A NEW ANNOTATION TOOL FOR MALARIA BASED ON INFERENCE OF PROBABILISTIC GENETIC NETWORKS

J. Barrera[1,] R. M. Cesar Jr. [1], D. C. Martins Jr.[1],
E. F Merino[2] , R. Z. N. Vêncio[1] , F. G. Leonardi[1,]
M. M. Yamamoto[2], C. A. B. Pereira[1], H. A. del Portillo[2]

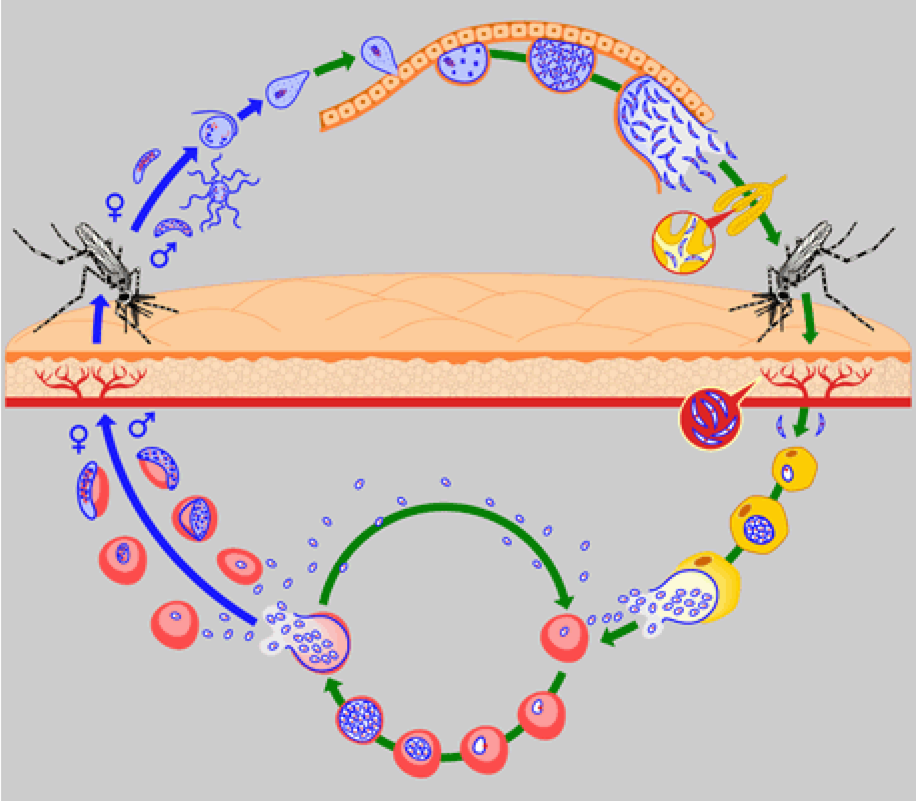
1- IME-USP; 2- ICB-USP

Layout

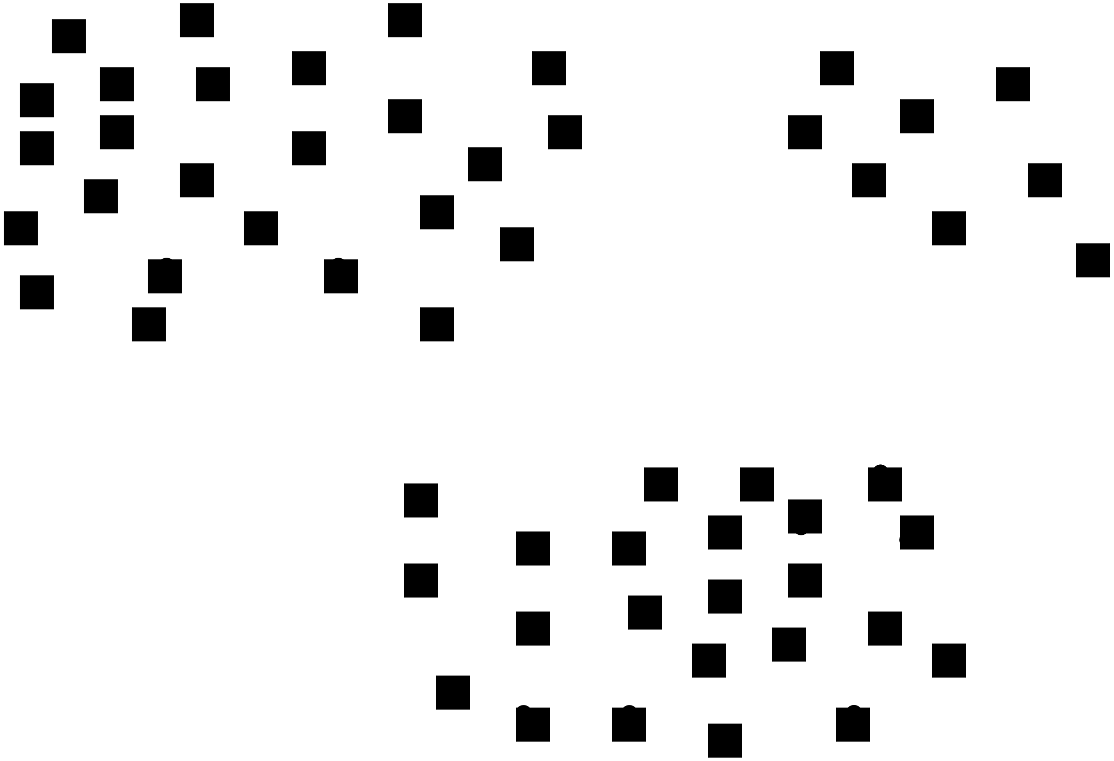
- Introduction
- Probabilistic genetic network (PGN)
- PGN design
- Data analysis pipeline
- Biological interpretation
- Future steps

Introduction

The life cycle of the malaria parasite

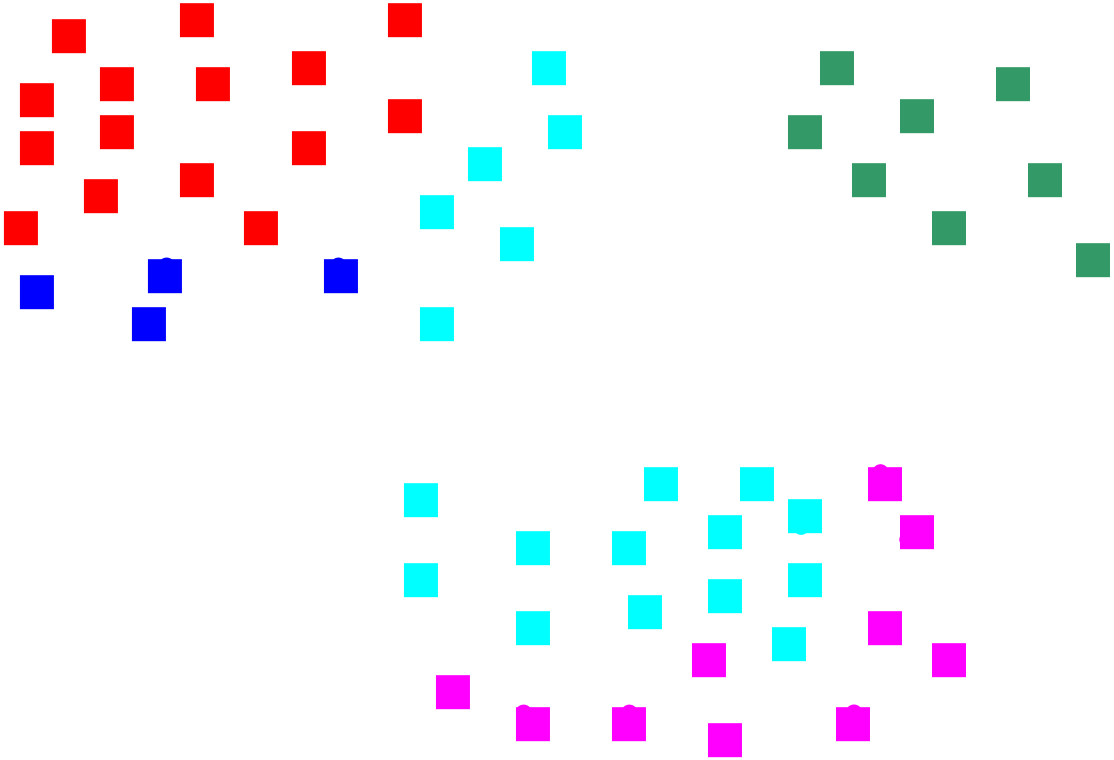


Malaria parasite genes with almost sinusoidal signals



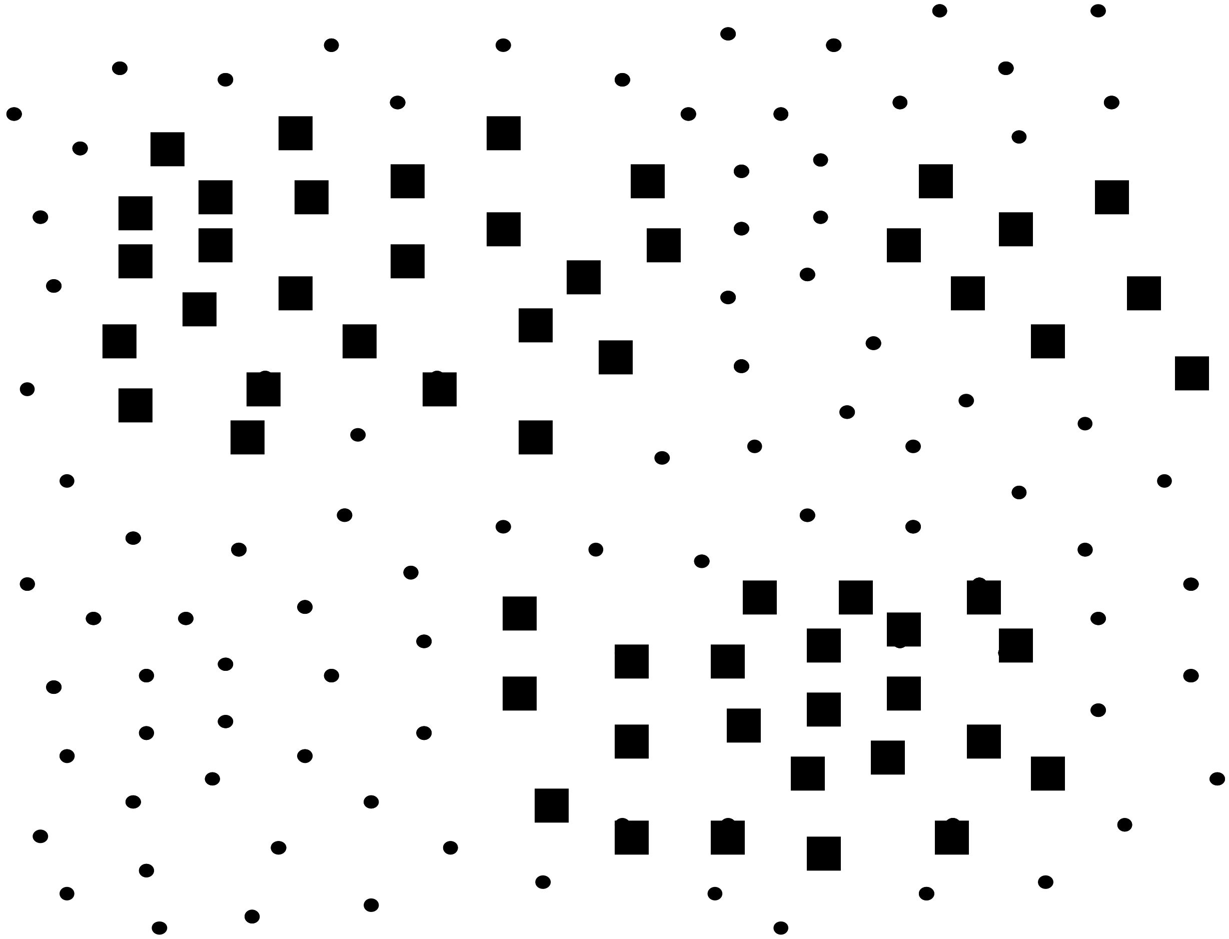
DeRisi, 2003.

Functional Classification

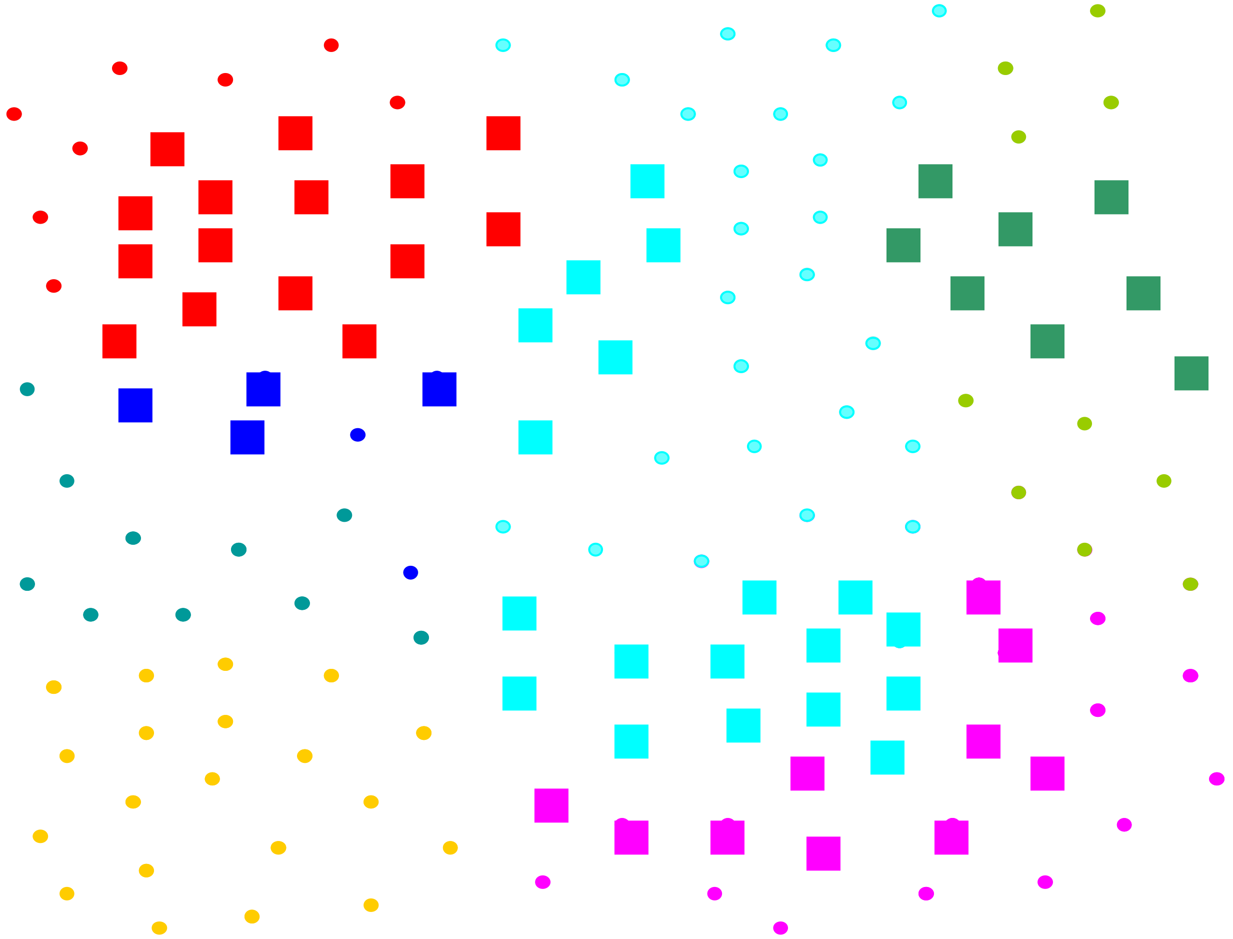


DeRisi, 2003.

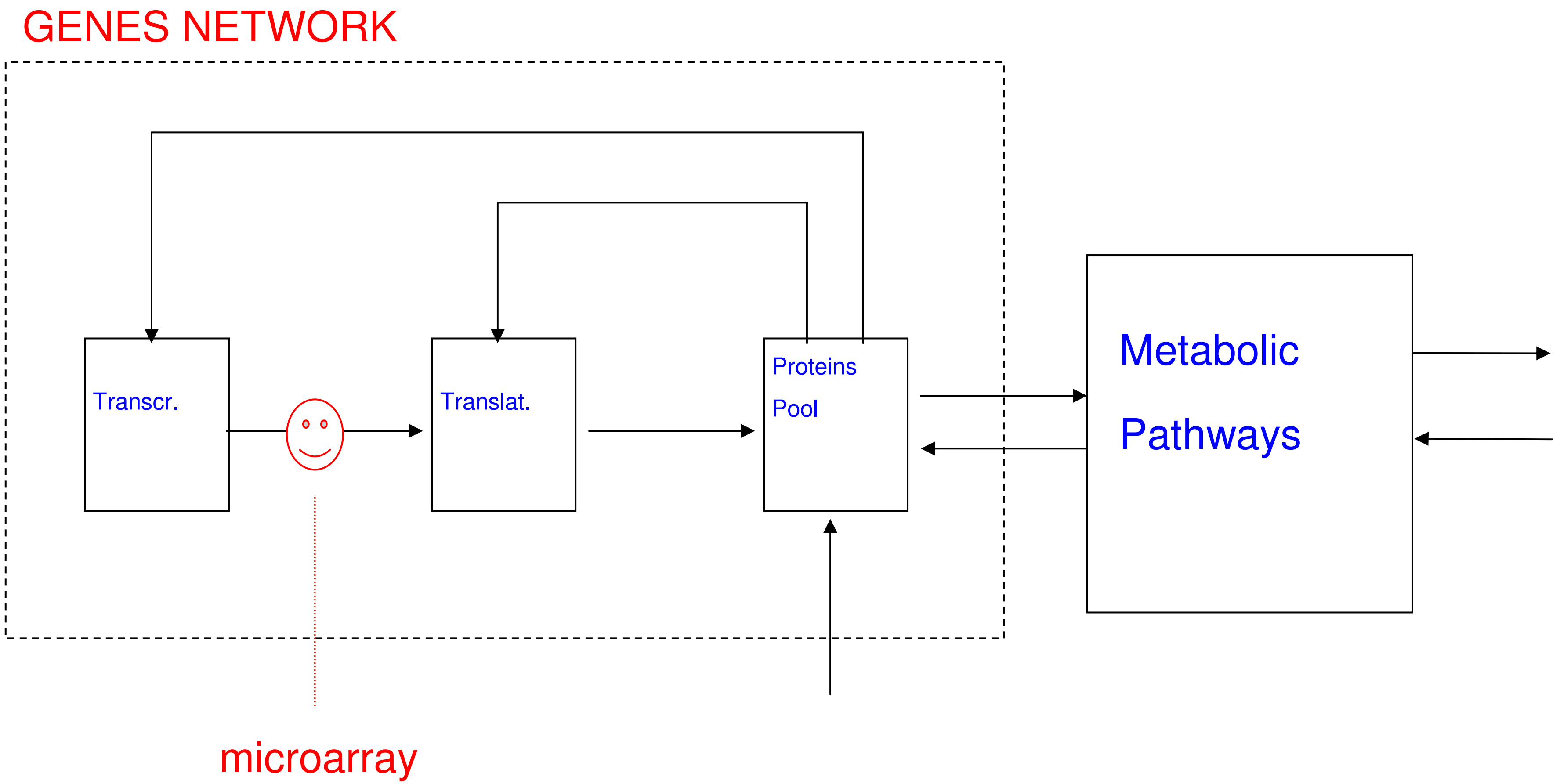
Malaria parasite genes



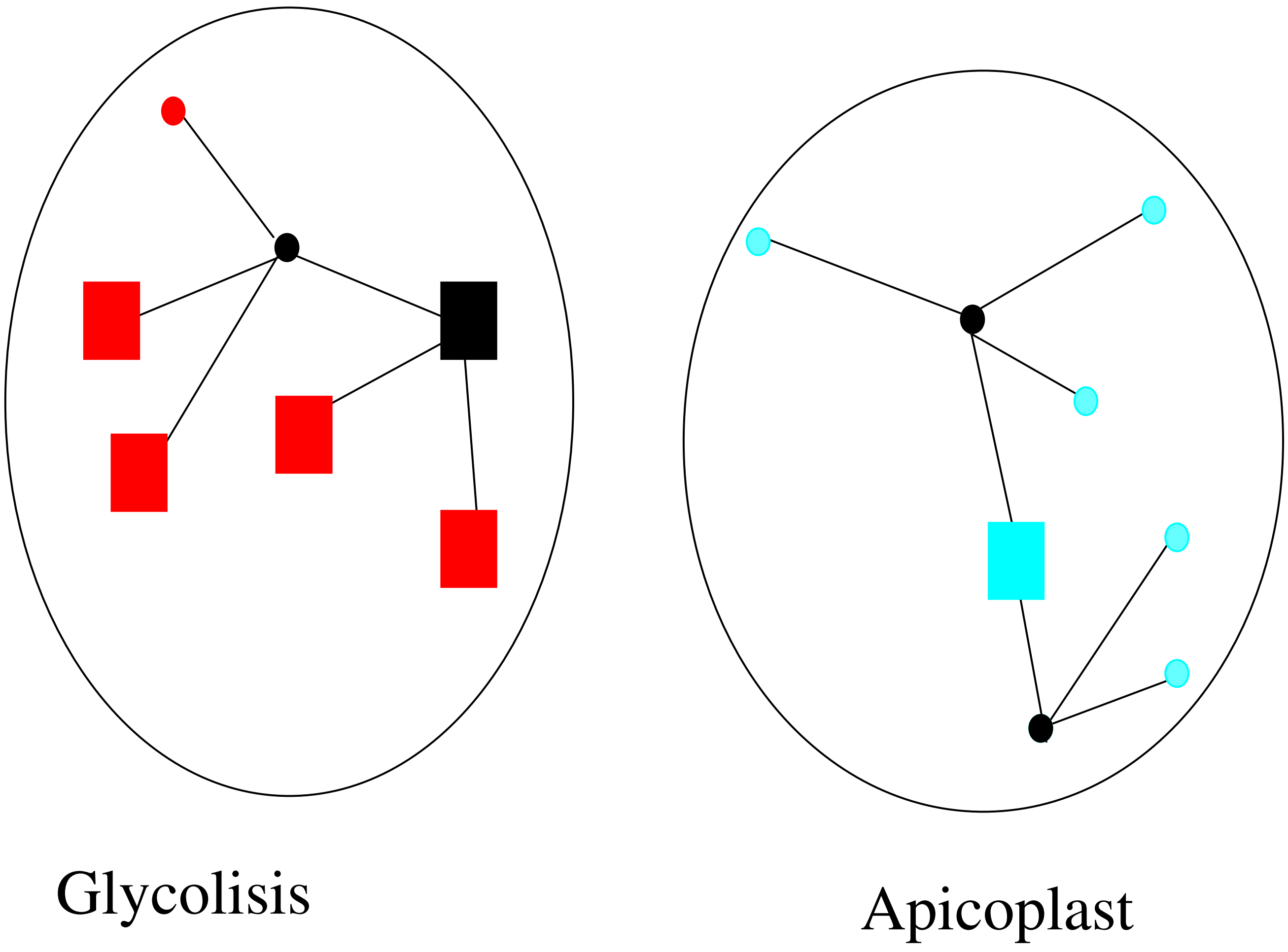
Functional Classification



Regulatory System



Interaction Graph



Glycolisis

Apicoplast

Probabilistic Genetic Network (PGN)

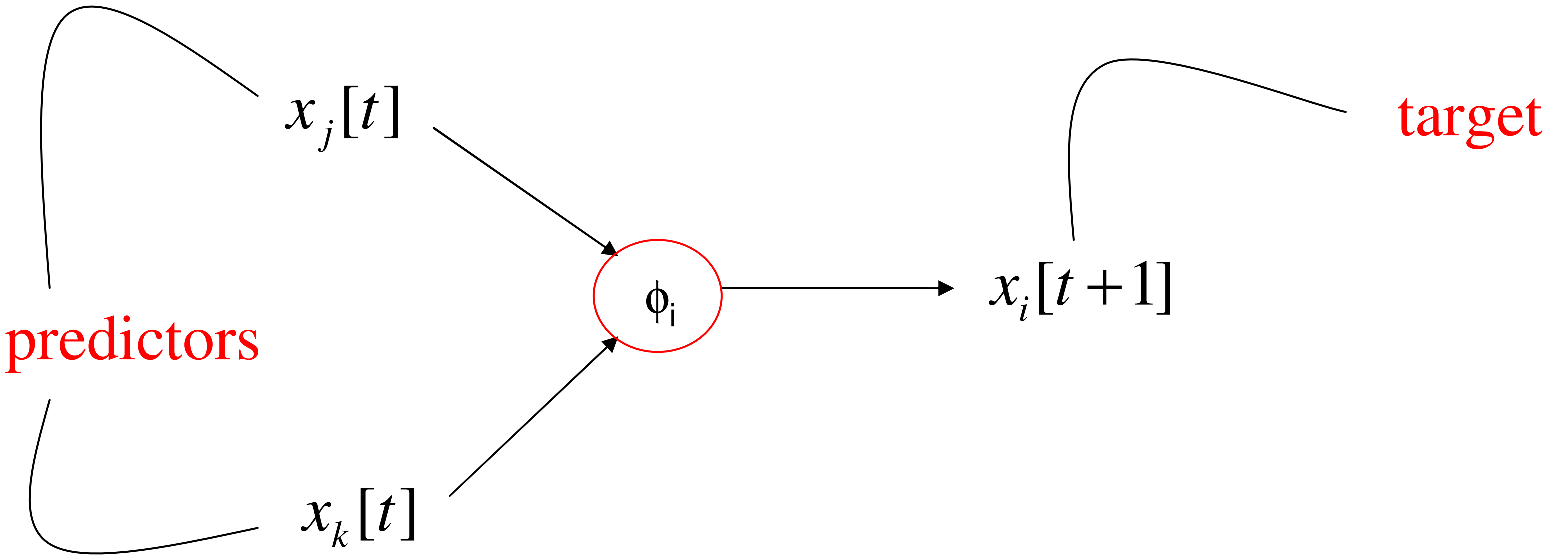
Expression of gene i at time t: $x_i[t] \in \{-1,0,+1\}$

State of the regulatory network at time t: $x[t] = \begin{bmatrix} x_1[t] \\ x_2[t] \\ . \\ . \\ x_n[t] \end{bmatrix}$

Network dynamics: $x[t+1] = \phi(x[t])$

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ . \\ . \\ . \\ \phi_n \end{bmatrix}$$

$$x_i[t+1] = \phi_i(x[t])$$



$x_j[t]$

$x_k[t]$

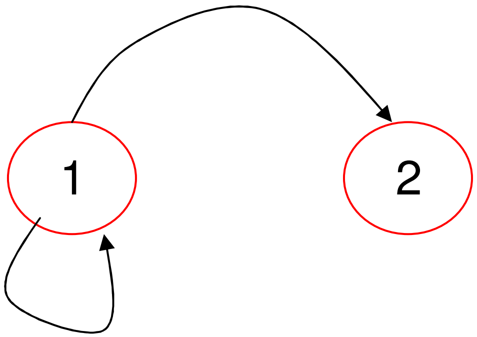
predictors

ϕ_i

$x_i[t+1]$

target

Example

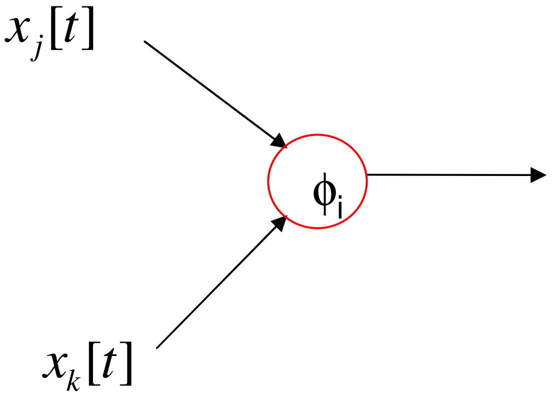


$$\phi_1(-1) = 0$$
$$\phi_1(0) = 1$$
$$\phi_1(1) = -1$$

$$\phi_2(-1) = -1$$
$$\phi_2(0) = 0$$
$$\phi_2(1) = 1$$

t	0	1	2	3	4	5	6	7	8
$x_1[t]$	-1	0	1	-1	0	1	-1	0	1
$x_2[t]$	1	-1	0	1	-1	0	1	-1	0

Probabilistic Genetic Network (PGN)



$$x_i[t+1] = \begin{cases} 1 & p(1 \mid x_j[t], x_k[t]) \\ 0 & p(0 \mid x_j[t], x_k[t]) \\ -1 & p(-1 \mid x_j[t], x_k[t]) \end{cases}$$

$$\exists\ y, z, w \in \{-1, 0, 1\},\ \ y \neq z \neq w :$$

$$p(y \mid x_j[t], x_k[t]) >> p(z \mid x_j[t], x_k[t]) + p(w \mid x_j[t], x_k[t])$$

This system

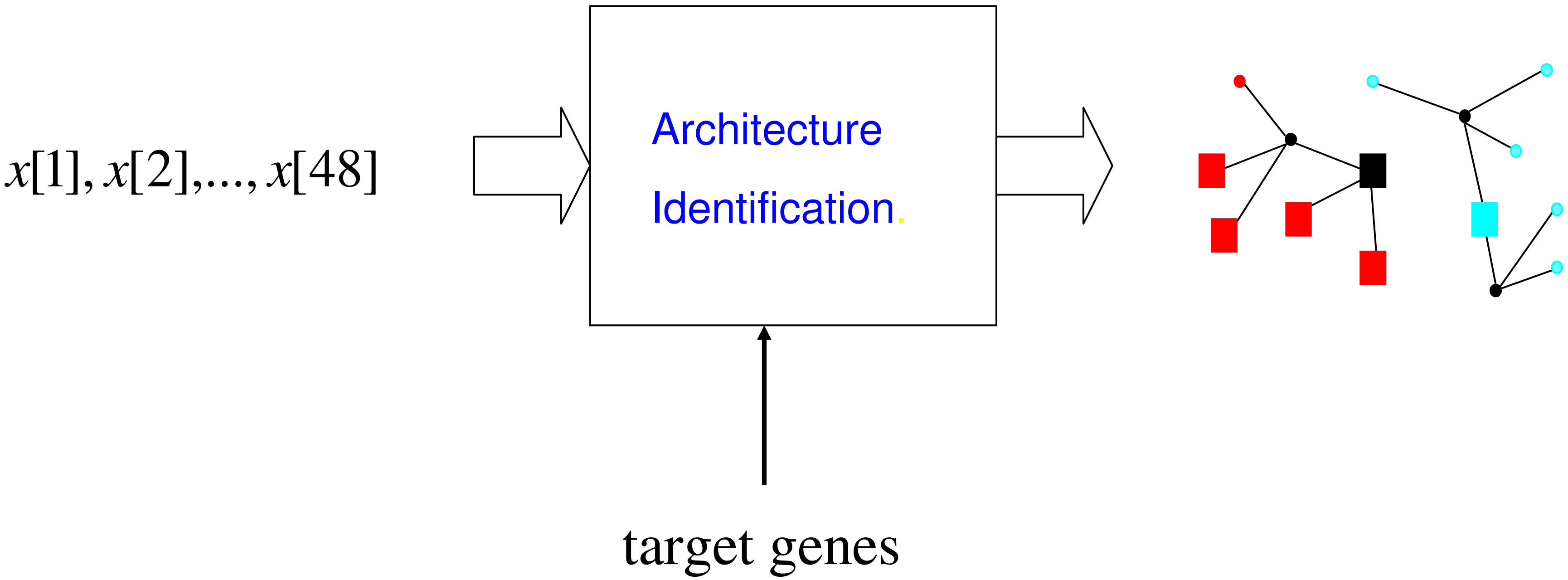
- depends just on the previous time
- is time translation invariant
- is a conditionally independent Markov chain

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

- is characterized by the conditional probabilities

$$p(x_i[t+1] \mid x[t])$$

PGN Design



$x[1], x[2],..., x[48]$

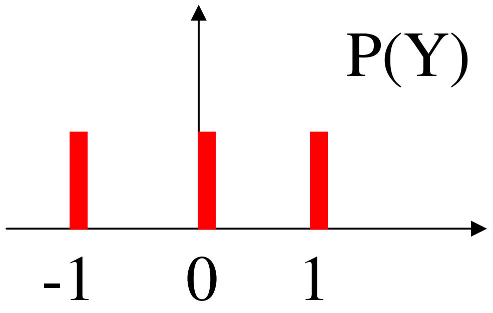
Architecture Identification.

target genes

Distribution of Y

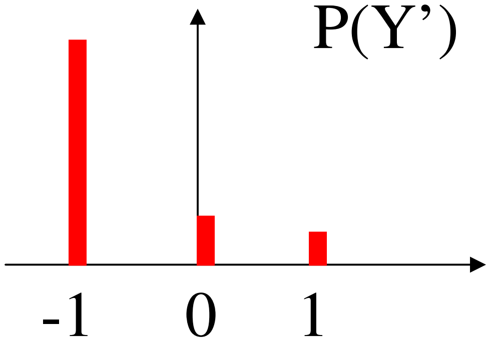
$$P : \{-1,0,1\} \rightarrow [0,1]$$

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

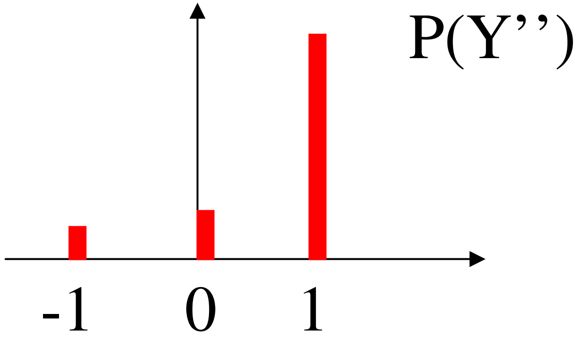


Entropy

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



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



Mutual information

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

Mean conditional entropy

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

Mean mutual information

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

Mean mutual information estimation

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

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

Estimation of P(Y|X)

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

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

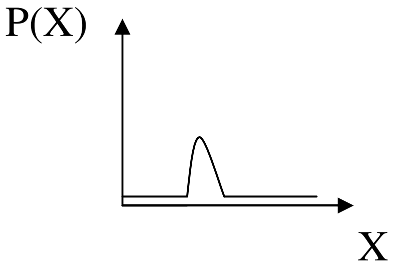
For a fixed parameter n

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

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

Estimation of P(X) for a fixed parameter n

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



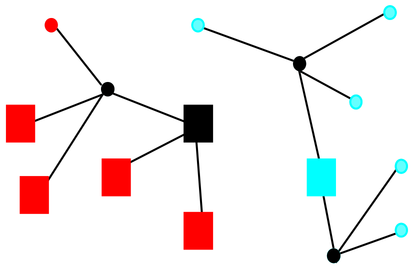
$$N^+ = \sum_{\#(X=(a,b))\geq n, \forall (a,b)} \#(X=(a,b)) \qquad N^- = \sum_{\#(X=(a,b))<n, \forall (a,b)} \#(X=(a,b))$$

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

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

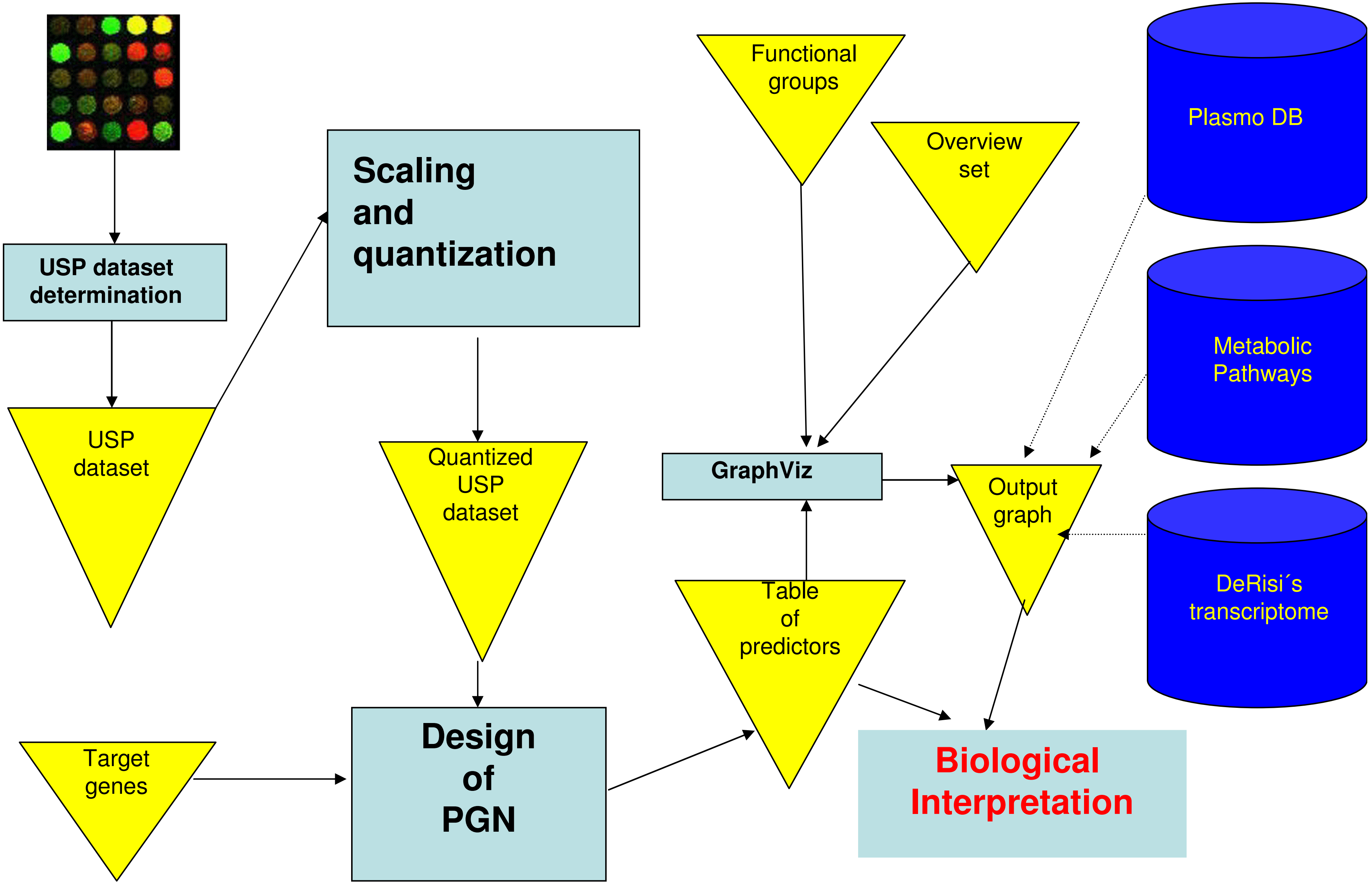
Buiding Interaction Graphes

- 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

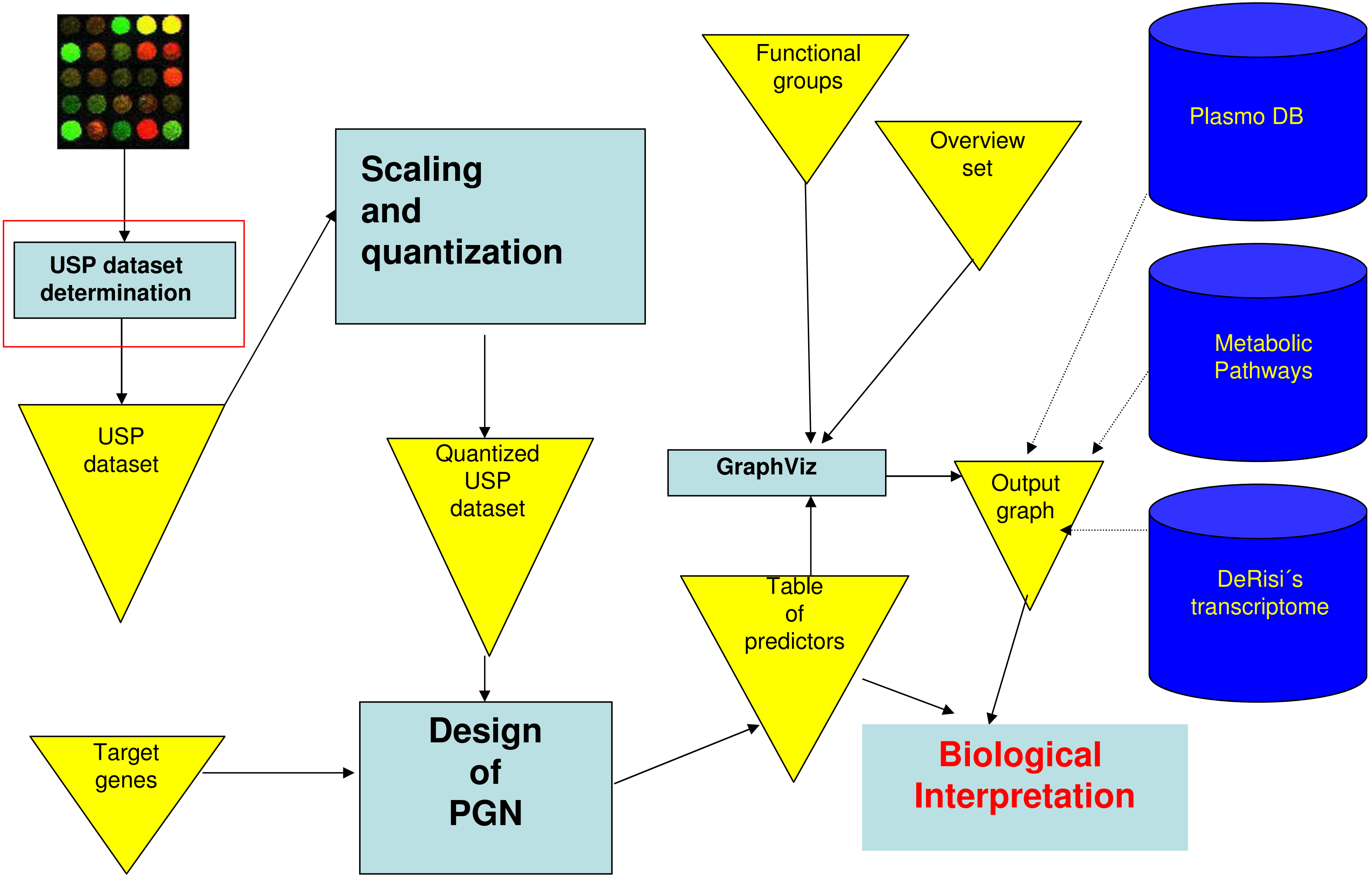


Data analysis pipeline

System architecture



System architecture



USP dataset determination

USP dataset

Scaling and quantization

Quantized USP dataset

Target genes

Design of PGN

Table of predictors

Functional groups

Overview set

GraphViz

Output graph

Plasmo DB

Metabolic Pathways

DeRisi´s transcriptome

Biological Interpretation

USP-dataset

- directly from original .gpr “raw” data;
- intensity = foreground mean - background median;
- mean for replicated time points;
- different definition of “weak” spots and elimination rules;
- consider ALL accepted oligos as unique entities (including almost sinusoidal).

USP-dataset: 6532 oligos

Overview dataset: 3719 oligos

Weak spots definition

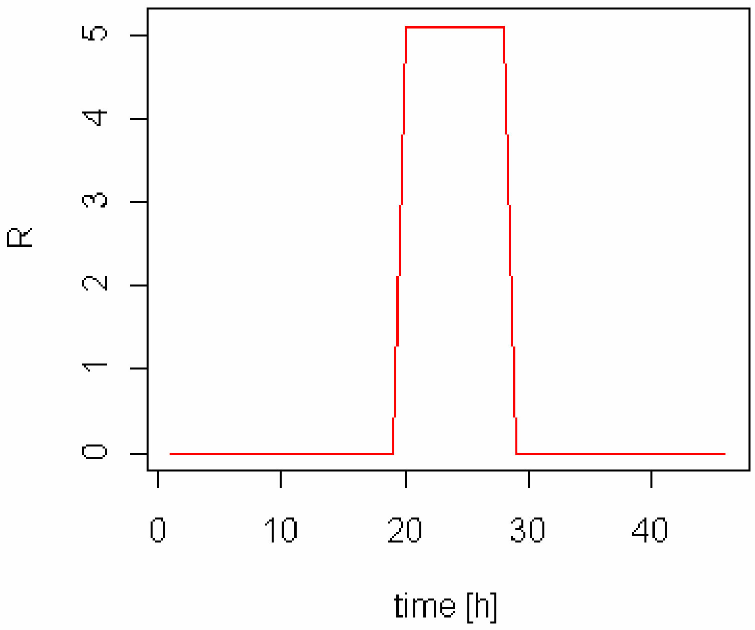
$\mathbf{X}$ = (0, 0, ... , 100, 100, ... , 100, 0, 0, ... , 0, 0)

$<\mathbf{X}>$ = 9 * 100 / 46 = 19.56

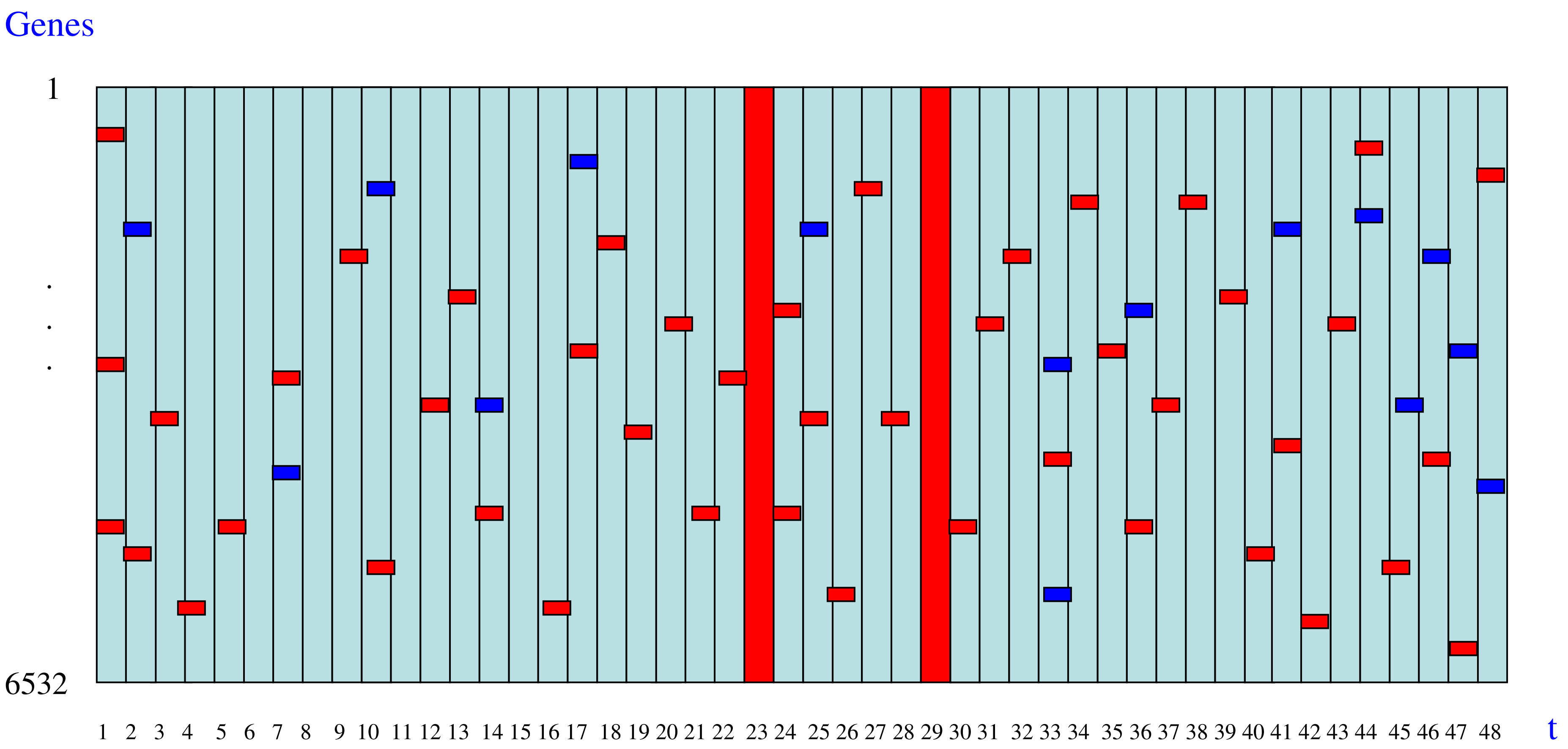
$\mathbf{R}$ = normalized cy5/cy3 = $\mathbf{X}/<\mathbf{X}>$ =

$\mathbf{R}$ = (0, 0, ... , 5.11, 5.11, ... , 5.11, 0, 0, ... , 0, 0)

$\log_2(\mathbf{R}) = (-\infty, -\infty, \ldots, 1.63, 1.63, \ldots, 1.63, -\infty, -\infty, \ldots, -\infty)$



Not amenable to Fourier analysis due to infinities.



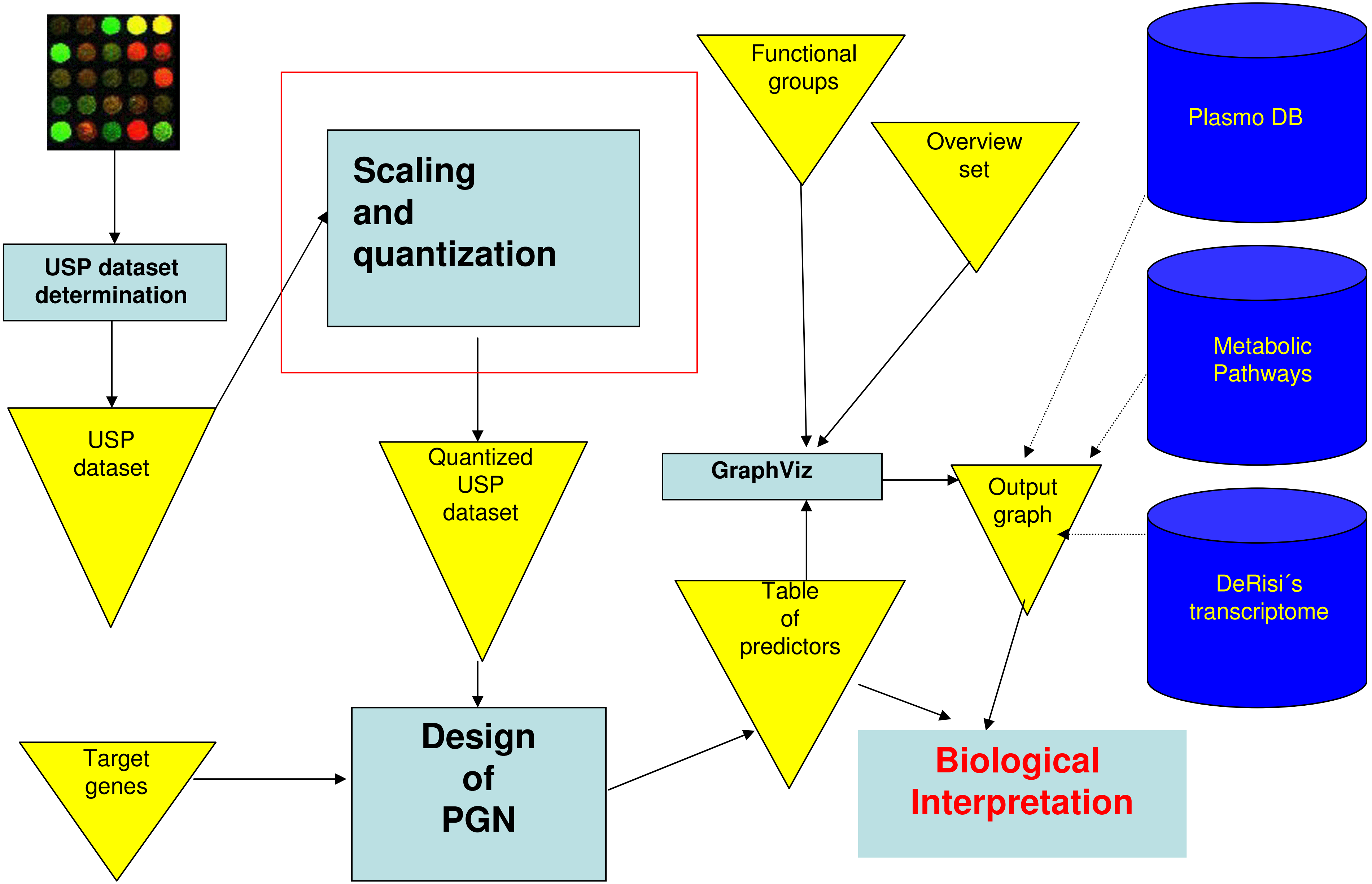
⬜ Good spots

🟦 Weak spots

🟥 Bad spots

NO INTERPOLATION

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

Scaling

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

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

of the □ spots

Scale normalization of the □ and 🟦 spots

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

Quantization

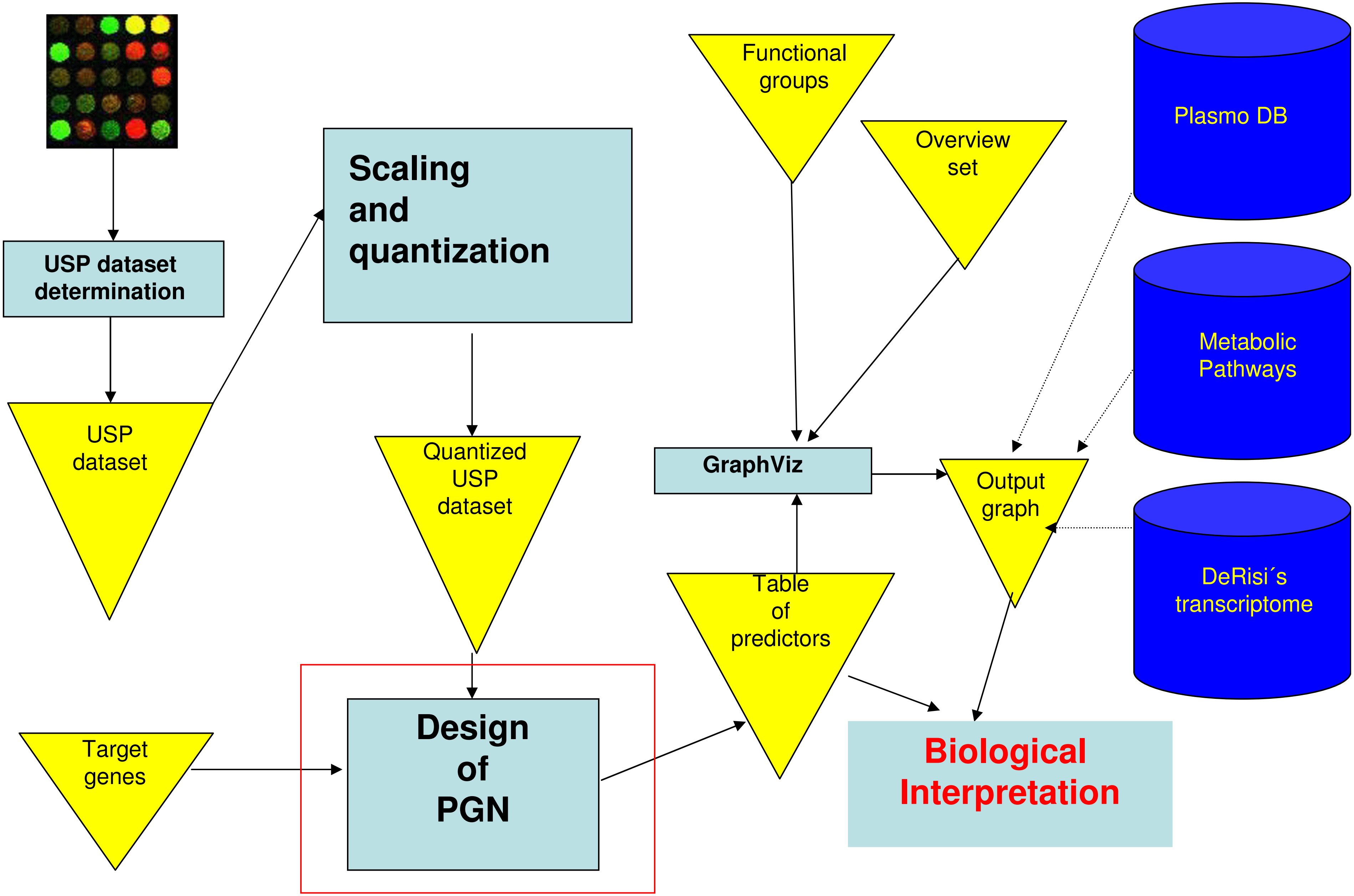
Let $n_i^+[t]$ and $n_i^-[t]$ denote, respectively, the normalized signals greater and lower than zero at t.

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

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

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

System architecture



USP dataset determination

Scaling and quantization

USP dataset

Quantized USP dataset

Target genes

Design of PGN

Functional groups

Overview set

GraphViz

Table of predictors

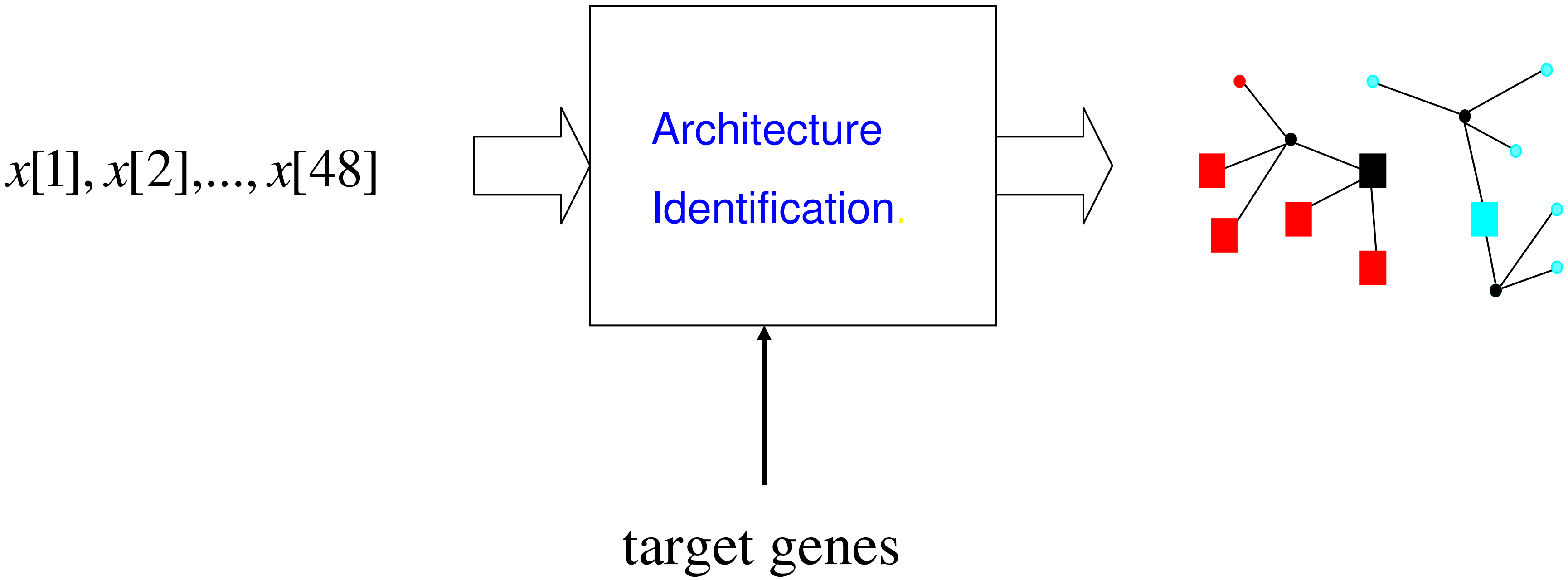
Output graph

Plasmo DB

Metabolic Pathways

DeRisi´s transcriptome

Biological Interpretation



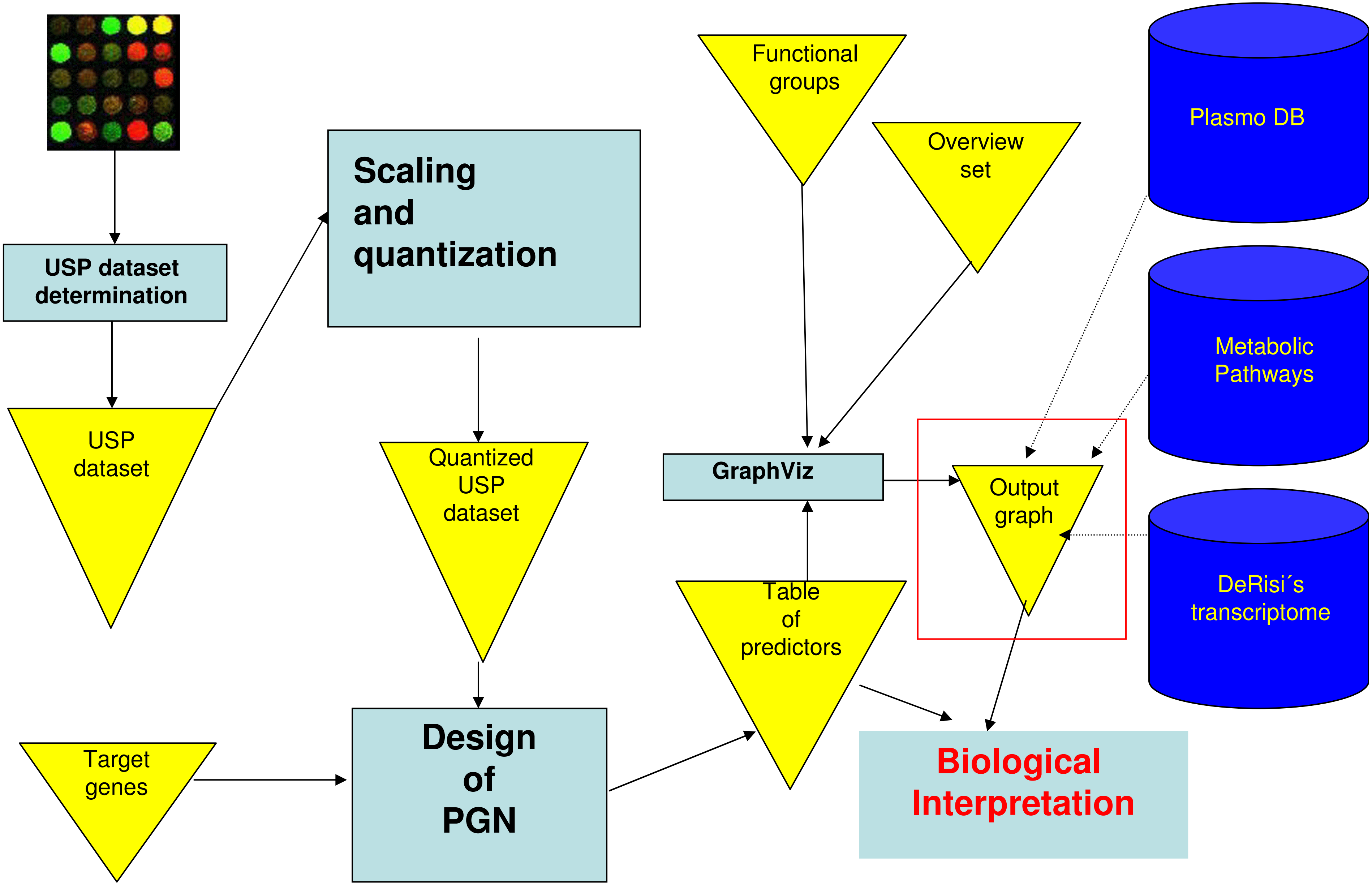
$x[1], x[2],..., x[48]$

Architecture

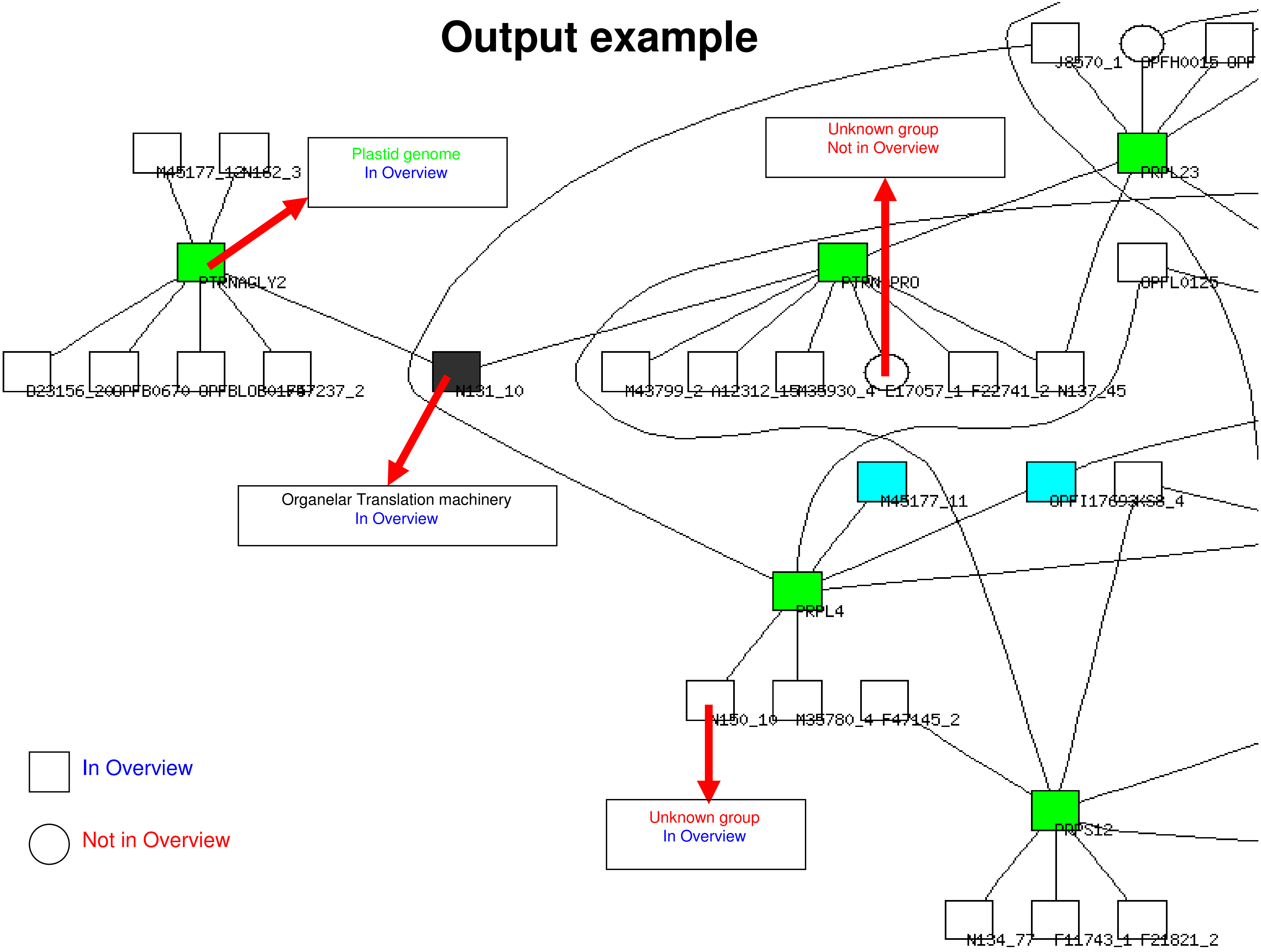
Identification.

target genes

System architecture



Output example



Plastid genome
In Overview

Unknown group
Not in Overview

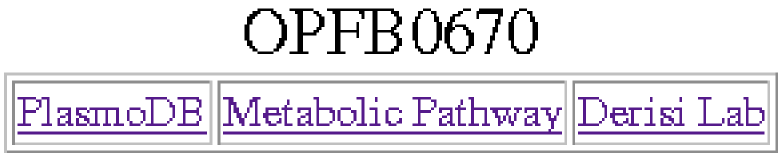
Organelar Translation machinery
In Overview

Unknown group
In Overview

□ In Overview

○ Not in Overview

Back



OPFB0670

PlasmoDB	Metabolic Pathway	Derisi Lab

P. falciparum PFB0330c

Back

Home Downloads Tools Queries BLAST History CDs & Links Browse Data Sources SRT Help

Plasmodium falciparum / CHR 2 / PFB0330c

cysteine protease, putative

Summary view

Add this gene to your History

Annotation
- Curated Annotation
- UserComments
- GO Process
- GO Component
- GO Function
- EC number
- RefSeqs
- Metabolic Pathways
- MR4 Reagents
- Ortholog Group
- Ortholog Views
- Orthologs
- BLASTP non-Pf (graphic)
- BLASTP other (graphic)
- BLASTP NRDB

Protein
- PDB structures
- Structural Models
- Features (graphic)
- Pfam
- PROSITE
- TM domains
- SignalP
- PlasmoAP
- Motifs (graphic)
- Motifs
- Proteomics (graphic)
- Mass spec. data

Expression
- Microarrays
- Developmental series (clone array)
- Developmental series (Affy array)
- Developmental series (glass slide array)
- Proteomics (graphic)
- Mass spec. data

Sequence
- DNA (graphic)
- Exons
- SNPs
- mRNA/RNA sequence
- Protein sequence

Annotation

back to top

Curated Annotation

*** *None* ***

P. falciparum Gene: PFB0330c

Back

ID: PFB0330C

Comment:

This gene was predicted and reviewed manually for the Oct. 3, 2002 Nature publication by Gardner et al.This gene has at least one intron

Superclasses: Genes -> UNCLASSIFIED

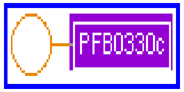
Chromosome: Chromosome 2

Map Position (centisomes): 31.287 [click to view in chromosome browser]

Map Position (nucleotides): 296,317 -> 297,583

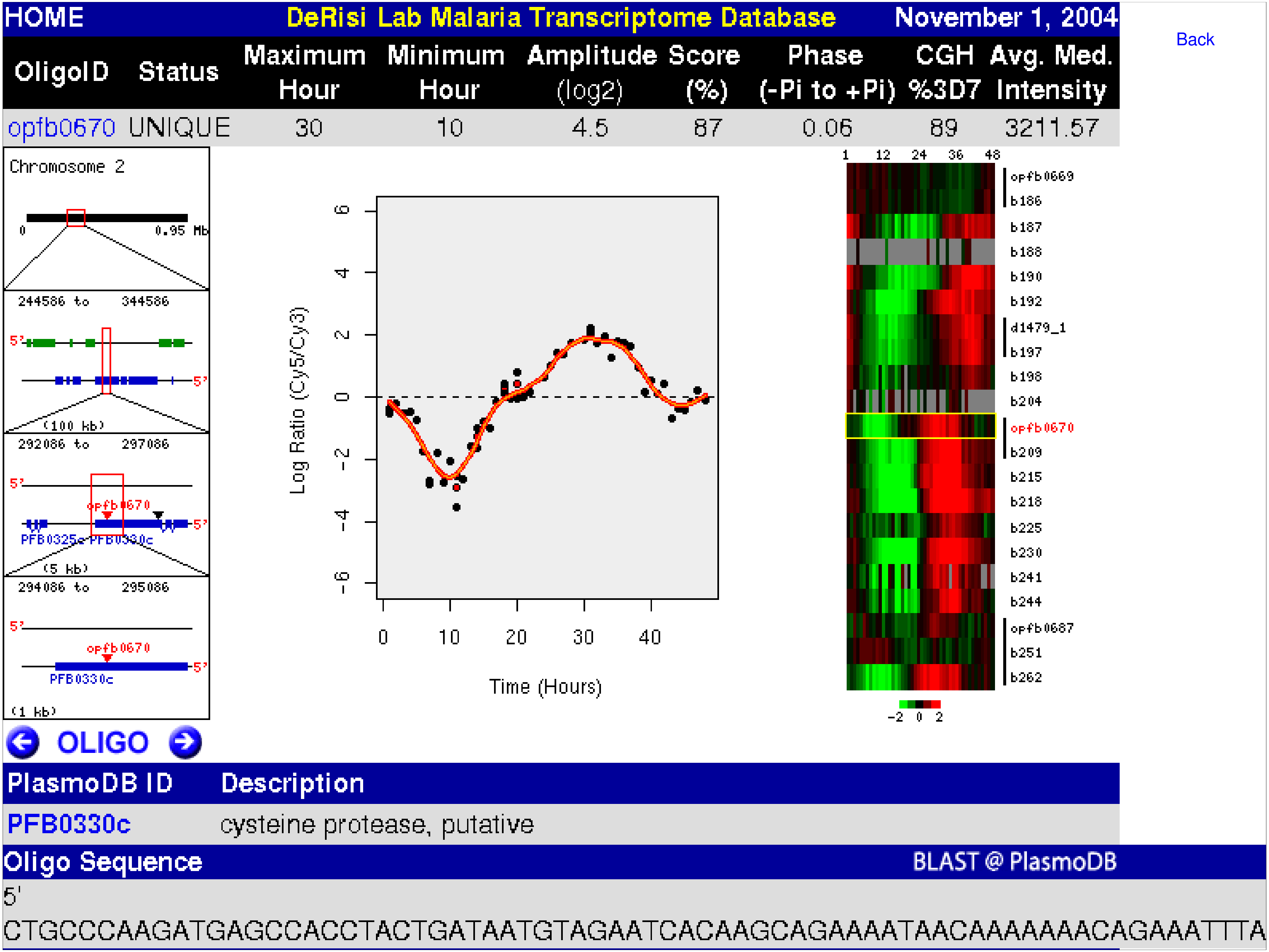
Products: cysteine protease, putative

Gene-Reaction Schematic: ?



Query Page | Advanced Query Page | BioCyc Home | Report Errors or Provide Feedback

SRI International Pathway Tools software, page generated on Mon Nov 1, 2004.



HOME DeRisi Lab Malaria Transcriptome Database November 1, 2004

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

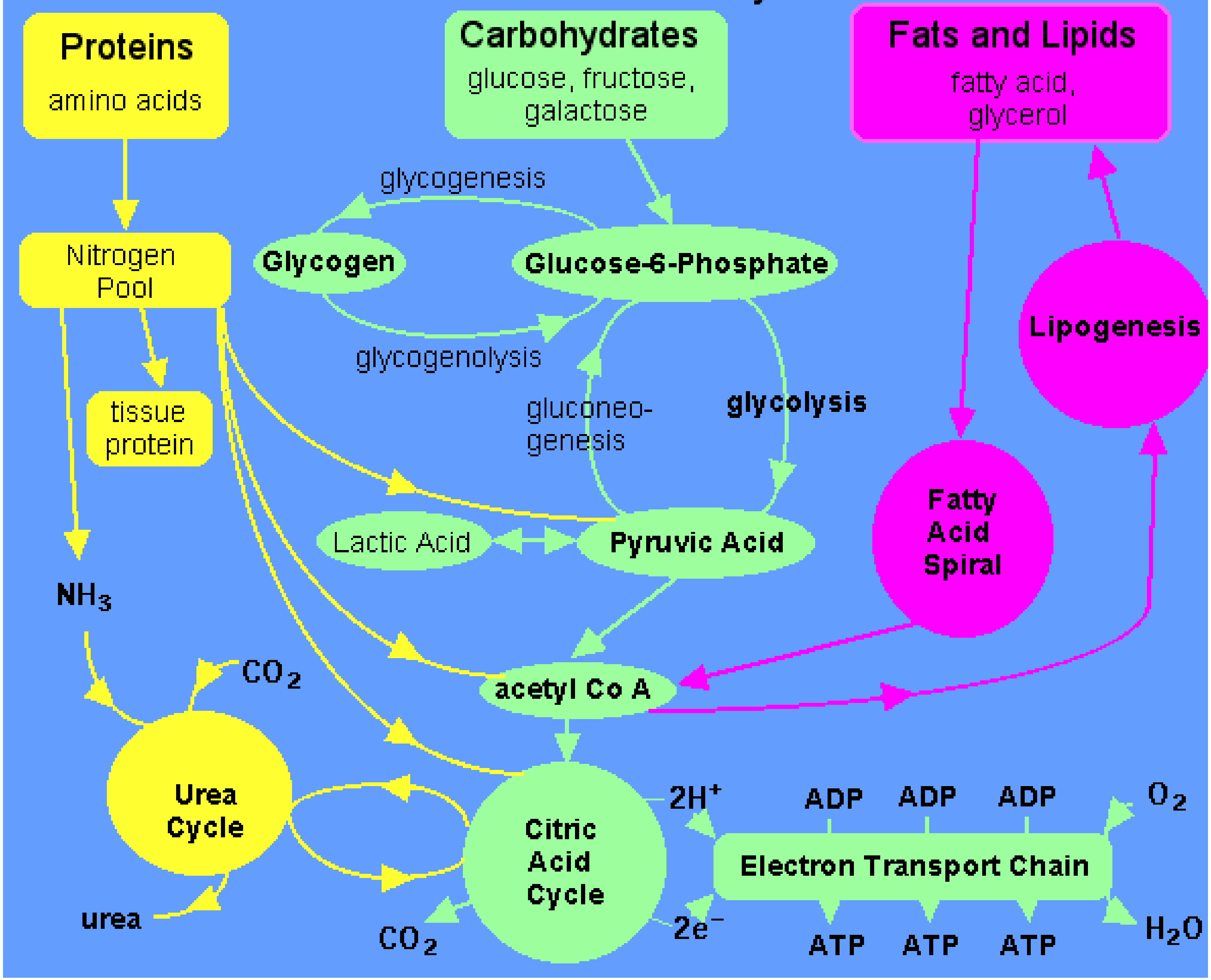
PlasmoDB ID	Description
PFB0330c	cysteine protease, putative

Oligo Sequence BLAST @ PlasmoDB

5'
CTGCCCAAGATGAGCCACCTACTGATAATGTAGAATCACAAGCAGAAAATAACAAAAAAACAGAAATTTA

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

NH_3

CO_2

acetyl Co A

Urea Cycle

Citric Acid Cycle

$2H^+$

ADP

ADP

ADP

O_2

Electron Transport Chain

urea

CO_2

$2e^-$

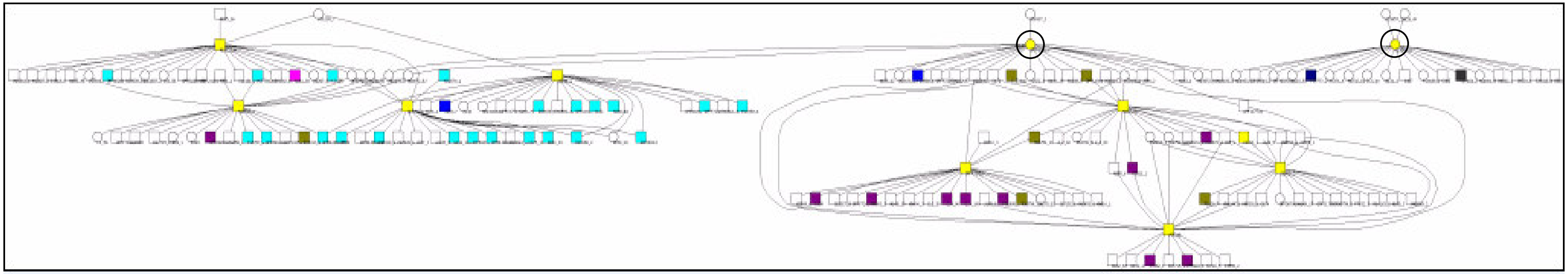
ATP

ATP

ATP

H_2O

Glycolytic PGN network (single genes)



● glycolysis

● transcription machinery

● cytoplasmic translation

● ribonucleotide synthesis

● deoxynucleotide synthesis

● DNA replication

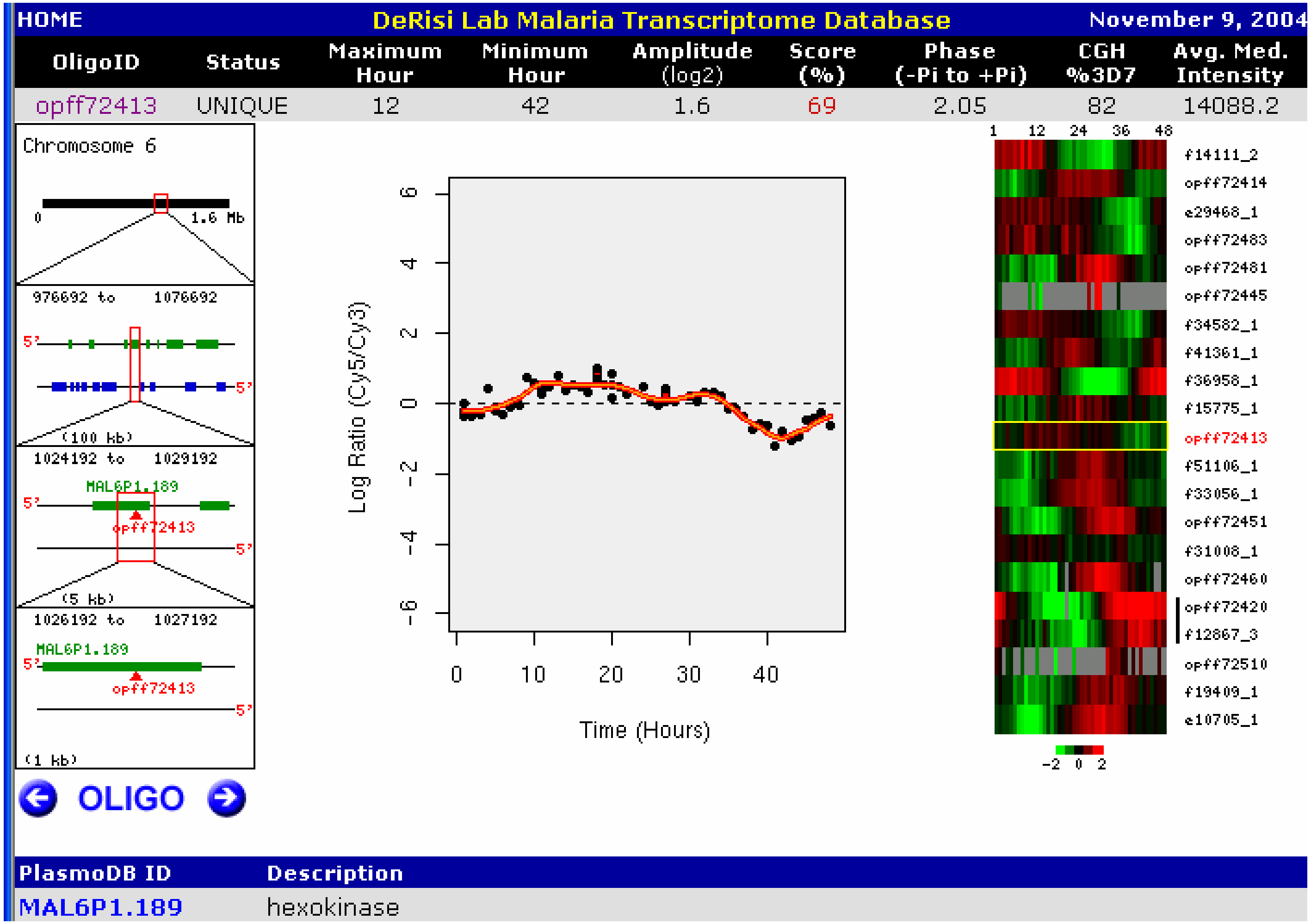
● proteoasome

● plastid genome

● merozoite invasion (kinases)

● actin myosoin motors

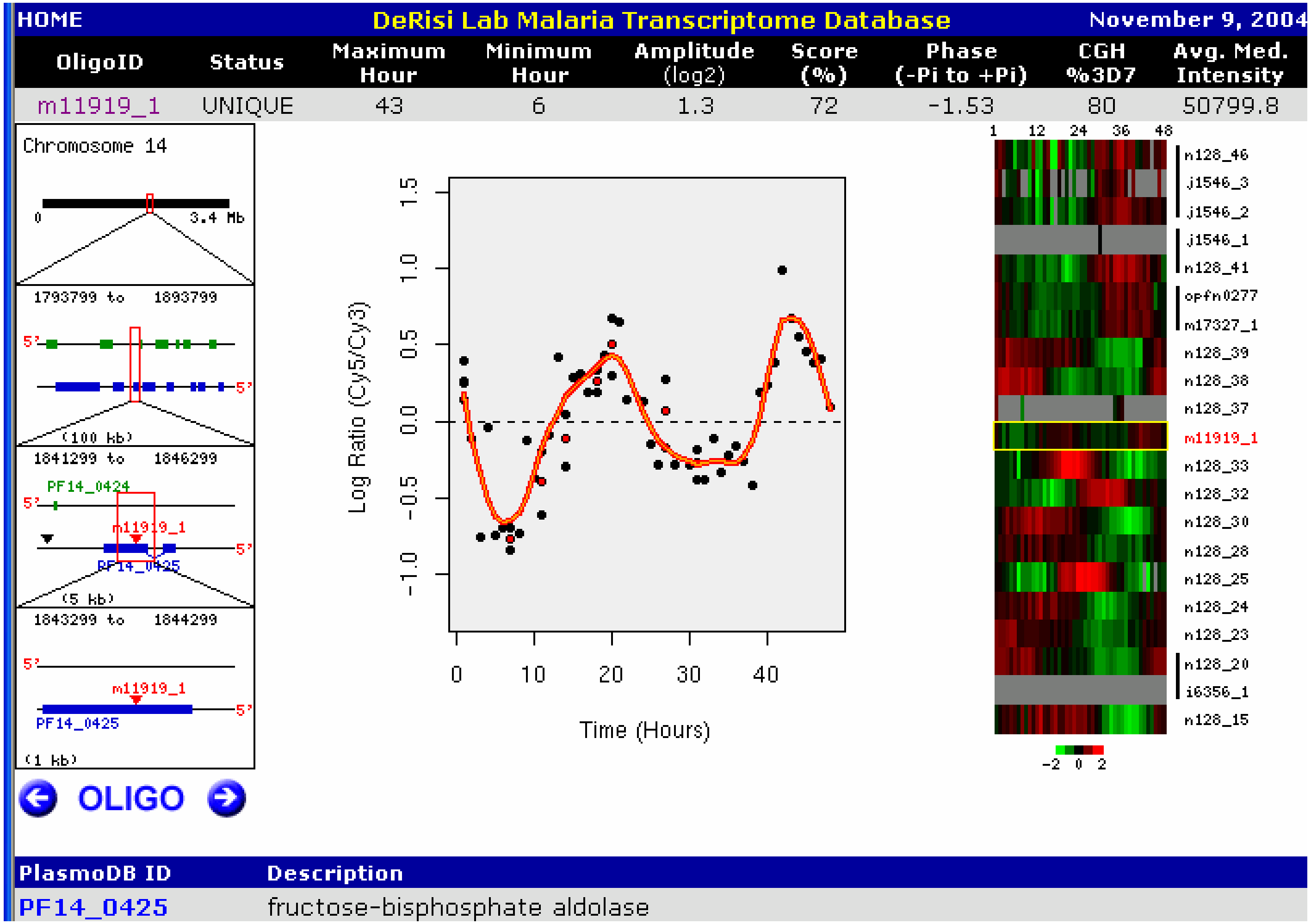
● early ring transcripts



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

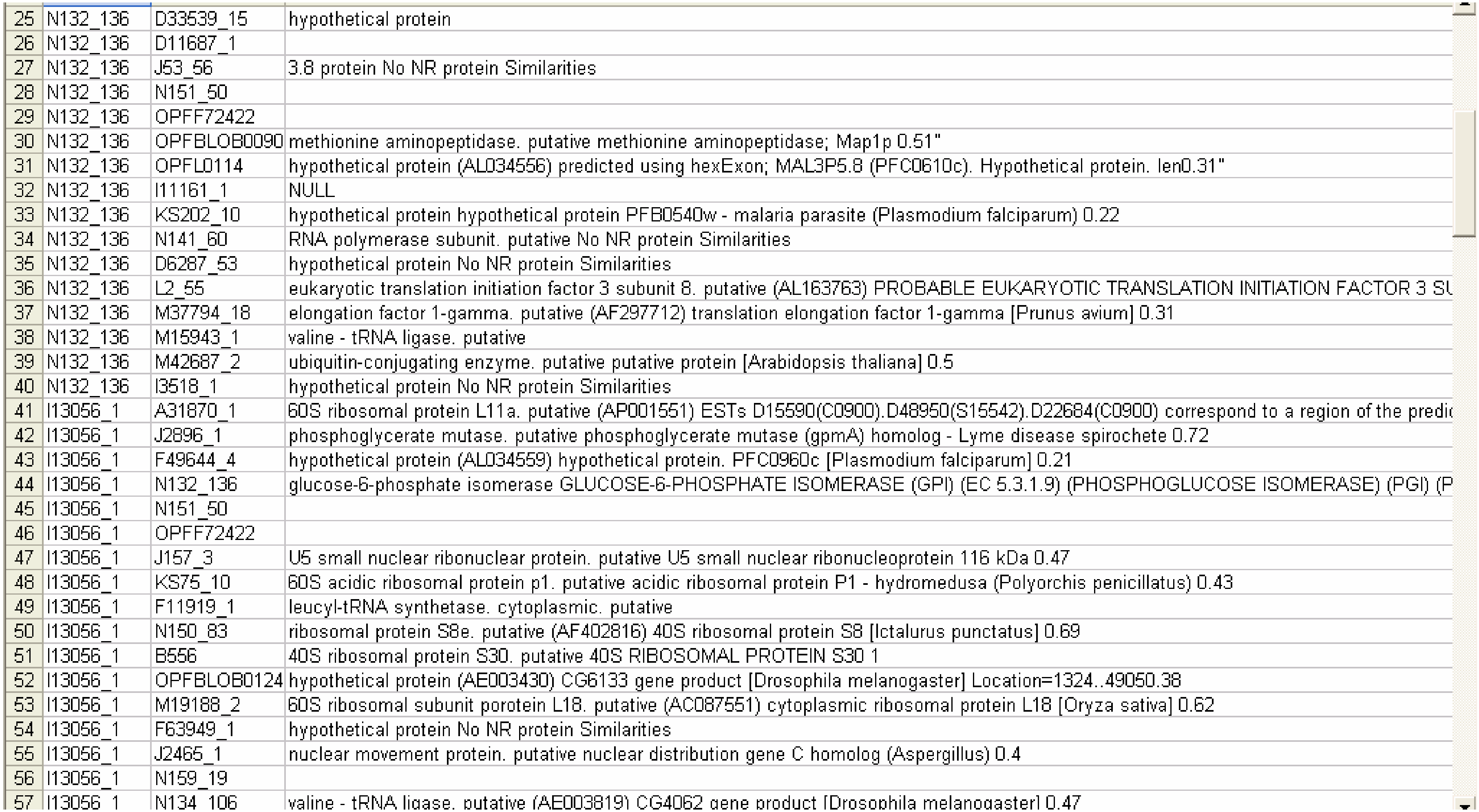
PlasmoDB ID	Description
MAL6P1.189	hexokinase



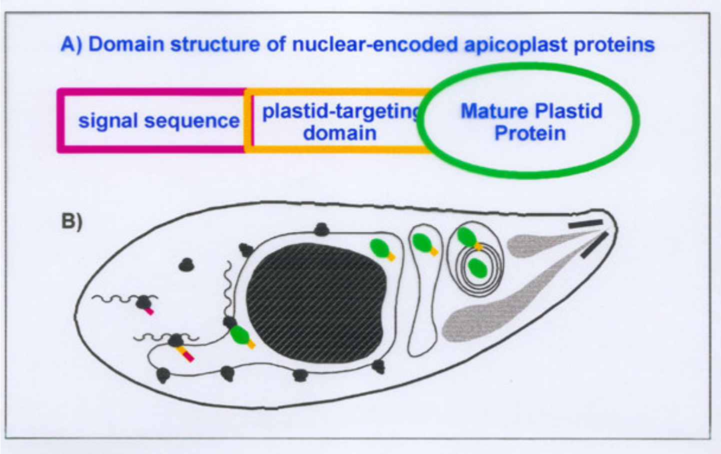
HOME | DeRisi Lab Malaria Transcriptome Database | November 9, 2004

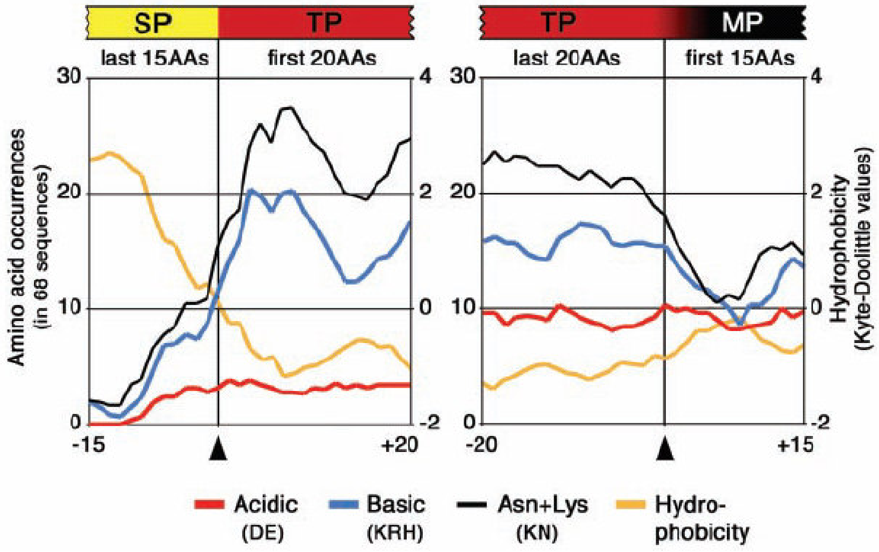
OligoID	Status	Maximum Hour	Minimum Hour	Amplitude (log2)	Score (%)	Phase (-Pi to +Pi)	CGH %3D7	Avg. Med. Intensity
m11919_1	UNIQUE	43	6	1.3	72	-1.53	80	50799.8

PlasmoDB ID	Description
PF14_0425	fructose-bisphosphate aldolase

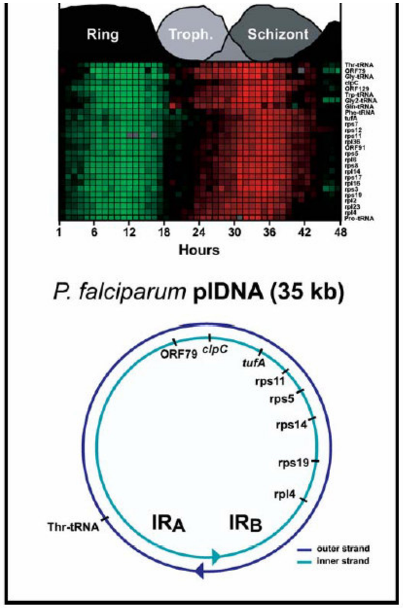


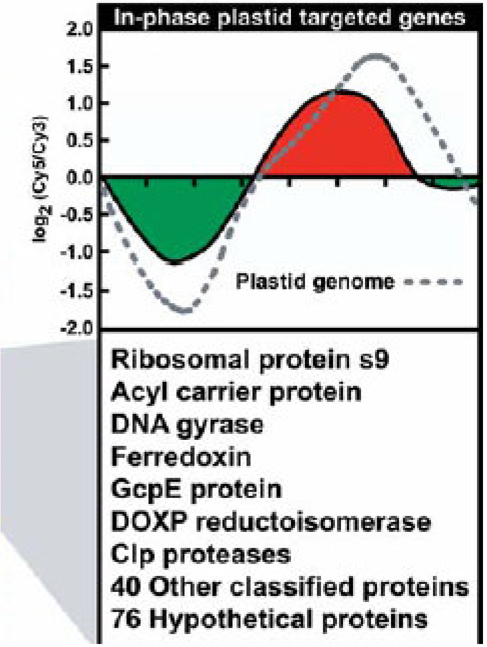
25	N132_136	D33539_15	hypothetical protein
26	N132_136	D11687_1	
27	N132_136	J53_56	3.8 protein No NR protein Similarities
28	N132_136	N151_50	
29	N132_136	OPFF72422	
30	N132_136	OPFBLOB0090	methionine aminopeptidase. putative methionine aminopeptidase; Map1p 0.51"
31	N132_136	OPFL0114	hypothetical protein (AL034556) predicted using hexExon; MAL3P5.8 (PFC0610c). Hypothetical protein. len0.31"
32	N132_136	I11161_1	NULL
33	N132_136	KS202_10	hypothetical protein hypothetical protein PFB0540w - malaria parasite (Plasmodium falciparum) 0.22
34	N132_136	N141_60	RNA polymerase subunit. putative No NR protein Similarities
35	N132_136	D6287_53	hypothetical protein No NR protein Similarities
36	N132_136	L2_55	eukaryotic translation initiation factor 3 subunit 8. putative (AL163763) PROBABLE EUKARYOTIC TRANSLATION INITIATION FACTOR 3 SU
37	N132_136	M37794_18	elongation factor 1-gamma. putative (AF297712) translation elongation factor 1-gamma [Prunus avium] 0.31
38	N132_136	M15943_1	valine - tRNA ligase. putative
39	N132_136	M42687_2	ubiquitin-conjugating enzyme. putative putative protein [Arabidopsis thaliana] 0.5
40	N132_136	I3518_1	hypothetical protein No NR protein Similarities
41	I13056_1	A31870_1	60S ribosomal protein L11a. putative (AP001551) ESTs D15590(C0900).D48950(S15542).D22684(C0900) correspond to a region of the predic
42	I13056_1	J2896_1	phosphoglycerate mutase. putative phosphoglycerate mutase (gpmA) homolog - Lyme disease spirochete 0.72
43	I13056_1	F49644_4	hypothetical protein (AL034559) hypothetical protein. PFC0960c [Plasmodium falciparum] 0.21
44	I13056_1	N132_136	glucose-6-phosphate isomerase GLUCOSE-6-PHOSPHATE ISOMERASE (GPI) (EC 5.3.1.9) (PHOSPHOGLUCOSE ISOMERASE) (PGI) (P
45	I13056_1	N151_50	
46	I13056_1	OPFF72422	
47	I13056_1	J157_3	U5 small nuclear ribonuclear protein. putative U5 small nuclear ribonucleoprotein 116 kDa 0.47
48	I13056_1	KS75_10	60S acidic ribosomal protein p1. putative acidic ribosomal protein P1 - hydromedusa (Polyorchis penicillatus) 0.43
49	I13056_1	F11919_1	leucyl-tRNA synthetase. cytoplasmic. putative
50	I13056_1	N150_83	ribosomal protein S8e. putative (AF402816) 40S ribosomal protein S8 [Ictalurus punctatus] 0.69
51	I13056_1	B556	40S ribosomal protein S30. putative 40S RIBOSOMAL PROTEIN S30 1
52	I13056_1	OPFBLOB0124	hypothetical protein (AE003430) CG6133 gene product [Drosophila melanogaster] Location=1324..49050.38
53	I13056_1	M19188_2	60S ribosomal subunit porotein L18. putative (AC087551) cytoplasmic ribosomal protein L18 [Oryza sativa] 0.62
54	I13056_1	F63949_1	hypothetical protein No NR protein Similarities
55	I13056_1	J2465_1	nuclear movement protein. putative nuclear distribution gene C homolog (Aspergillus) 0.4
56	I13056_1	N159_19	
57	I13056_1	N134_106	valine - tRNA ligase. putative (AE003819) CG4062 gene product [Drosophila melanogaster] 0.47





550 apicoplast proteins

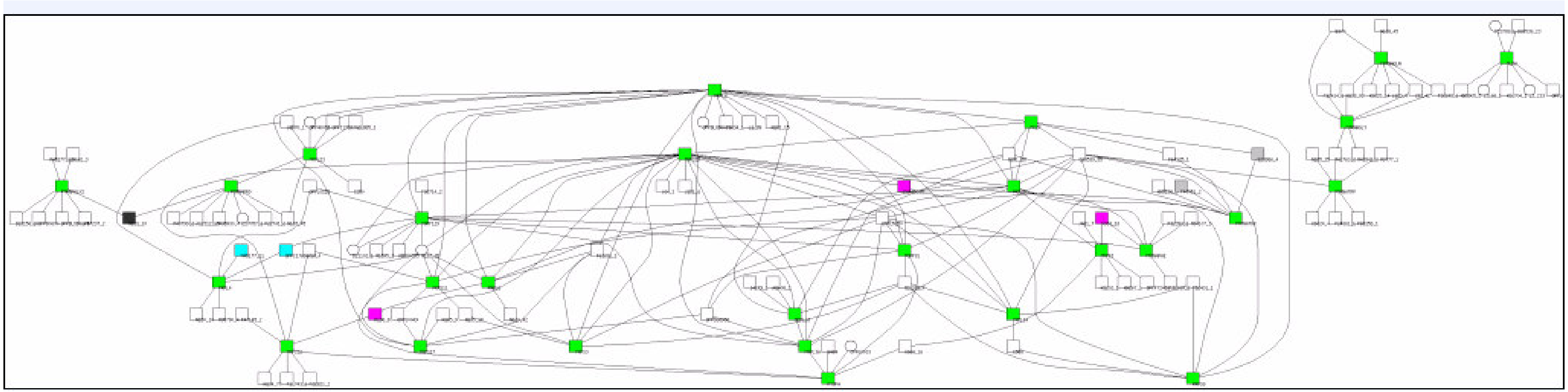




Ribosomal protein s9
Acyl carrier protein
DNA gyrase
Ferredoxin
GcpE protein
DOXP reductoisomerase
Clp proteases
40 Other classified proteins
76 Hypothetical proteins

124 apicoplast proteins

Apicoplast PGN network (single genes)



● glycolysis

● transcription machinery

● cytoplasmic translation

● ribonucleotide synthesis

● deoxynucleotide synthesis

● DNA replication

● proteoasome

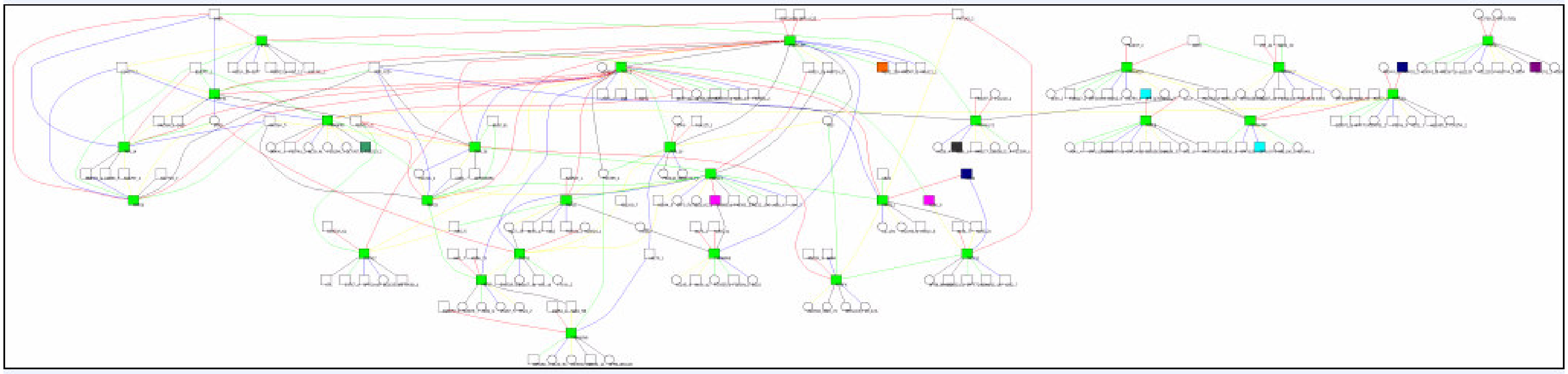
● plastid genome

● merozoite invasion

● actin myosoin motors

● mitochondrial

Apicoplast PGN network (double genes)



● glycolysis

● transcription machinery

● cytoplasmic translation

● ribonucleotide synthesis

● deoxynucleotide synthesis

● DNA replication

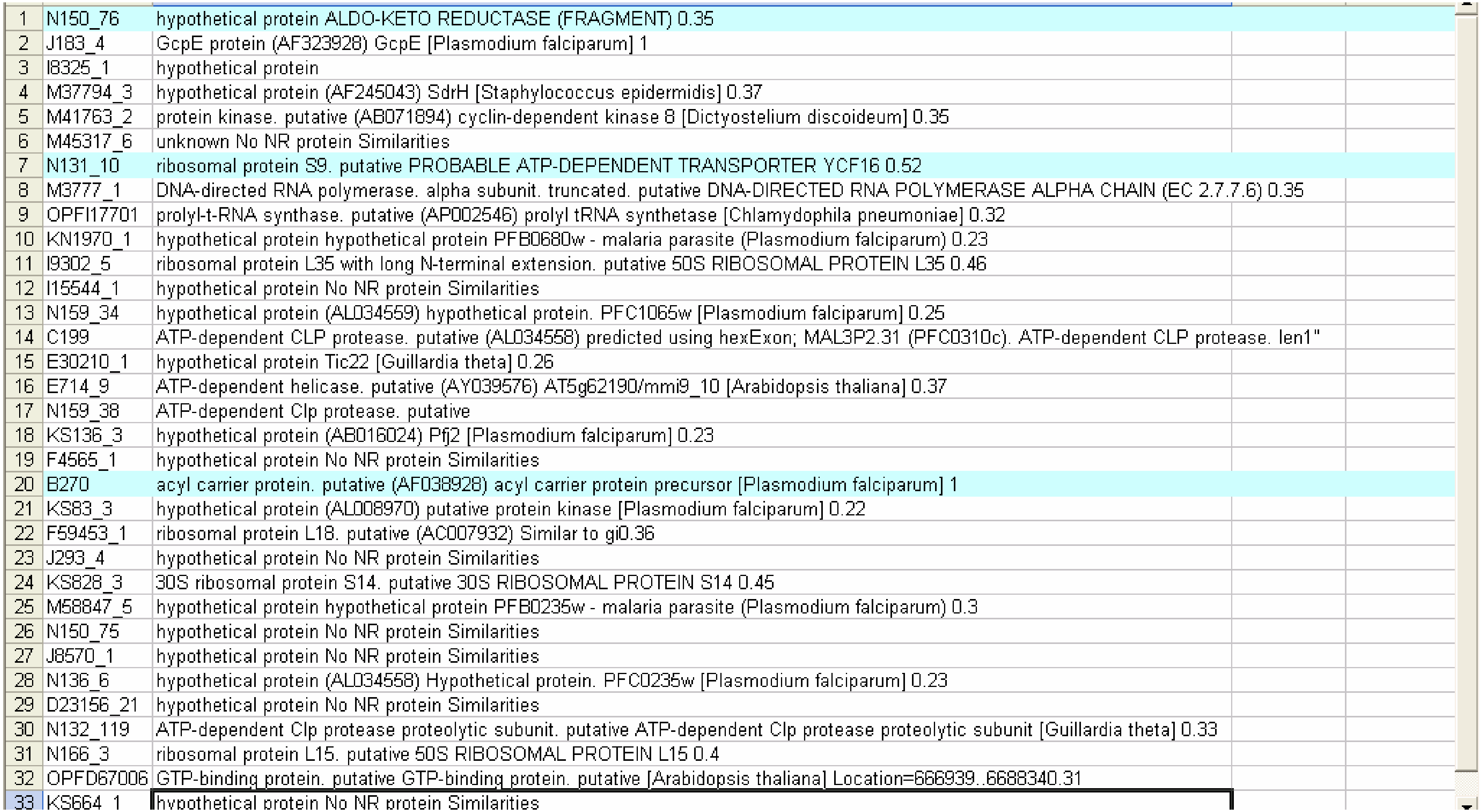
● proteoasome

● plastid genome

● merozoite invasion

● actin myosoin motors

● early ring transcripts



1	N150_76	hypothetical protein ALDO-KETO REDUCTASE (FRAGMENT) 0.35
2	J183_4	GcpE protein (AF323928) GcpE [Plasmodium falciparum] 1
3	I8325_1	hypothetical protein
4	M37794_3	hypothetical protein (AF245043) SdrH [Staphylococcus epidermidis] 0.37
5	M41763_2	protein kinase. putative (AB071894) cyclin-dependent kinase 8 [Dictyostelium discoideum] 0.35
6	M45317_6	unknown No NR protein Similarities
7	N131_10	ribosomal protein S9. putative PROBABLE ATP-DEPENDENT TRANSPORTER YCF16 0.52
8	M3777_1	DNA-directed RNA polymerase. alpha subunit. truncated. putative DNA-DIRECTED RNA POLYMERASE ALPHA CHAIN (EC 2.7.7.6) 0.35
9	OPFI17701	prolyl-t-RNA synthase. putative (AP002546) prolyl tRNA synthetase [Chlamydophila pneumoniae] 0.32
10	KN1970_1	hypothetical protein hypothetical protein PFB0680w - malaria parasite (Plasmodium falciparum) 0.23
11	I9302_5	ribosomal protein L35 with long N-terminal extension. putative 50S RIBOSOMAL PROTEIN L35 0.46
12	I15544_1	hypothetical protein No NR protein Similarities
13	N159_34	hypothetical protein (AL034559) hypothetical protein. PFC1065w [Plasmodium falciparum] 0.25
14	C199	ATP-dependent CLP protease. putative (AL034558) predicted using hexExon; MAL3P2.31 (PFC0310c). ATP-dependent CLP protease. len1"
15	E30210_1	hypothetical protein Tic22 [Guillardia theta] 0.26
16	E714_9	ATP-dependent helicase. putative (AY039576) AT5g62190/mmi9_10 [Arabidopsis thaliana] 0.37
17	N159_38	ATP-dependent Clp protease. putative
18	KS136_3	hypothetical protein (AB016024) Pfj2 [Plasmodium falciparum] 0.23
19	F4565_1	hypothetical protein No NR protein Similarities
20	B270	acyl carrier protein. putative (AF038928) acyl carrier protein precursor [Plasmodium falciparum] 1
21	KS83_3	hypothetical protein (AL008970) putative protein kinase [Plasmodium falciparum] 0.22
22	F59453_1	ribosomal protein L18. putative (AC007932) Similar to gi0.36
23	J293_4	hypothetical protein No NR protein Similarities
24	KS828_3	30S ribosomal protein S14. putative 30S RIBOSOMAL PROTEIN S14 0.45
25	M58847_5	hypothetical protein hypothetical protein PFB0235w - malaria parasite (Plasmodium falciparum) 0.3
26	N150_75	hypothetical protein No NR protein Similarities
27	J8570_1	hypothetical protein No NR protein Similarities
28	N136_6	hypothetical protein (AL034558) Hypothetical protein. PFC0235w [Plasmodium falciparum] 0.23
29	D23156_21	hypothetical protein No NR protein Similarities
30	N132_119	ATP-dependent Clp protease proteolytic subunit. putative ATP-dependent Clp protease proteolytic subunit [Guillardia theta] 0.33
31	N166_3	ribosomal protein L15. putative 50S RIBOSOMAL PROTEIN L15 0.4
32	OPFD67006	GTP-binding protein. putative GTP-binding protein. putative [Arabidopsis thaliana] Location=666939..6688340.31
33	KS664_1	hypothetical protein No NR protein Similarities
"

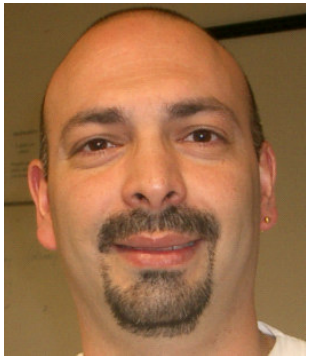
Future steps

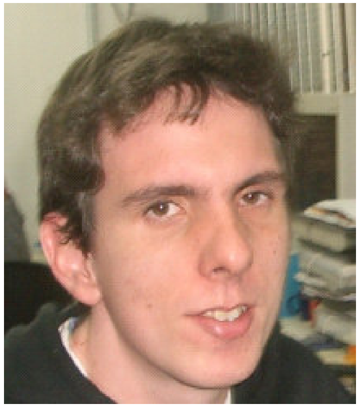
Network model generalization

- divide data in three time intervals: rings, trophozoites, schizonts
- build a network for each interval
- consider larger target sets, including unknown
- consider dependences of two or three previous times

Gene model and estimation alternatives

- Find corregulated genes by signal clustering
- create equivalent classes of corregulated genes
- gene expression depends on a linear combination of inputs
- use parallel processing





J. Barrera, R.M. Cesar Jr., C. P. Pereira, D. Martins,
R. Z. Vencio, E. F. Merino, M. M. Yamamoto

FAPESP