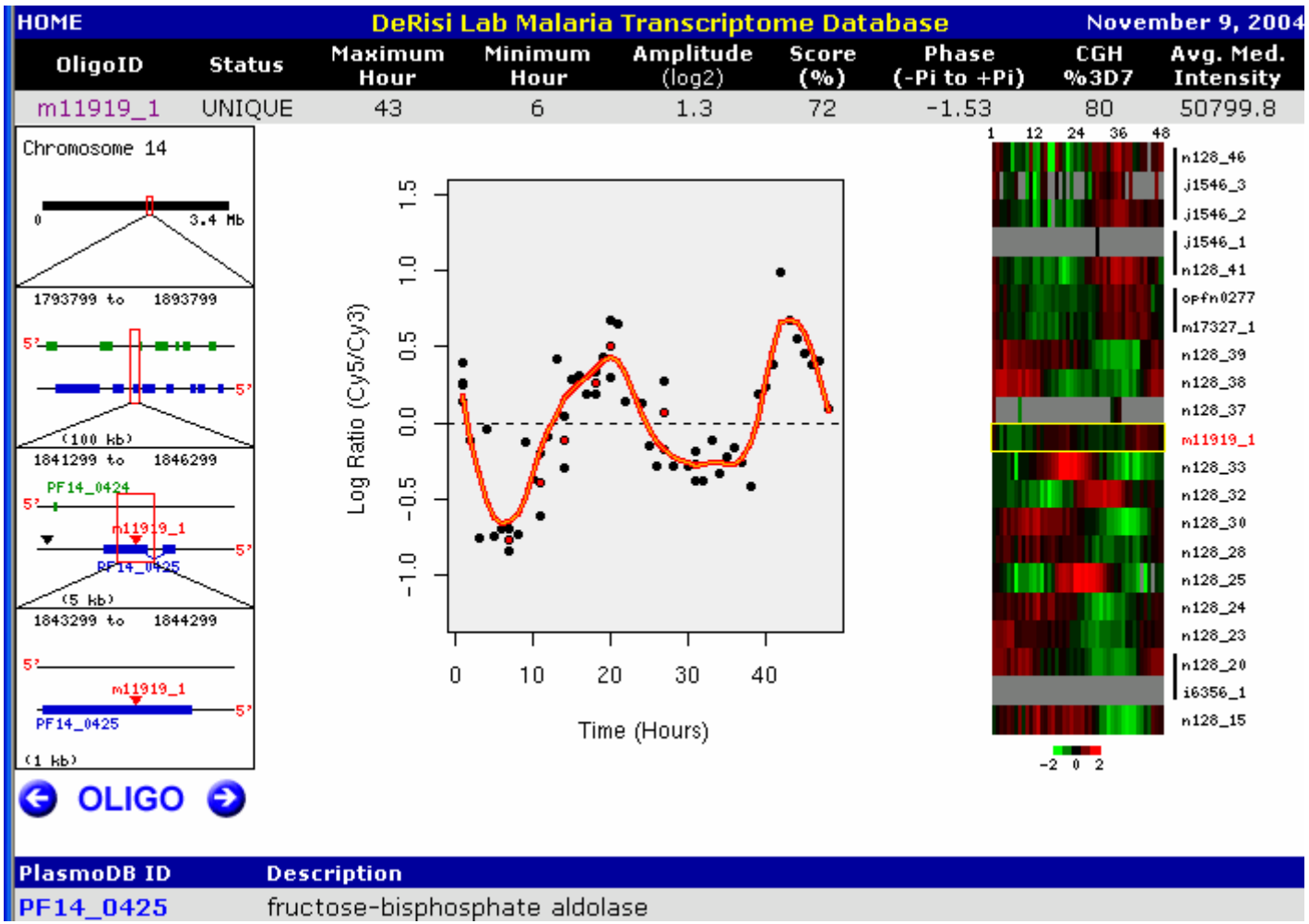
plastid genome

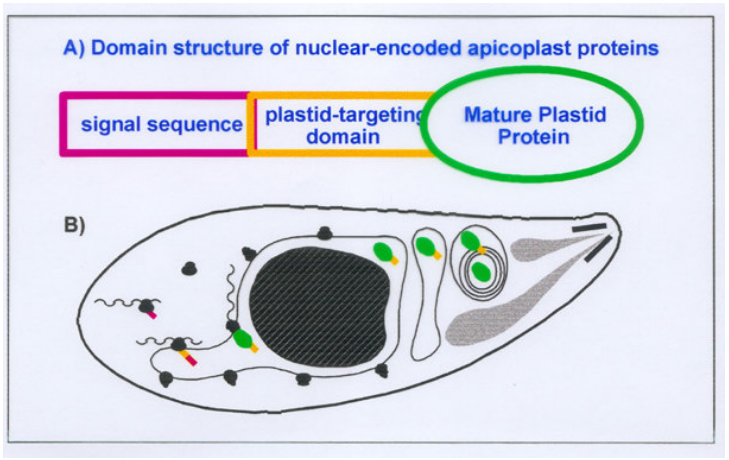
merozoite invasion (kinases)

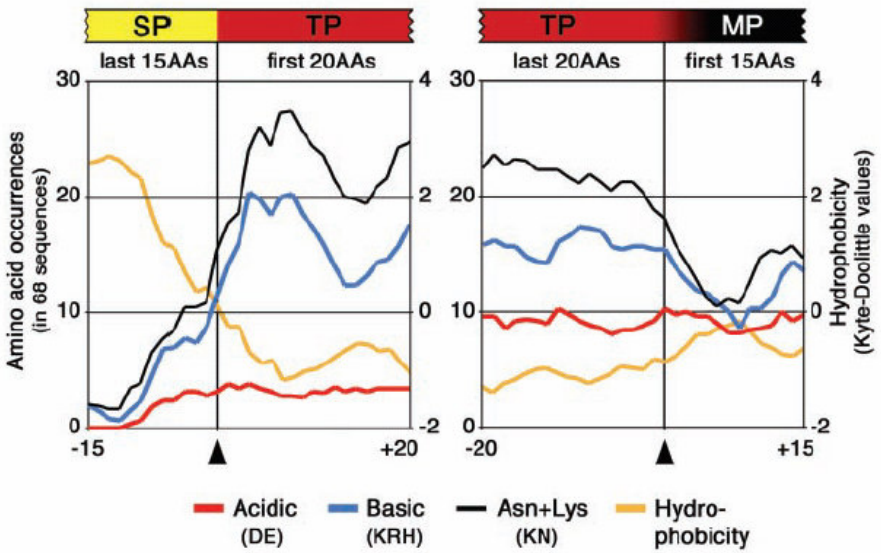
actin myosoin motors

early ring transcripts

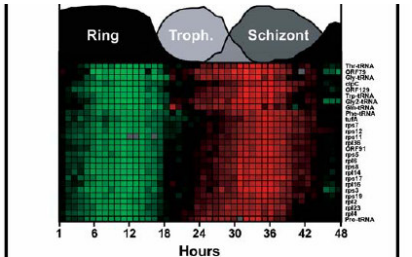

HOME
DeRisi Lab Malaria Transcriptome Database
November 9, 2004
OligoID
Status
Maximum Hour
Minimum Hour
Amplitude (log2)
Score (%)
Phase (-Pi to +Pi)
CGH %3D7
Avg. Med. Intensity
opff72413
UNIQUE
12
42
1.6
69
2.05
82
14088.2
Chromosome 6
0
1.6 Mb
976692 to 1076692
5'
5'
(100 kb)
1024192 to 1029192
MAL6P1.189
5'
opff72413
5'
(5 kb)
1026192 to 1027192
MAL6P1.189
5'
opff72413
5'
(1 kb)
OLIGO
Log Ratio (Cy5/Cy3)
6
4
2
0
-2
-4
-6
0
10
20
30
40
Time (Hours)
1
12
24
36
48
f14111_2
opff72414
e29468_1
opff72483
opff72481
opff72445
f34582_1
f41361_1
f36958_1
f15775_1
opff72413
f51106_1
f33056_1
opff72451
f31008_1
opff72460
opff72420
f12867_3
opff72510
f19409_1
e10705_1
-2 0 2
PlasmoDB ID
Description
MAL6P1.189
hexokinase


HOME
DeRisi Lab Malaria Transcriptome Database
November 9, 2004
OligoID
Status
Maximum Hour
Minimum Hour
Amplitude (log2)
Score (%)
Phase (-Pi to +Pi)
CGH %3D7
Avg. Med. Intensity
m11919_1
UNIQUE
43
6
1.3
72
-1.53
80
50799.8
Chromosome 14
0
3.4 Mb
1793799 to 1893799
(100 kb)
1841299 to 1846299
PF14_0424
m11919_1
PF14_0425
(5 kb)
1843299 to 1844299
m11919_1
PF14_0425
(1 kb)
OLIGO
Log Ratio (Cy5/Cy3)
Time (Hours)
PlasmoDB ID
Description
PF14_0425
fructose-bisphosphate aldolase

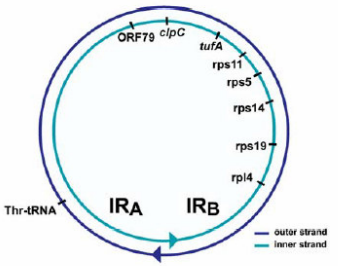


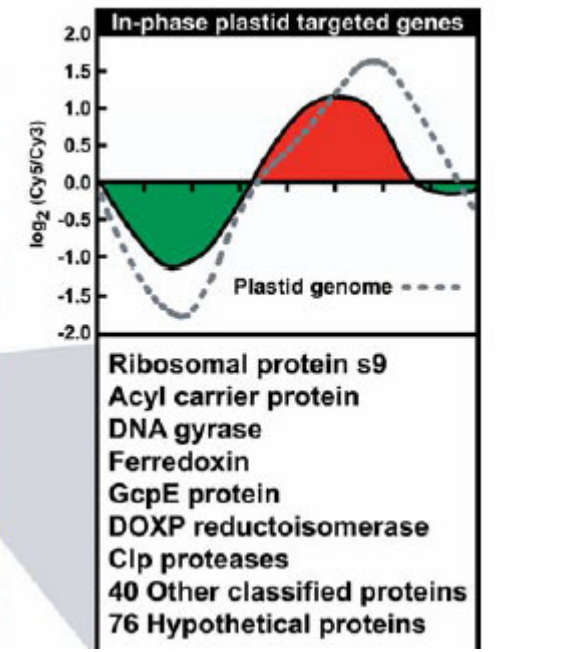


550 apicoplast proteins



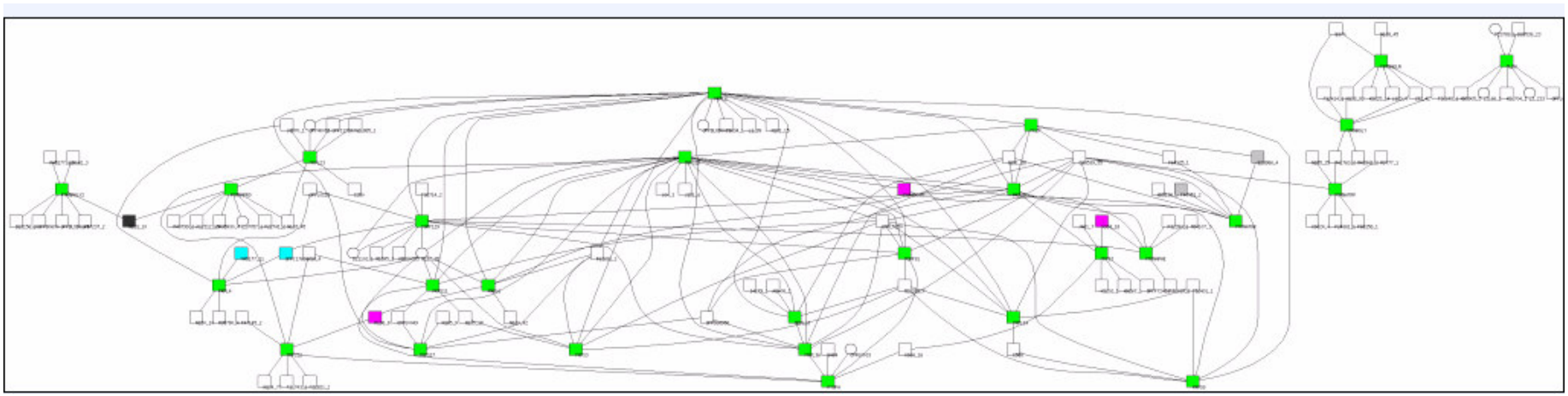
P. falciparum **plDNA (35 kb)**





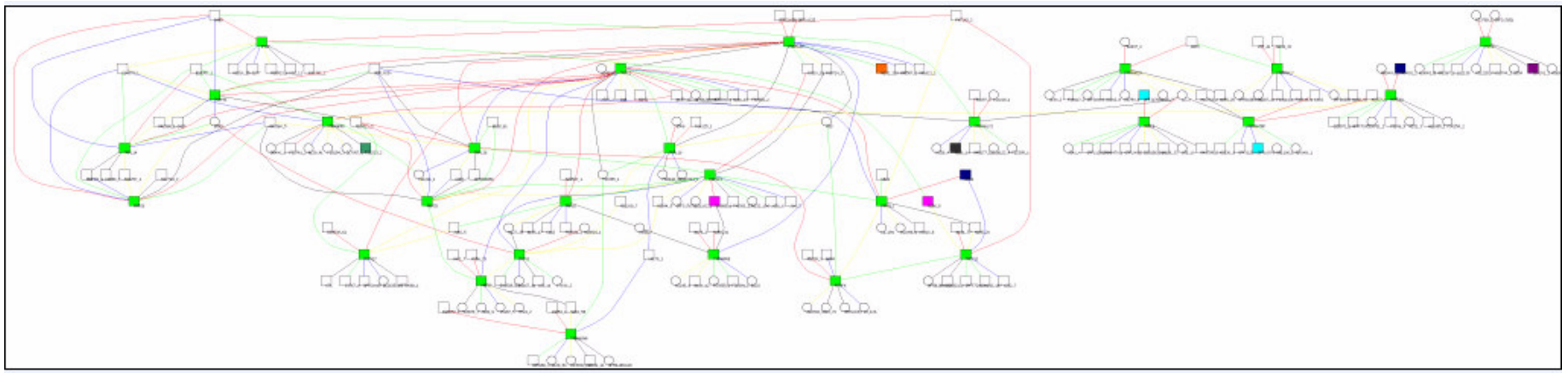
124 apicoplast proteins

Apicoplast PGN network (single genes)



- glycolysis
- transcription machinery
- cytoplasmic translation
- ribonucleotide synthesis
- deoxynucleotide synthesis
- DNA replication
- proteoasome
- plastid genome
- merozoite invasion
- actin myosoin motors
- mitochondrial

Apicoplast PGN network (double genes)



glycolysis

transcription machinery

cytoplasmic translation

ribonucleotide synthesis

deoxynucleotide synthesis

DNA replication

proteoasome

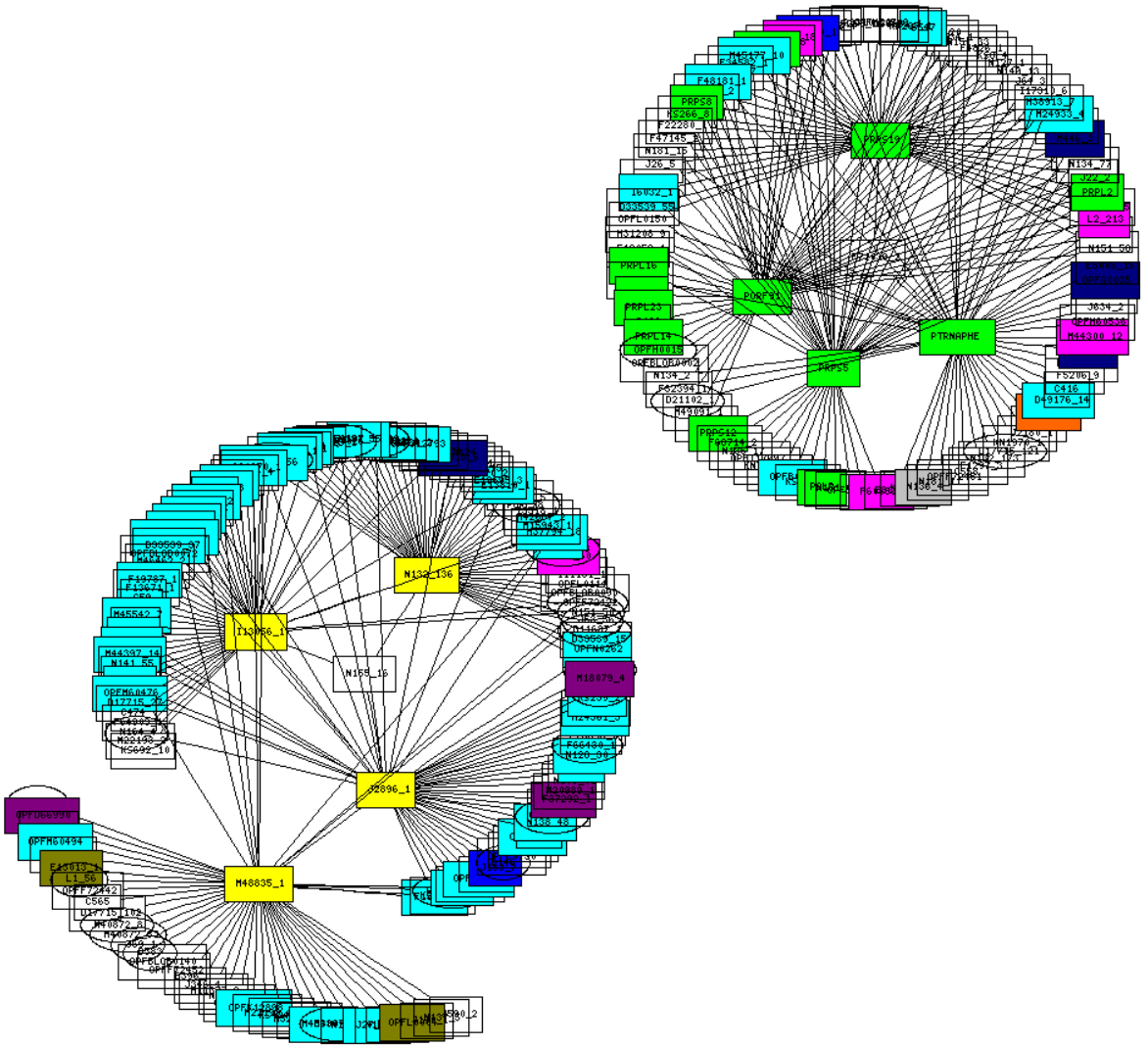
plastid genome

merozoite invasion

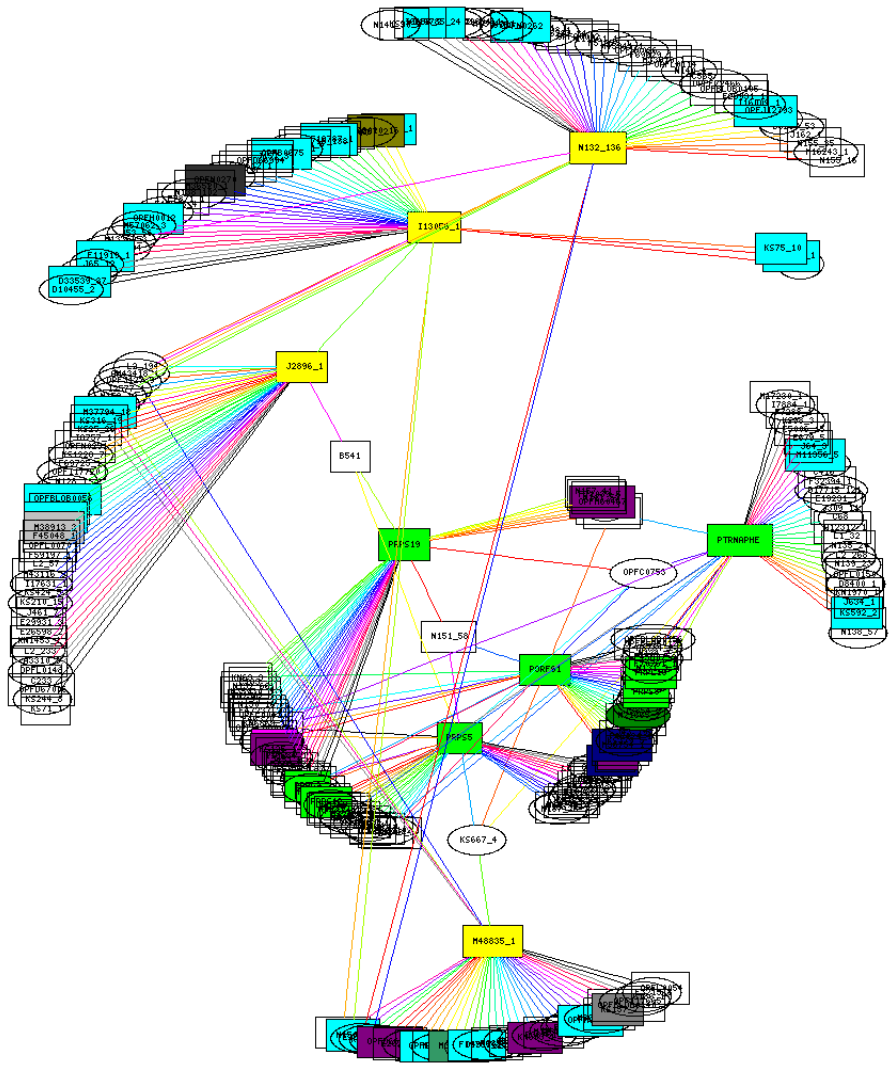
actin myosoin motors

early ring transcripts

Subsystems identification



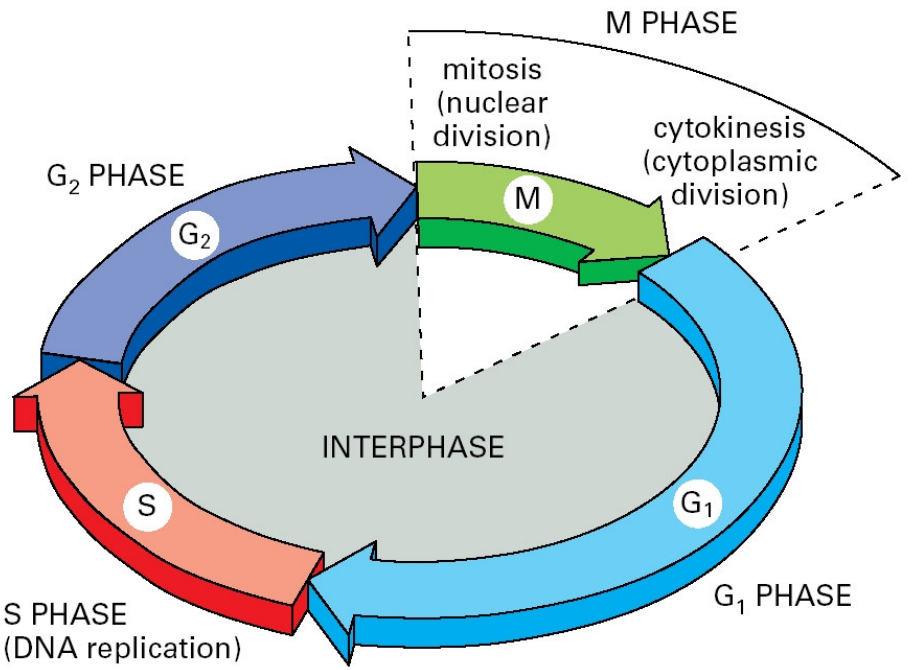
Subsystems identification



Layout

- Introduction
- Probabilistic Genetic Networks (PGN)
- Estimation of PGNs
- Architecture estimation
- Malaria
- Cell Cycle
- Future works

Phases of the Cell Cycle



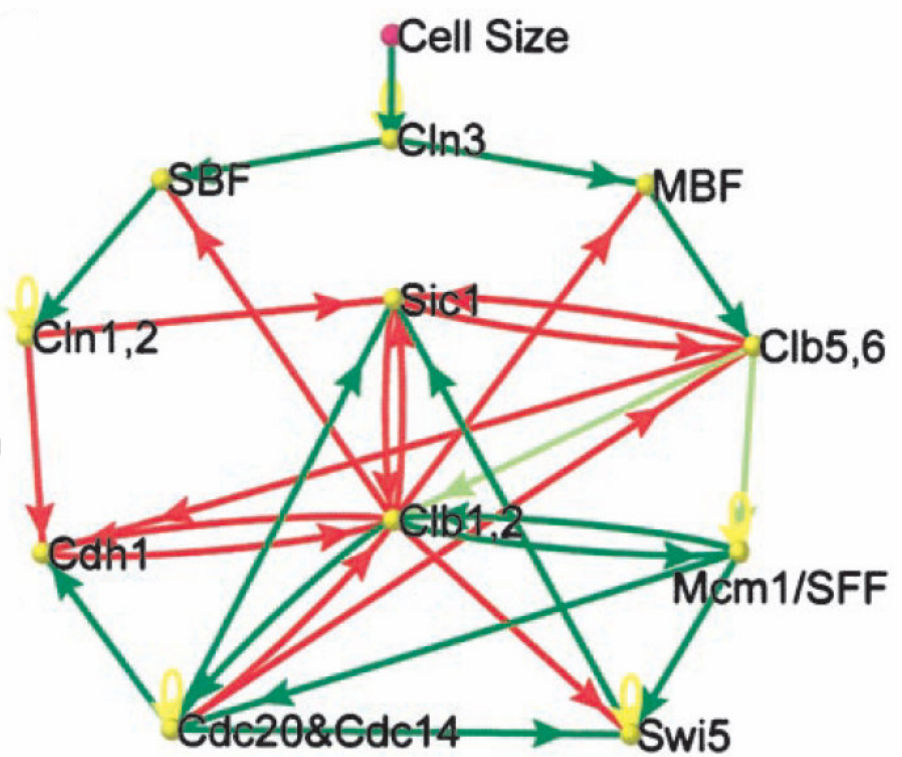
(*)

Model: architecture and dynamics

Li, 2004.

Transition Function

$$S_i(t+1) = \begin{cases} 1, & \sum_j a_{ij} S_j(t) > 0 \\ 0, & \sum_j a_{ij} S_j(t) < 0 \\ S_i(t), & \sum_j a_{ij} S_j(t) = 0 \end{cases}$$

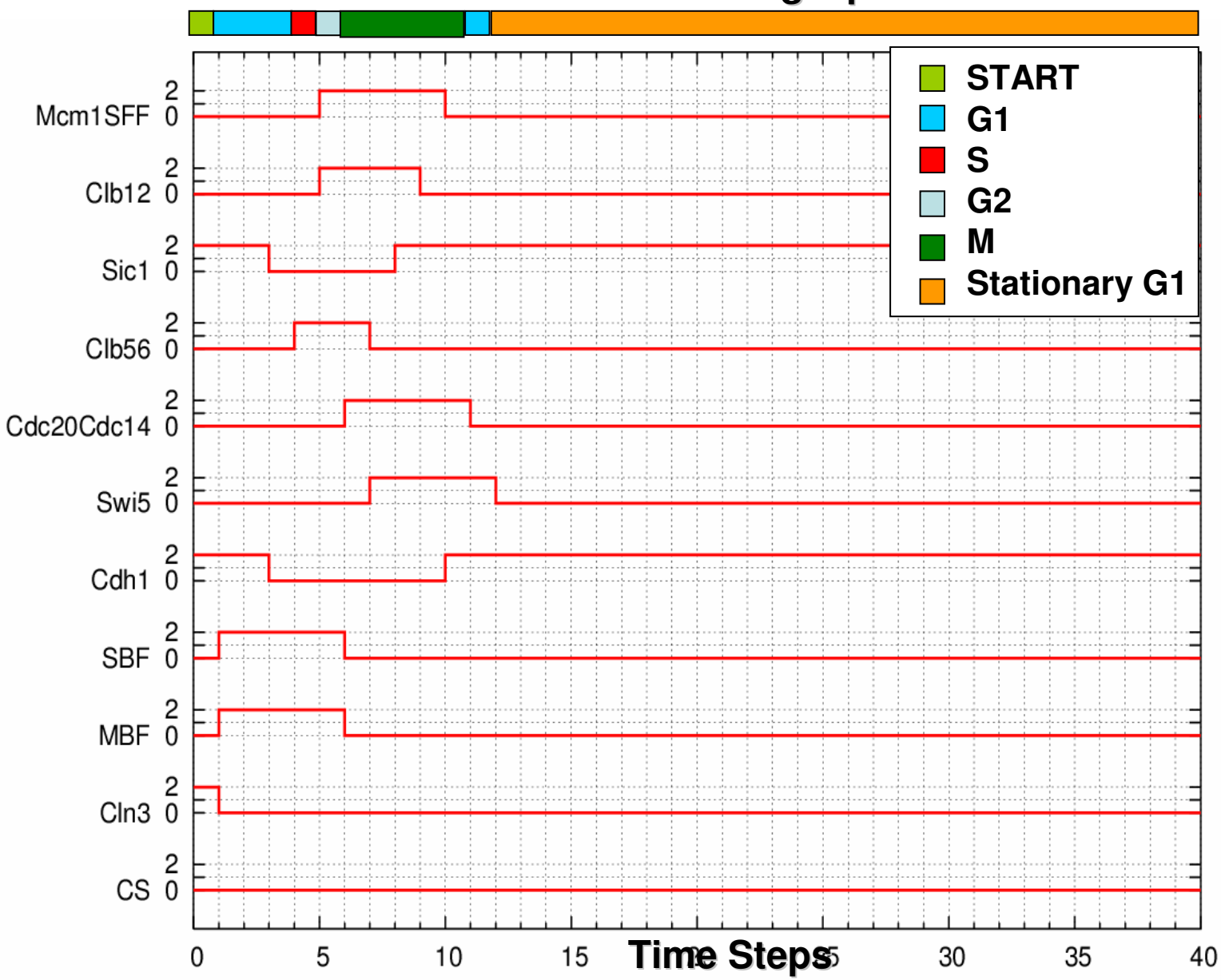


$a_{ij} = a_g$ **green arrow from *i* to *j***

$a_{ij} = a_r$ **red arrow from *i* to *j***

Simulation parameters: $a_g = -a_r = 1$

Deterministic



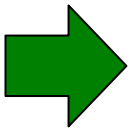
Binary

1 → 2

0 → 0

3 Levels (0, 1, 2)

$\sum_j a_{ij} S_j(0)$	$S_j(t+1)$	
	$S_j(t)=0$	$S_j(t)=1$
⋮	⋮	⋮
3	1	1
2	1	1
1	1	1
0	0	1
-1	0	0
-2	0	0
-3	0	0
⋮	⋮	⋮



$\sum_j a_{ij} S_j(0)$	$y_j(t+1)$		
	$S_j(t)=0$	$S_j(t)=1$	$S_j(t)=2$
⋮	⋮	⋮	⋮
3	2	2	2
2	2	2	2
1	1	2	2
0	0	1	2
-1	0	0	1
-2	0	0	0
-3	0	0	0
⋮	⋮	⋮	⋮

Stochastic Transition Function

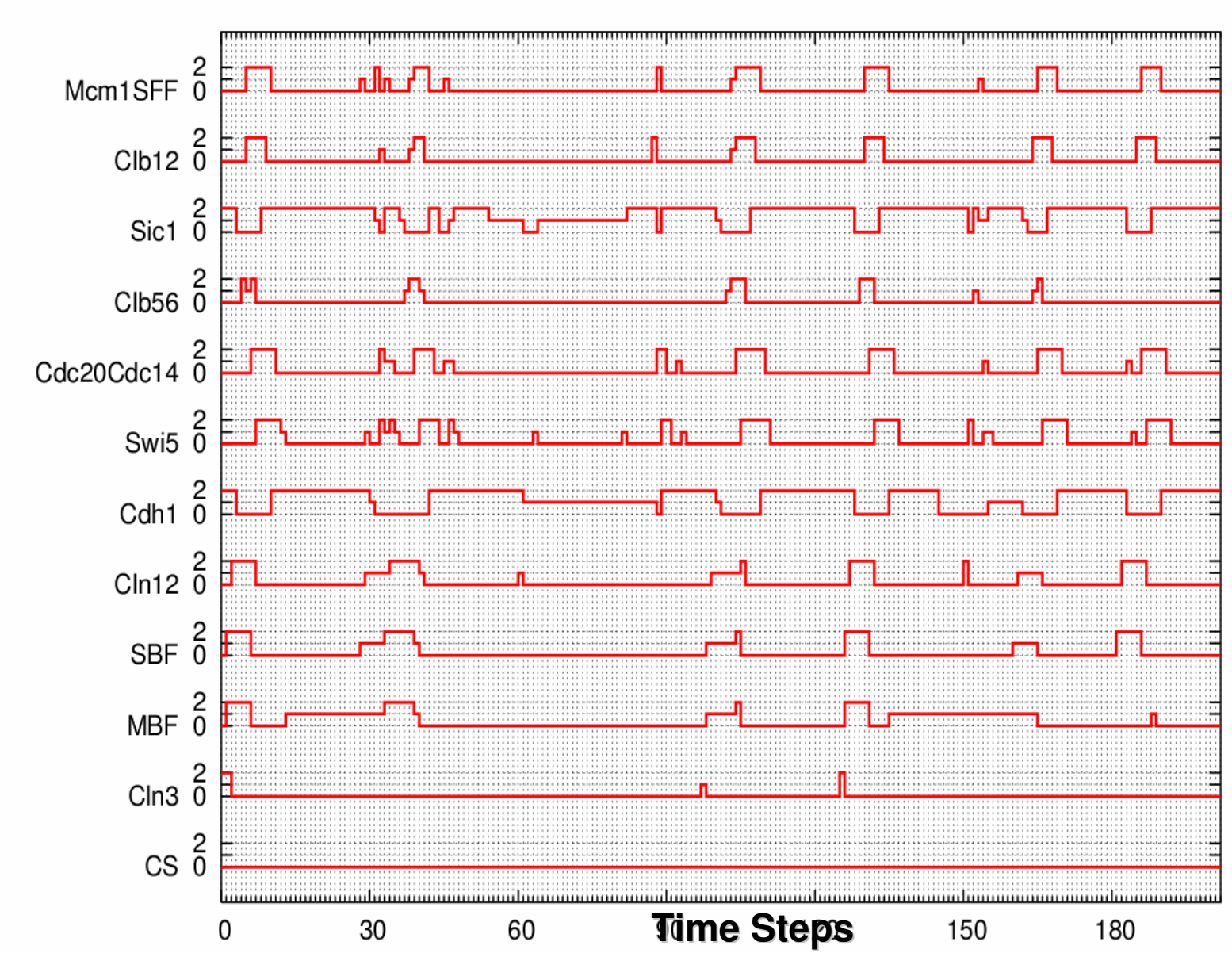
$\sum_j a_{ij} S_j(0)$	$y_j(t+1)$		
	$S_j(t)=0$	$S_j(t)=1$	$S_j(t)=2$
⋮	⋮	⋮	⋮
3	2	2	2
2	2	2	2
1	1	2	2
0	0	1	2
-1	0	0	1
-2	0	0	0
-3	0	0	0
⋮	⋮	⋮	⋮

$$x_i(t+1) = \begin{cases} y_i(t+1) & \text{with} \quad P = 0.99 \\ a & \text{with} \quad P = 0.005 \\ b & \text{with} \quad P = 0.005 \end{cases}$$

where $a, b \in \{0,1,2\} - \{y_i(t+1)\}$

PGN with $P = 0.99$

One single pulse of $CS = 2$ at $t = -1$



Our gene model

Total input signal driving a generic variable

$$x_i(t) \in \{0, 1, 2\} \quad (1 \leq i \leq N)$$

$$d_i(t) = \sum_{j=1}^{N} \sum_{k=1}^{m} a_{ji}^{k} x_j(t-k)$$

Driving function for x_i

m**: memory of the system**

a_{ji}^{k}**: weight for variable** x_j **at time** $t-k$

Our gene model

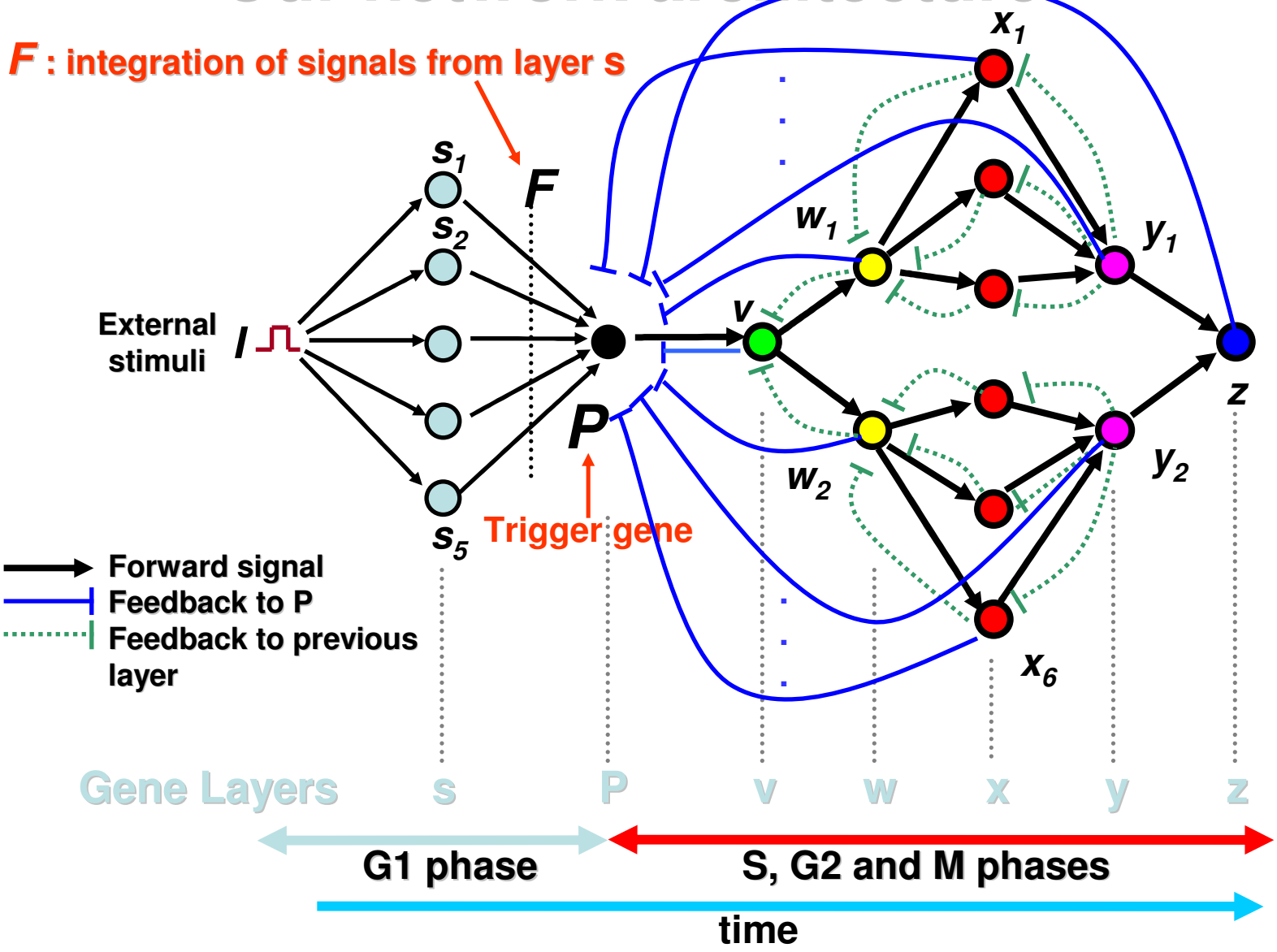
$$y_i(t+1) = \begin{cases} 2 & if \quad d_i(t) \geq th_{i2} \\ 1 & if \quad th_{i1} \leq d_i(t) \leq th_{i2} \\ 0 & if \quad d_i(t) < th_{i1} \end{cases}$$

Stochastic Transition Function

$$x_i(t) = \begin{cases} y_i(t) & \text{with} \quad P \approx 1 \\ a & \text{with} \quad (1-P)/2 \\ b & \text{with} \quad (1-P)/2. \end{cases}$$

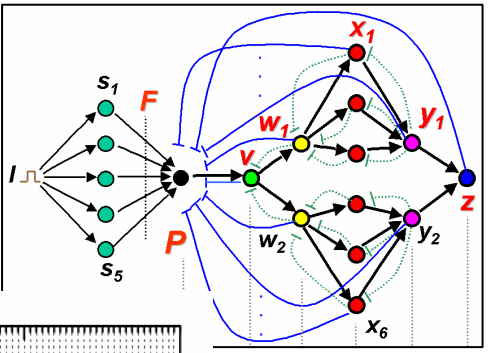
$$a, b \in \{0, 1, 2\} - \{y_i(t)\}$$

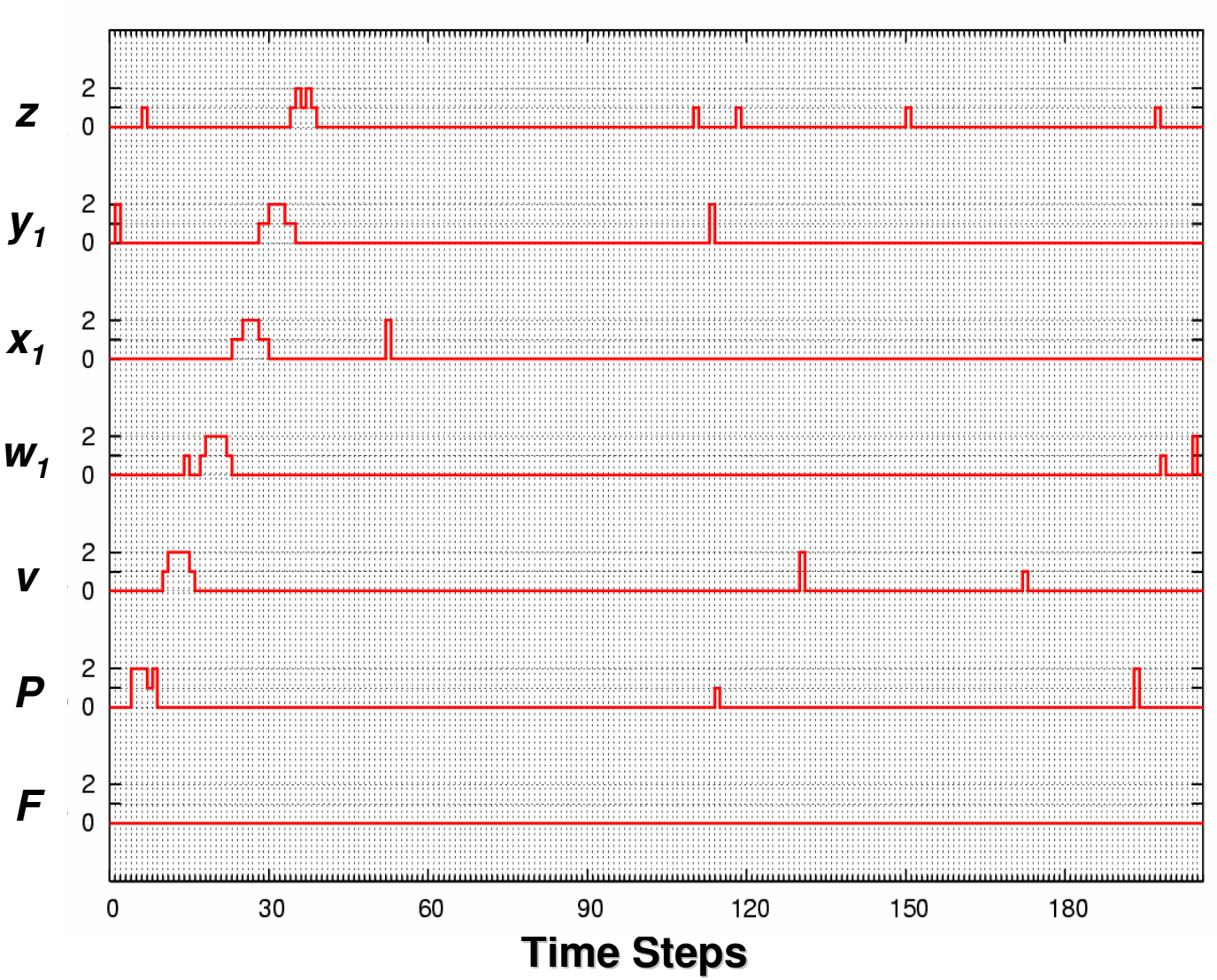
Our network architecture



PGN with $P = 0.99$

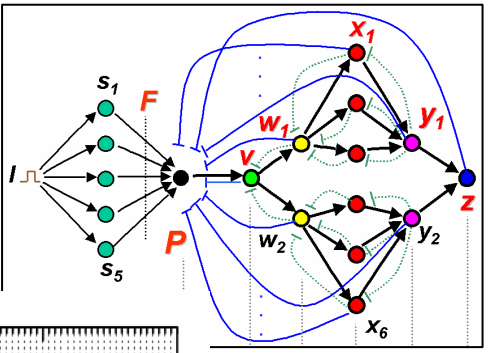
One single pulse of $F = 2$ at $t = -1$

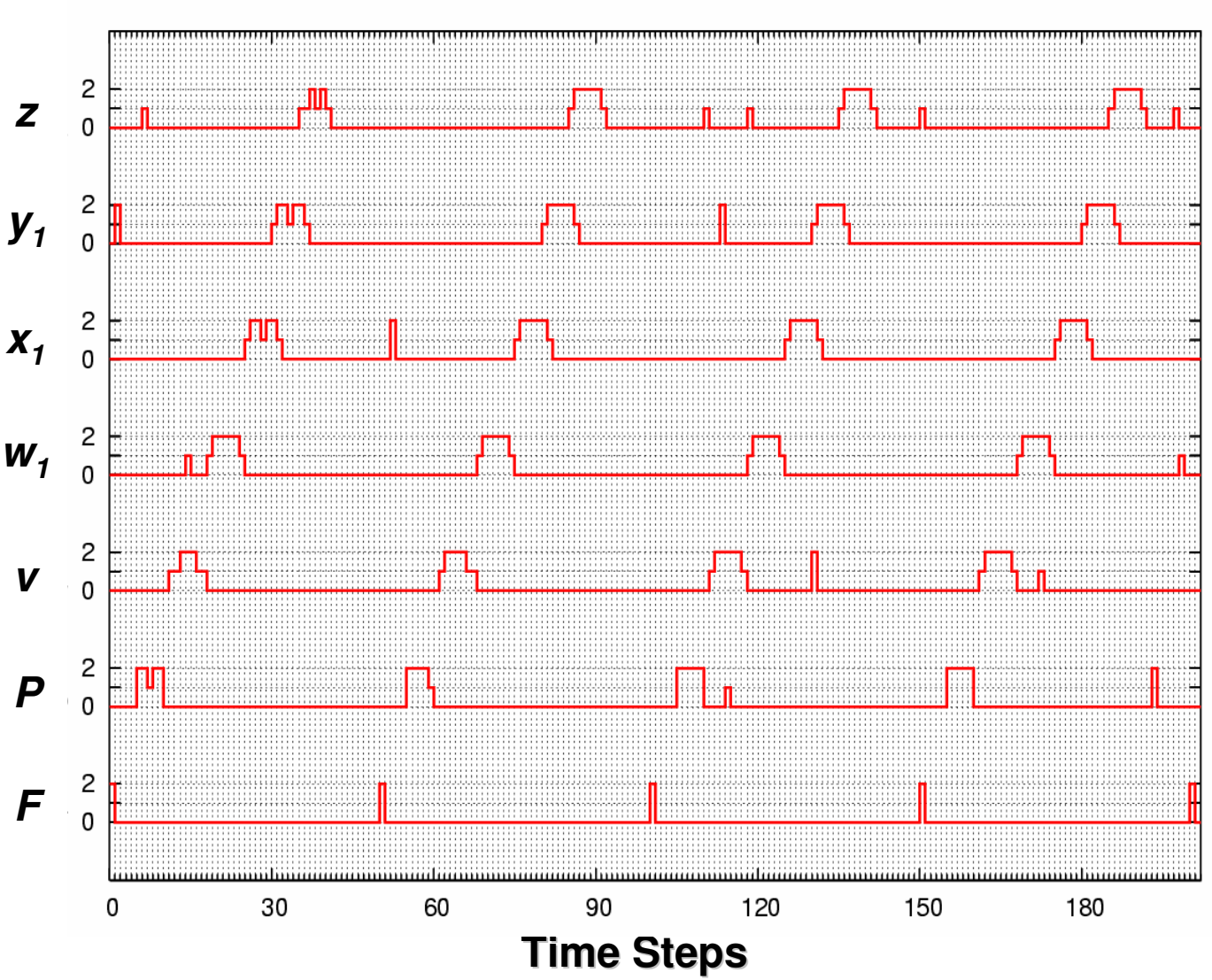




PGN with $P = 0.99$

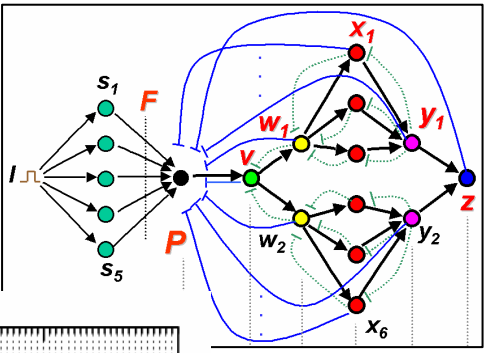
Signal F = period 50 oscillator

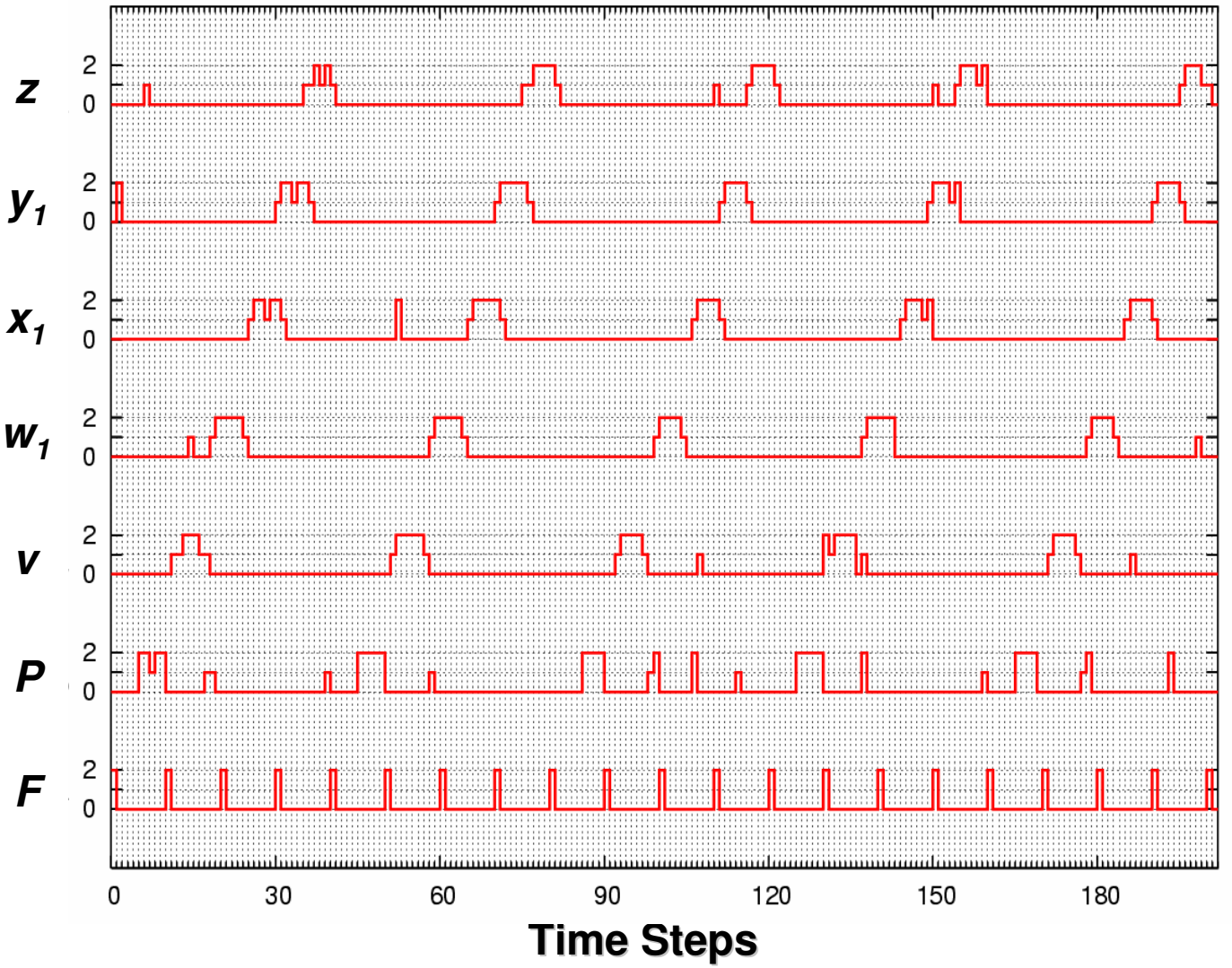




PGN with $P = 0.99$

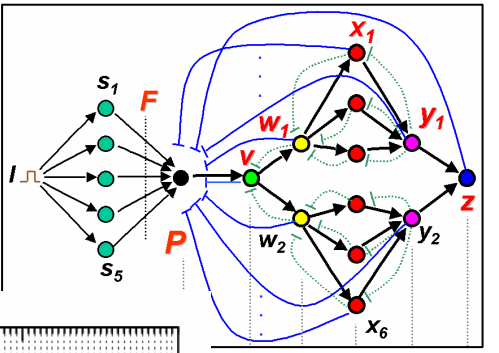
Signal F = period 10 oscillator

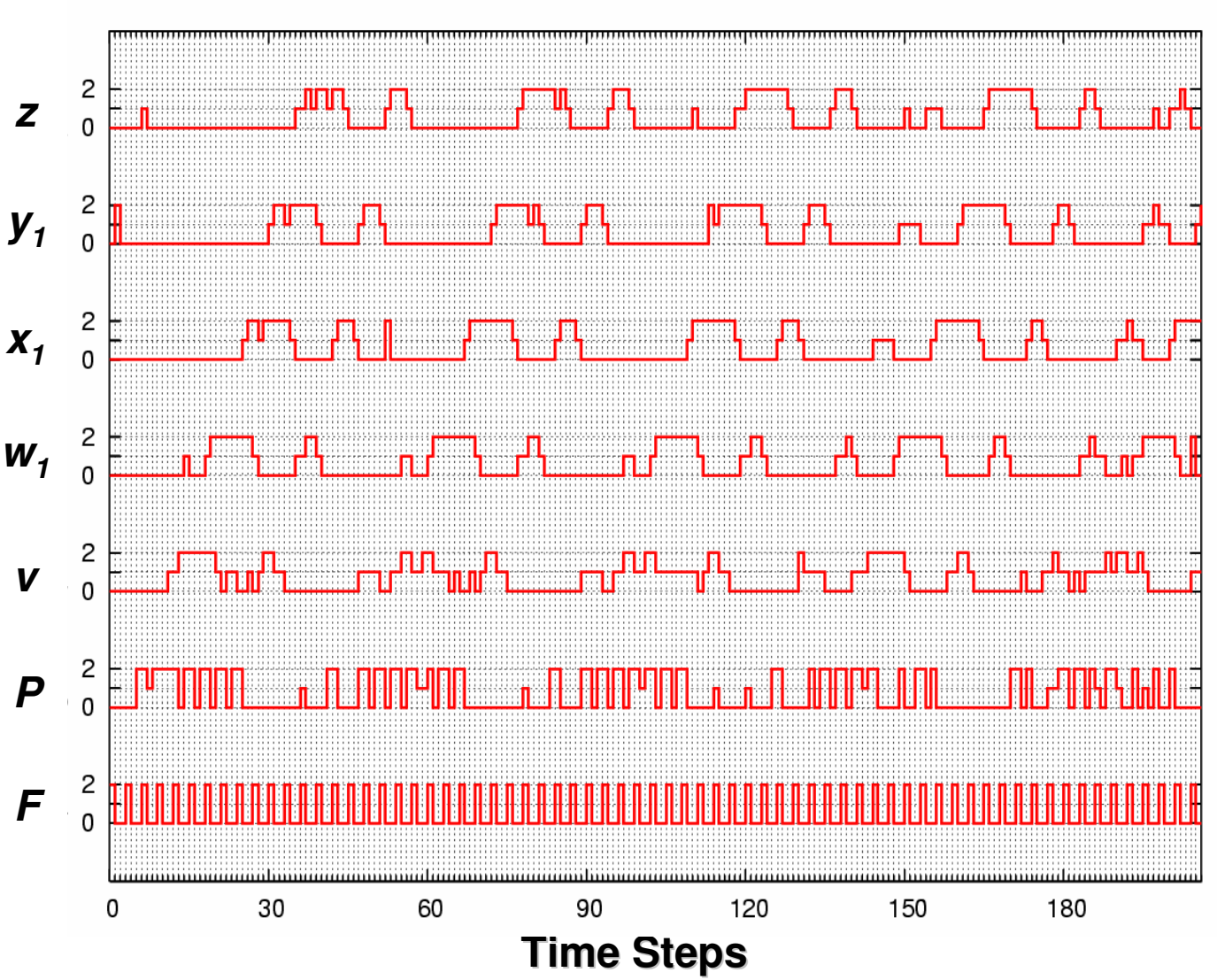




PGN with $P = 0.99$

Signal F = period 3 oscillator





Layout

- Introduction
- Probabilistic Genetic Networks (PGN)
- Estimation of PGNs
- Architecture estimation
- Malaria
- Cell Cycle
- Future works

- Estimating cell cycle network from data

  Use the developed model as constraint

- Add new constraints

  Representing phenomena like stability, robustness, protein interactions

- Create dynamical criteria for network estimation

  A kind of measure on the time sequence distribution

- Design dedicated search algorithms for particular partitions spaces

  Different partitions may have common parts that does not need to be calculated again.

- Design smart search algorithms for feature selection

  These algorithms should learn while walking through the search space