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]

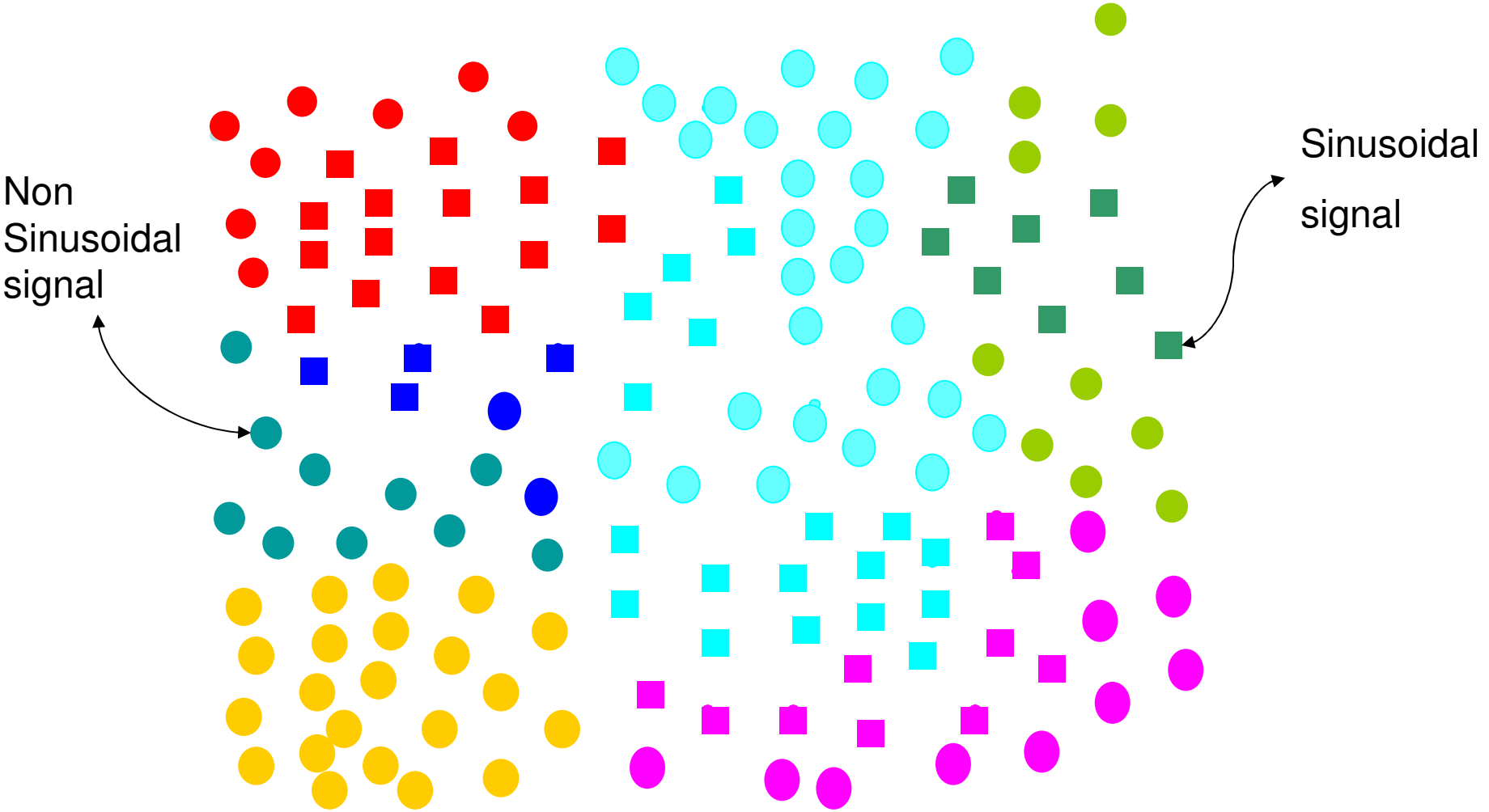
UNIVERSITY OF SAO PAULO, BRAZIL
1- Institute of Mathematics and Statistics
2- Institute of Biomedical Sciences

Layout

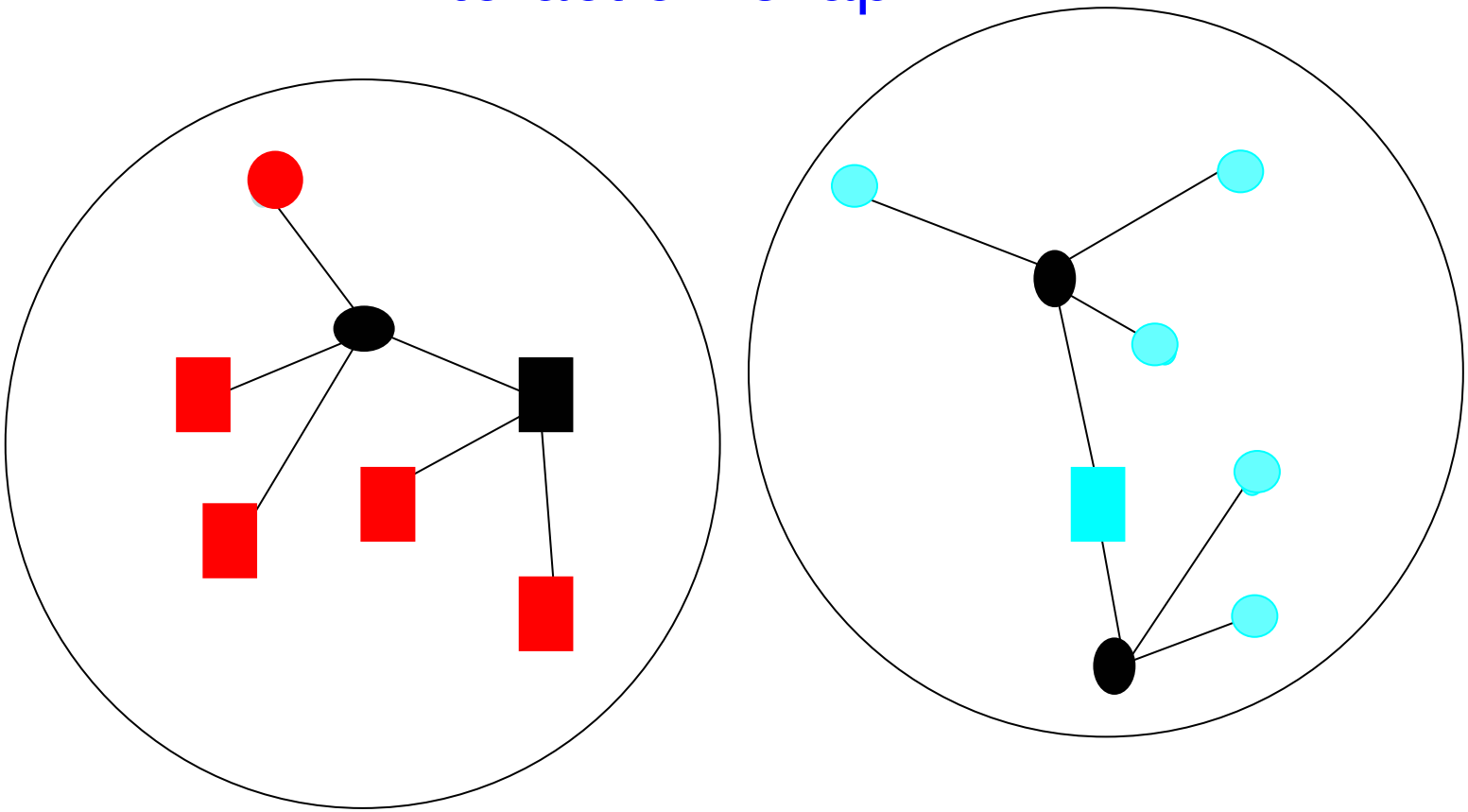
- Introduction
- Probabilistic genetic network (PGN)
- PGN design
- Data analysis pipeline
- Biological interpretation

Introduction

Functional Classification



Interaction Graph



glycolysis

plastid genome

Probabilistic Genetic Network (PGN)

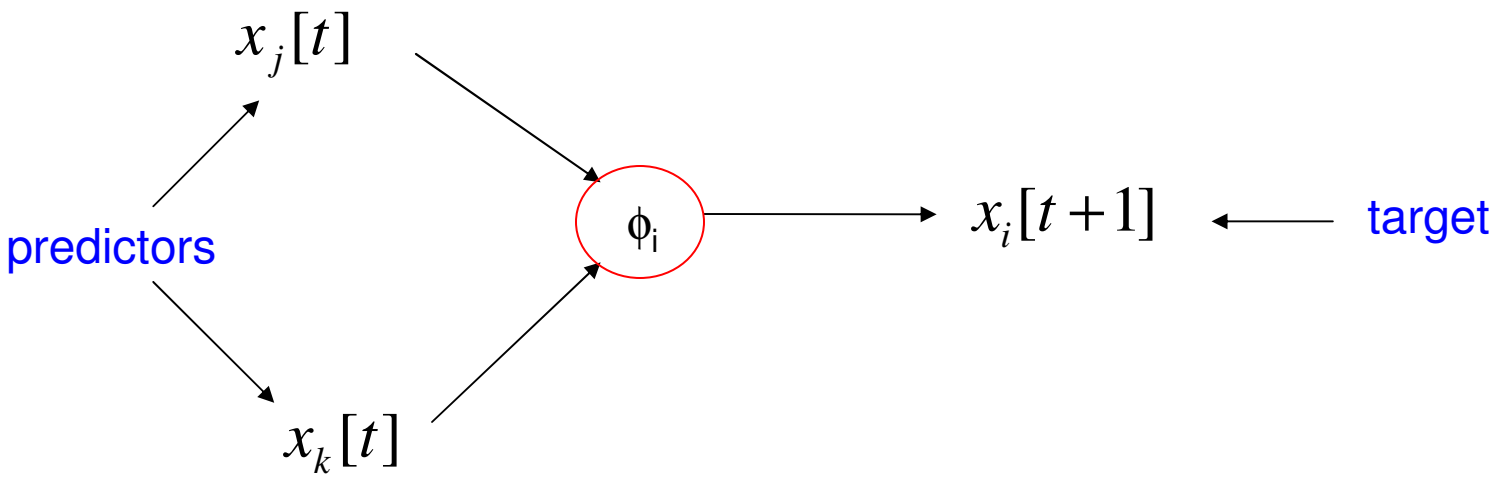
Expression of gene i at time t: $x_i[t] \in \{-1,0,+1\}$

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

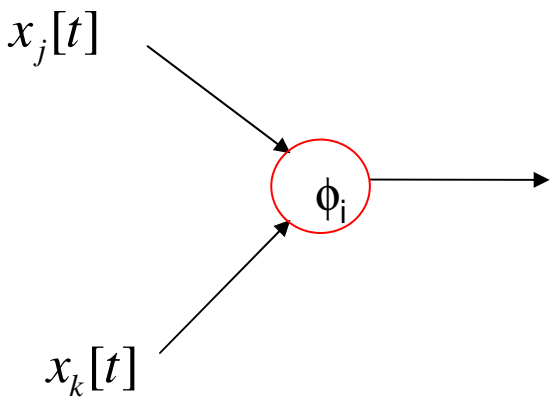
Network dynamics: $x[t+1] = \phi(x[t])$

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

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



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

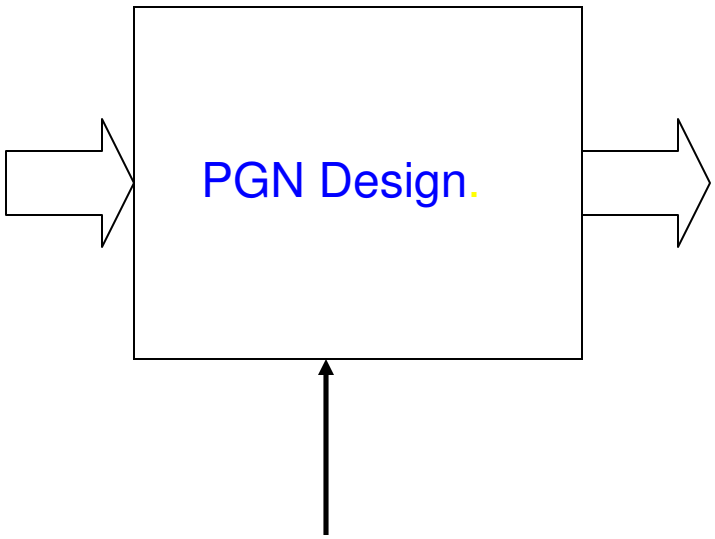
$$P(x[t+1] \mid x[t]) = \prod_{i=1}^{n} p(x_i[t+1] \mid x[t])$$

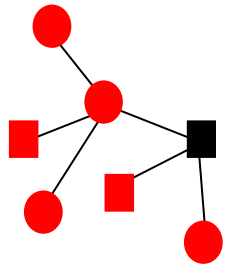
- is characterized by the conditional probabilities

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

PGN Design

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

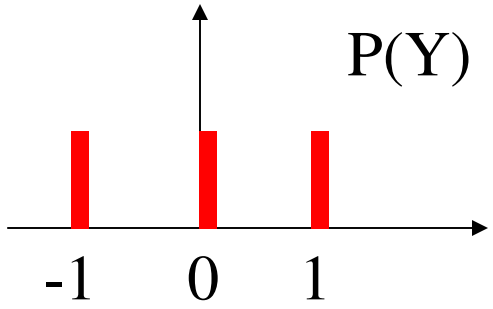




Distribution of Y

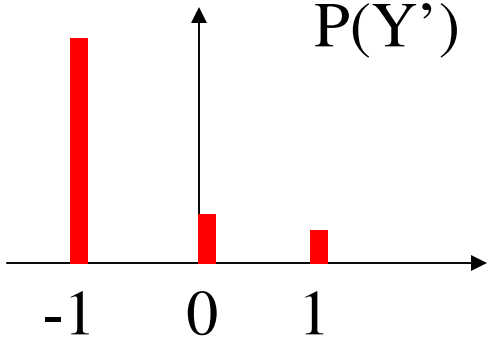
$$P : \{-1,0,1\} \rightarrow [0,1]$$

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



Entropy

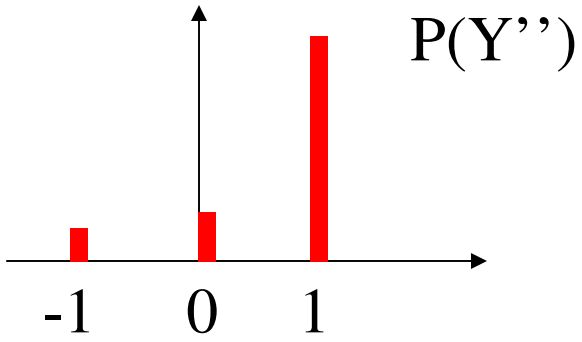
$$H(Y) = -\sum_{y \in \{-1,0,1\}} P(y) \log P(y)$$



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

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

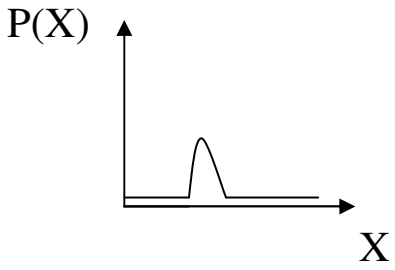
For a fixed parameter n

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

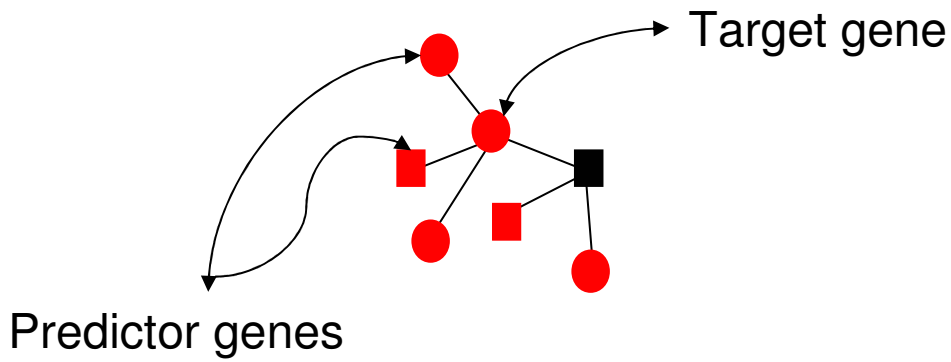
USP

Building Interactions Graphs

- For each target gene, rank all predictors by their mean estimated mutual information;

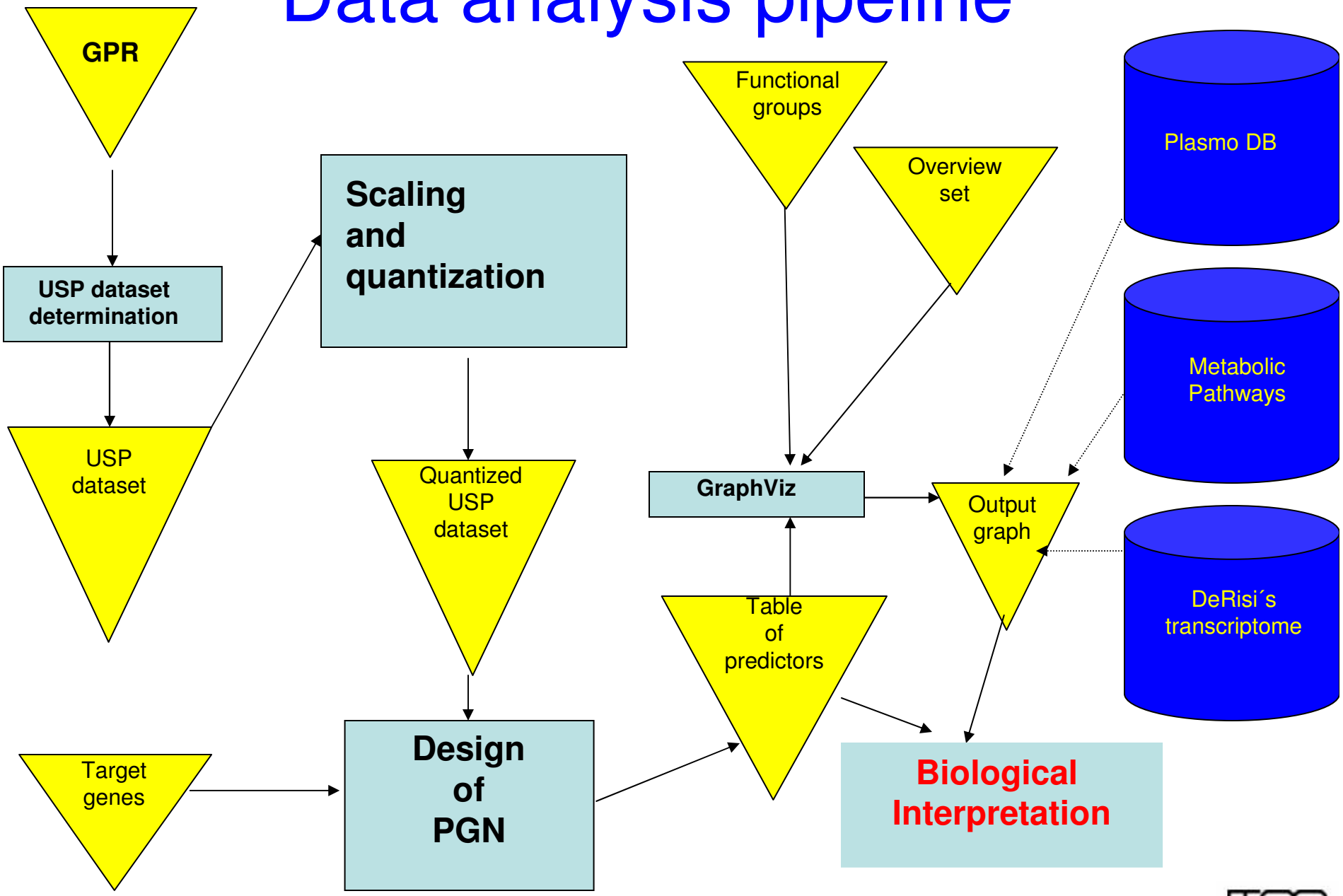
- Choose best predictors;

- Design the interaction graph



Data analysis pipeline

Data analysis pipeline



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;
- no interpolation used;
- consider ALL accepted oligos as unique entities (including almost sinusoidal).

USP-dataset: 6532 oligos

Overview dataset: 3719 oligos

Weak spots definition

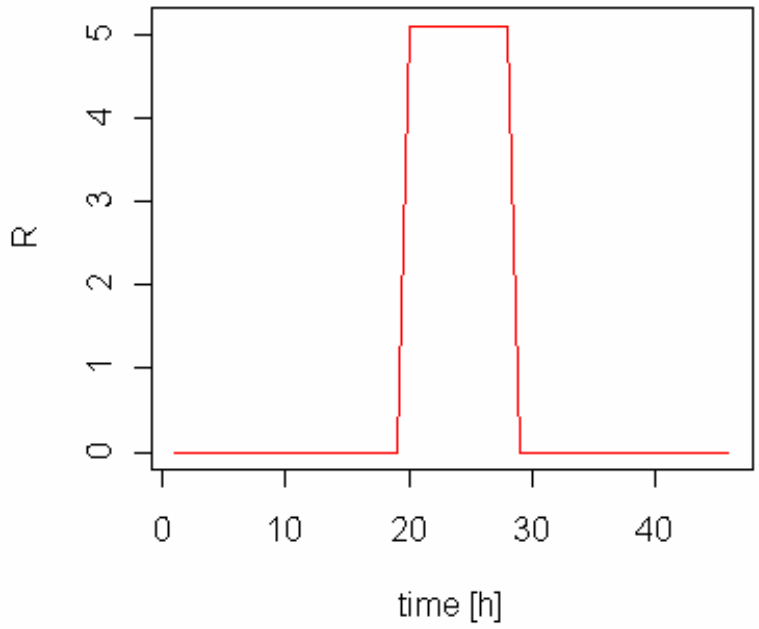
$\mathbf{X}$ = (0, 0, ... , 100, 100, ... , 100, 0, 0, ... , 0, 0)

$\langle \mathbf{X} \rangle$ = 9 * 100 / 46 = 19.56

$\mathbf{R}$ = normalized cy5/cy3 = $\mathbf{X}/\langle \mathbf{X} \rangle$ =

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

$\log_2(\mathbf{R})$ = (-∞, -∞, ... , 1.63, 1.63, ..., 1.63, -∞, -∞, ... , -∞)



Not amenable to Fourier analysis due to infinities.

Scaling

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

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

normal transform

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

Quantization

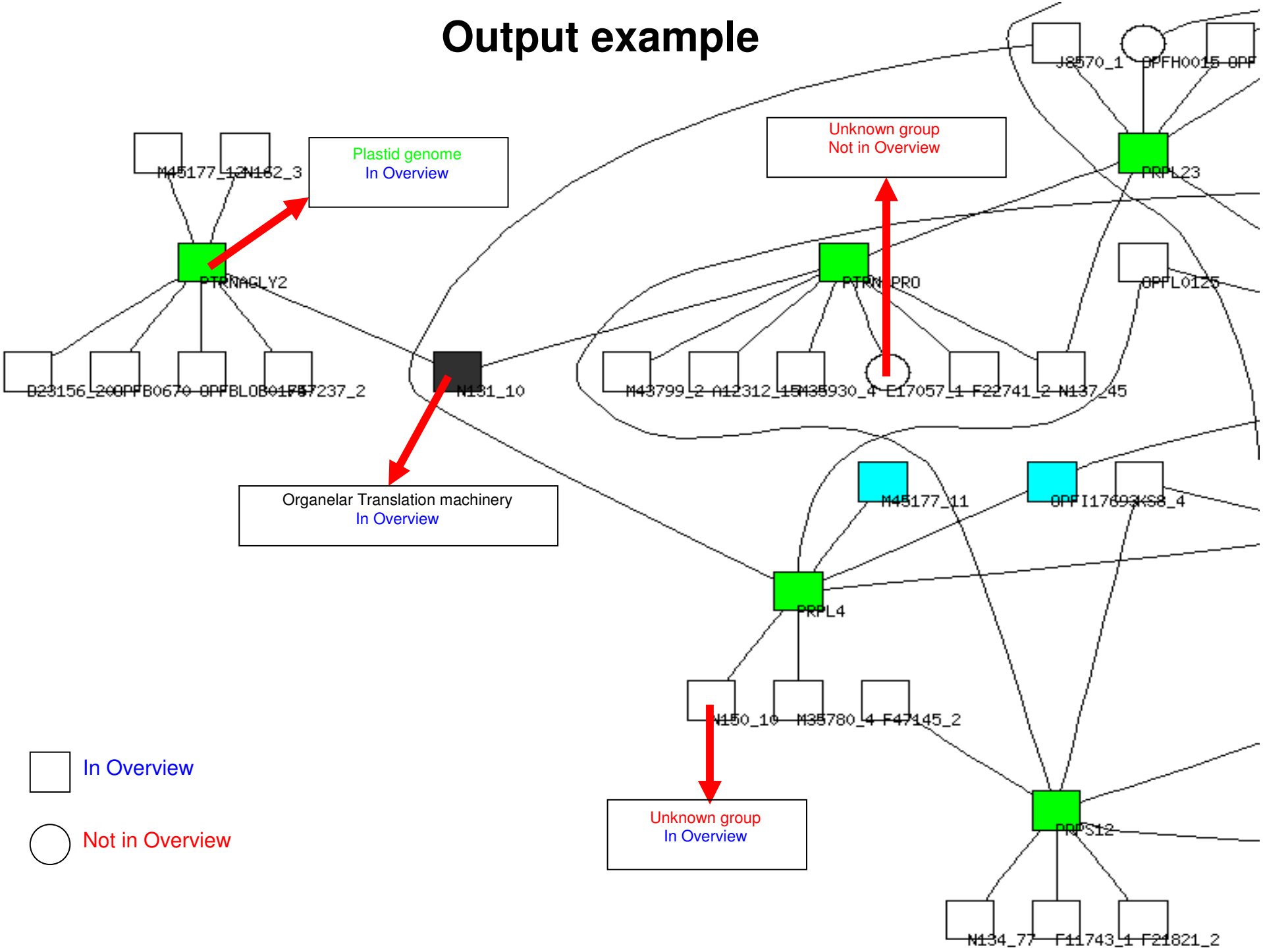
Let $n_i^+[t]$ and $n_i^-[t]$ denote, respectively, the normalized signals greater and lower than zero at t..

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 $x_i[t] = 0$

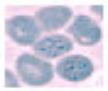
If $n_i^-[t] < \hat{E}[n_i^-[t]]$, then $x_i[t] = -1$

Output example

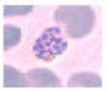


OPFB0670

PlasmoDB	Metabolic Pathway	Derisi Lab



P. falciparum PFB0330c



***Plasmodium falciparum* / CHR 2 / PFB0330c**

cysteine protease, putative

Summary view

Add this gene to your History

Annotation	Protein	Expression	Sequence
Curated Annotation	PDB structures	Microarrays	DNA (graphic)
UserComments	Structural Models	Developmental series (clone array)	Exons
GO Process	Features (graphic)	Developmental series (Affy array)	SNPs
GO Component	Pfam	Developmental series (glass slide array)	mRNA/RNA sequence
GO Function	PROSITE	Proteomics (graphic)	Protein sequence
EC number	TM domains	Mass spec. data	
RefSeqs	SignalP		
Metabolic Pathways	PlasmoAP		
MR4 Reagents	Motifs (graphic)		
Ortholog Group	Motifs		
Ortholog Views	Proteomics (graphic)		
Orthologs	Mass spec. data		
BLASTP non-Pf (graphic)			
BLASTP other (graphic)			
BLASTP NRDB			

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

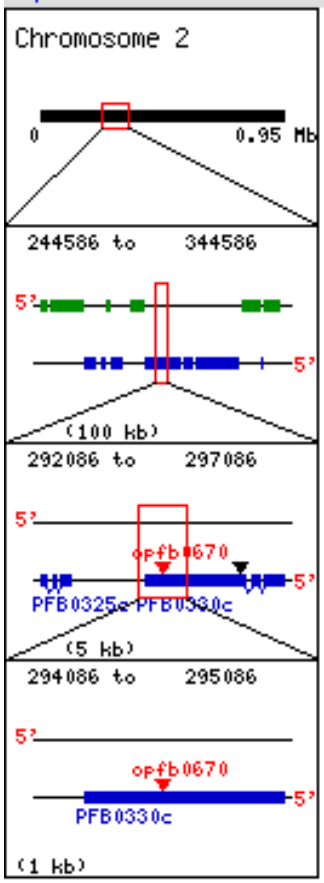
Products: cysteine protease, putative

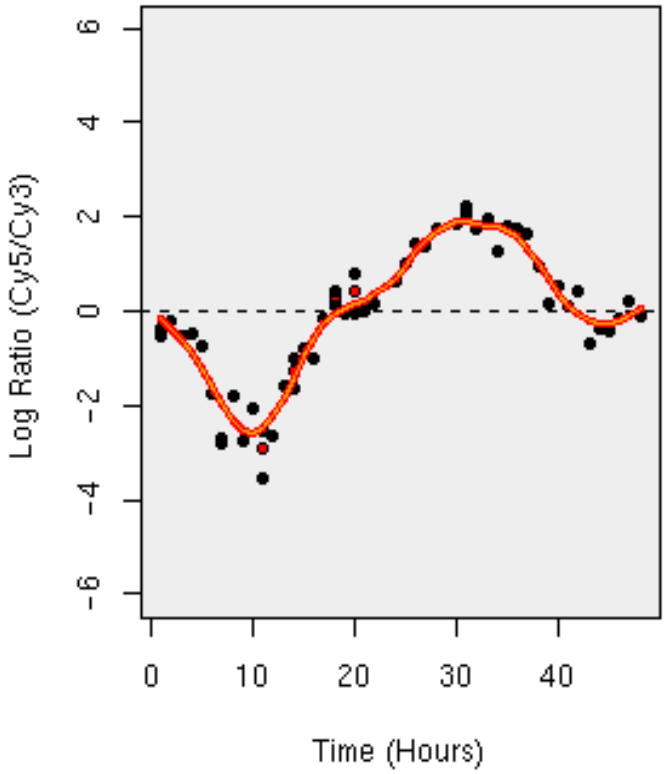
Gene-Reaction Schematic:

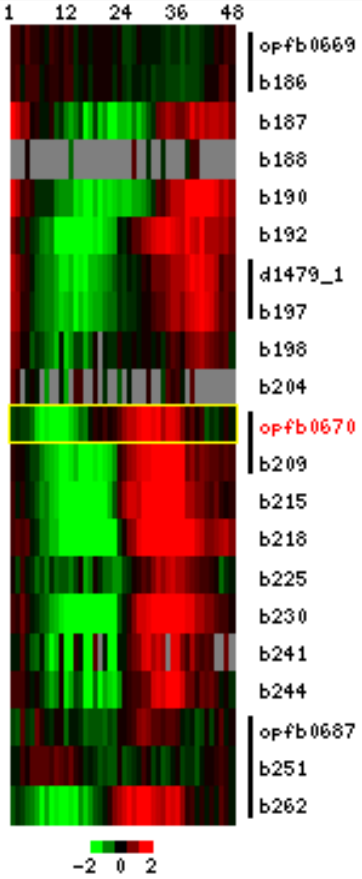
Query Page | Advanced Query Page | BioCyc Home | Report Errors or Provide Feedback

HOME | DeRisi Lab Malaria Transcriptome Database | November 1, 2004

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

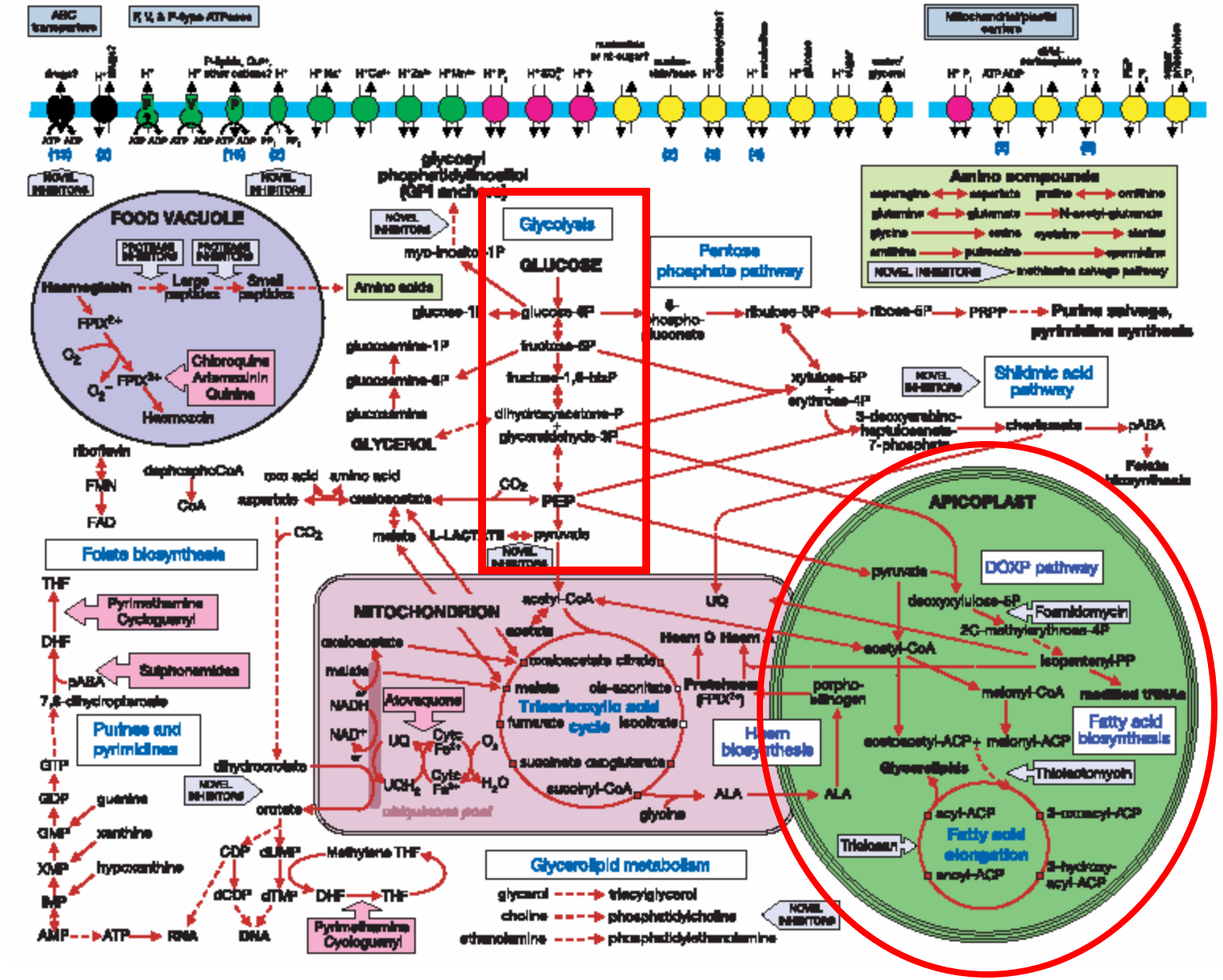
PlasmoDB ID	Description
PFB0330c	cysteine protease, putative

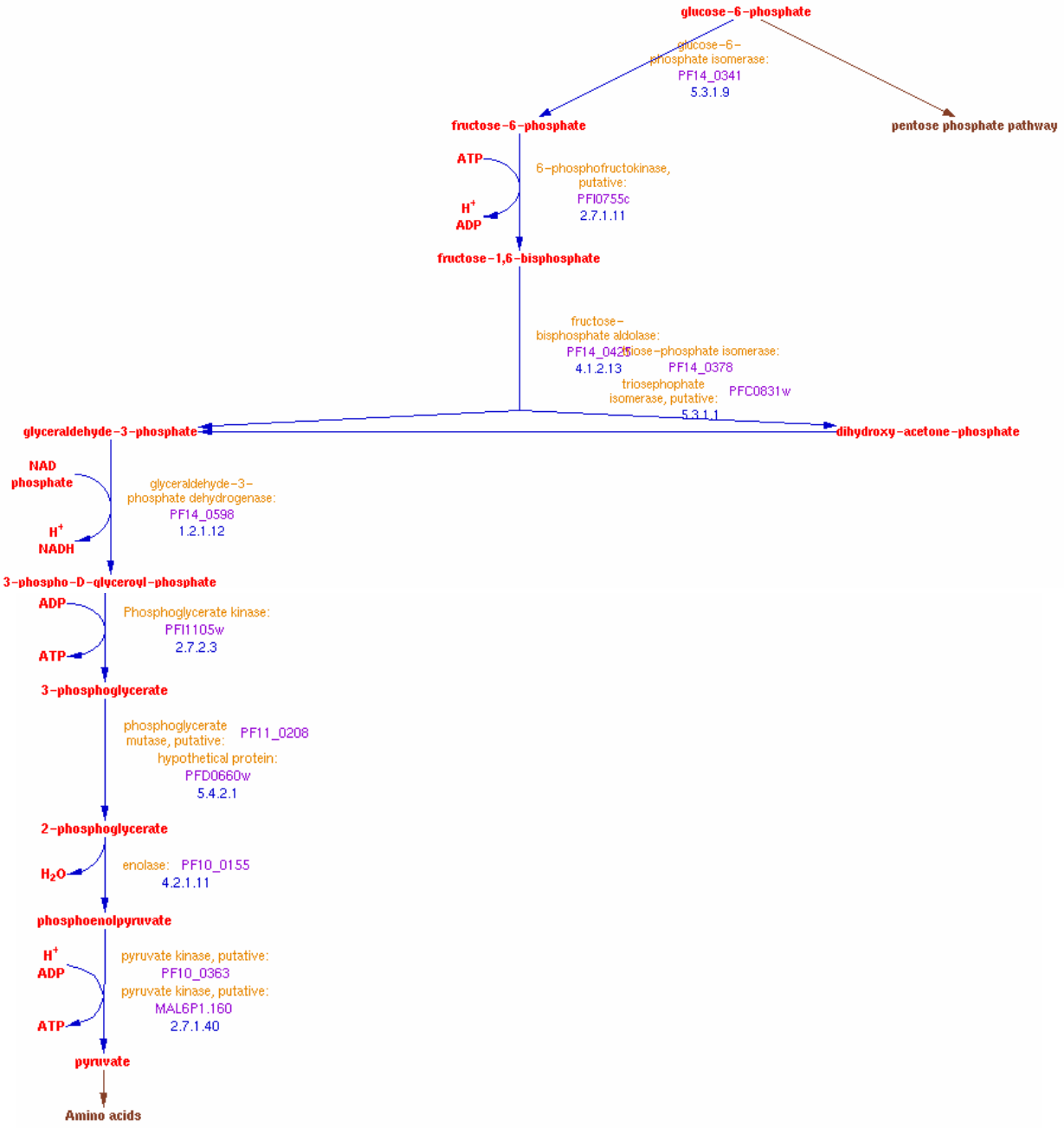
Oligo Sequence BLAST @ PlasmoDB

5'

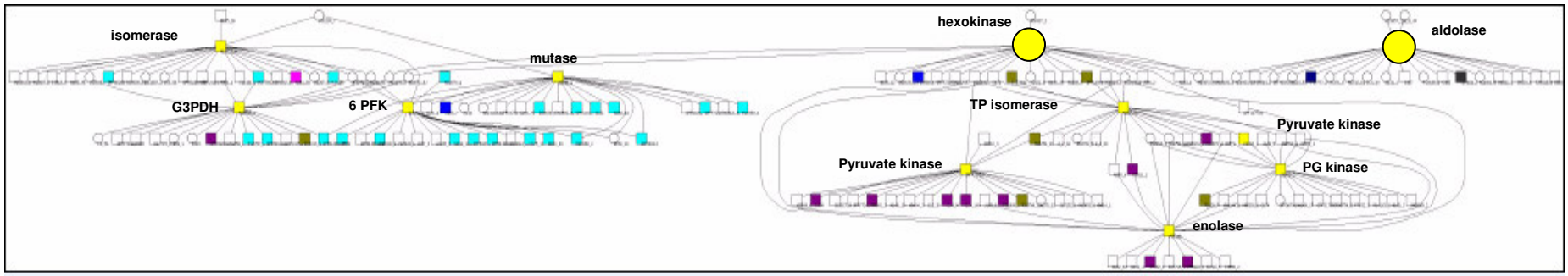
CTGCCCAAGATGAGCCACCTACTGATAATGTAGAATCACAAGCAGAAAATAACAAAAAAACAGAAATTTA

Biological Interpretation


ABC transporters
P, V, & F-type ATPases
Mitochondrial/plastid carriers
NOVEL INHIBITORS
FOOD VACUOLE
Haemoglobin
Large peptides
Small peptides
Chloroquine
Artemisinin
Quinine
Haemozoin
glycosyl phosphatidylinositol (GPI anchor)
Glycolysis
GLUCOSE
Amino acids
Pentose phosphate pathway
Amino compounds
Shikimic acid pathway
Purine salvage, pyrimidine synthesis
GLYCEROL
riboflavin
FMN
FAD
CoA
Folate biosynthesis
Pyrimethamine
Cycloguanyl
Sulphonamides
pABA
Purines and pyrimidines
NOVEL INHIBITORS
MITOCHONDRION
Tricarboxylic acid cycle
Atovaquone
Heme biosynthesis
APICOPLAST
DOXP pathway
Fosmidomycin
Fatty acid biosynthesis
Thiolactomycin
Triclosan
Fatty acid elongation
Glycerolipid metabolism
glycerol → triacylglycerol
choline → phosphatidylcholine
ethanolamine → phosphatidylethanolamine


glucose-6-phosphate
glucose-6-phosphate isomerase:
PF14_0341
5.3.1.9
fructose-6-phosphate
pentose phosphate pathway
ATP
H+
ADP
6-phosphofructokinase, putative:
PFI0755c
2.7.1.11
fructose-1,6-bisphosphate
fructose-bisphosphate aldolase:
PF14_0425
4.1.2.13
triosephophate isomerase:
PF14_0378
triosephophate isomerase, putative:
PFC0831w
5.3.1.1
glyceraldehyde-3-phosphate
dihydroxy-acetone-phosphate
NAD
phosphate
H+
NADH
glyceraldehyde-3-phosphate dehydrogenase:
PF14_0598
1.2.1.12
3-phospho-D-glyceroyl-phosphate
ADP
ATP
Phosphoglycerate kinase:
PFI1105w
2.7.2.3
3-phosphoglycerate
phosphoglycerate mutase, putative:
PF11_0208
hypothetical protein:
PFD0660w
5.4.2.1
2-phosphoglycerate
H2O
enolase:
PF10_0155
4.2.1.11
phosphoenolpyruvate
H+
ADP
ATP
pyruvate kinase, putative:
PF10_0363
pyruvate kinase, putative:
MAL6P1.160
2.7.1.40
pyruvate
Amino acids

Glycolytic PGN network (single genes)



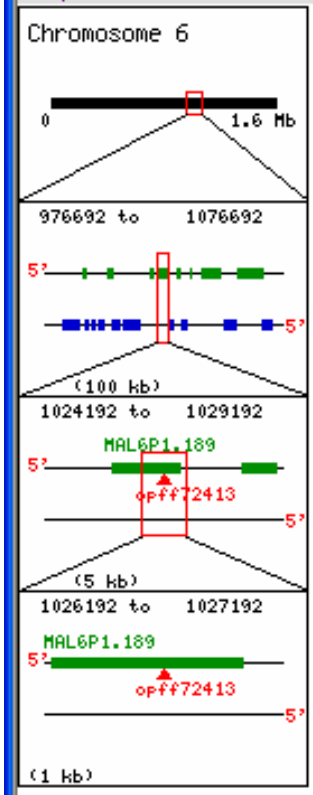
- glycolysis
- transcription machinery
- cytoplasmic translation
- ribonucleotide synthesis
- deoxynucleotide synthesis
- DNA replication
- proteoasome
- plastid genome
- kinases
- actin myosoin motors
- mitochondrial

HOME

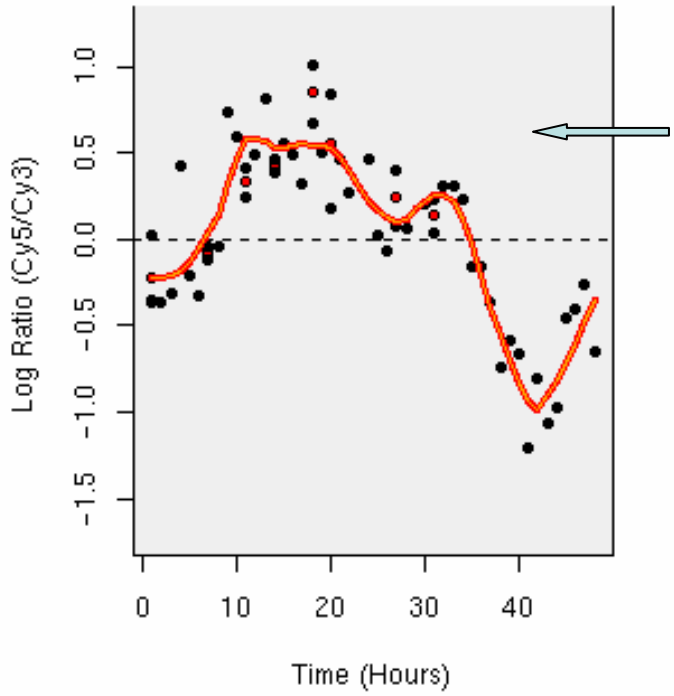
DeRisi Lab Malaria Transcriptome Database

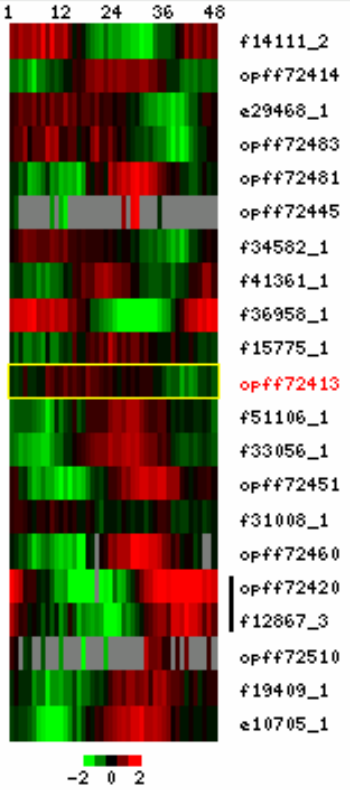
November 11, 2004

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



OLIGO





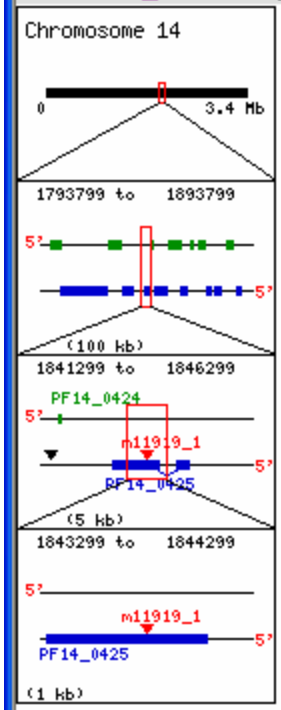
PlasmoDB ID	Description
MAL6P1.189	hexokinase

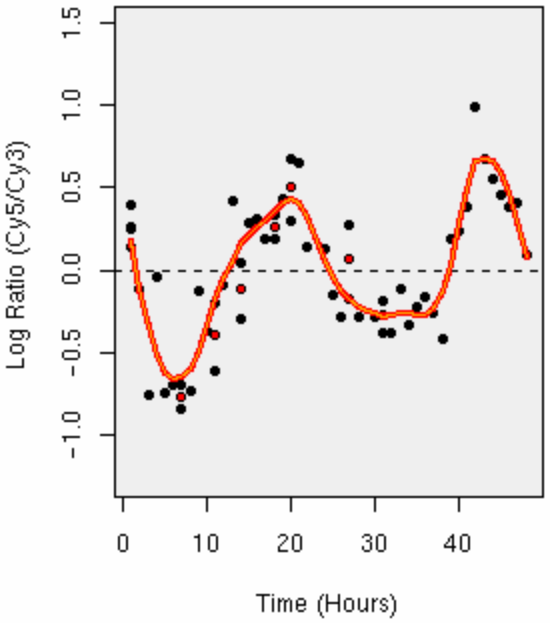
HOME

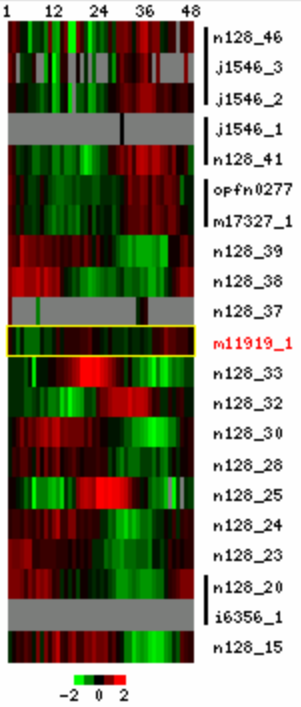
DeRisi Lab Malaria Transcriptome Database

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





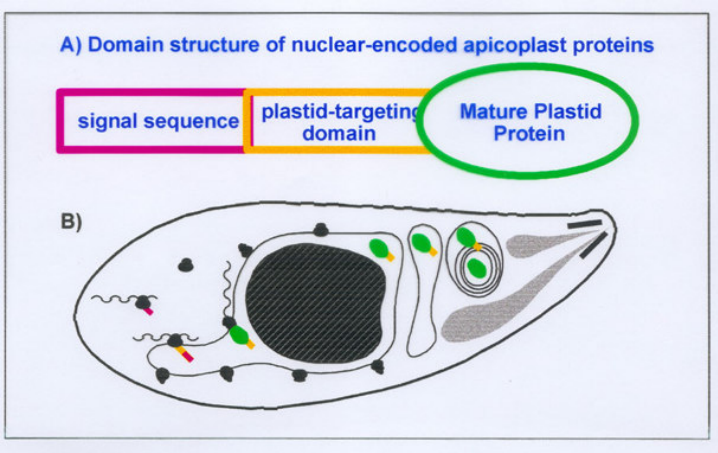


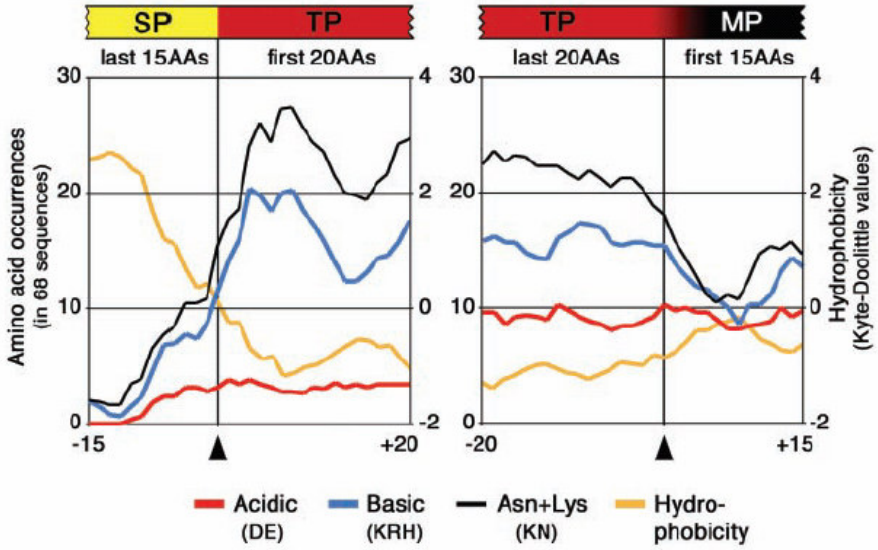
OLIGO

PlasmoDB ID	Description
PF14_0425	fructose-bisphosphate aldolase

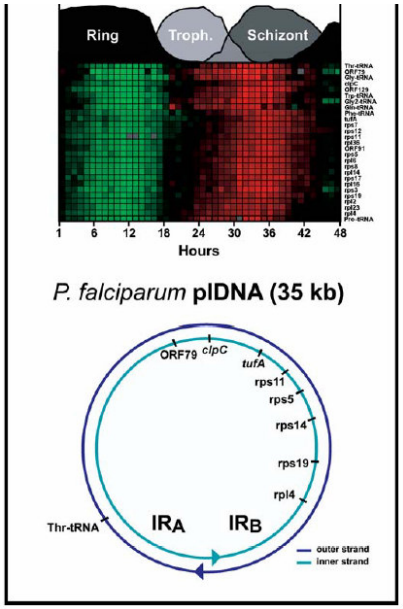
No TCA genes

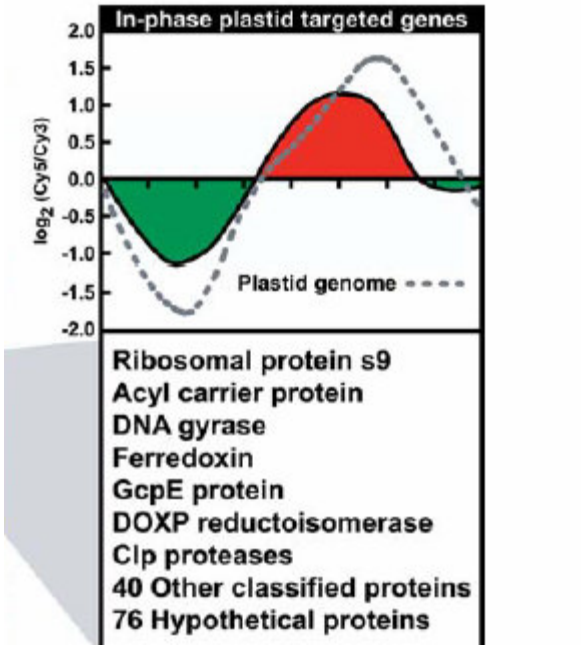
25	N132_136	D33539_15	hypothetical protein
26	N132_136	D11687_1	
27	N132_136	J53_56	3.8 protein No NR protein Similarities
28	N132_136	N151_50	
29	N132_136	OPFF72422	
30	N132_136	OPFBLOB0090	methionine aminopeptidase. putative methionine aminopeptidase; Map1p 0.51"
31	N132_136	OPFL0114	hypothetical protein (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 predi
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



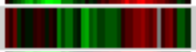


550 apicoplast proteins

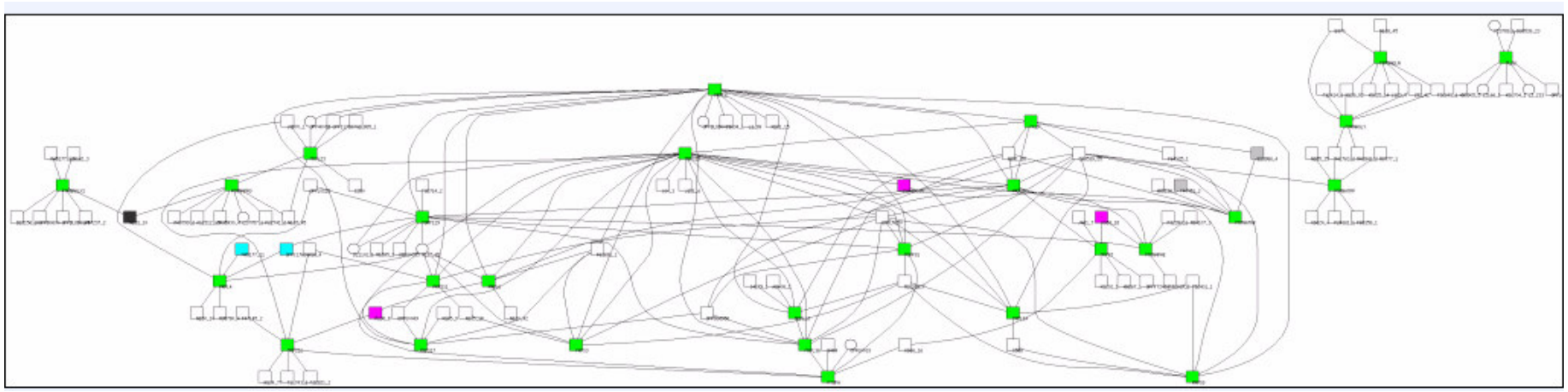




124 apicoplast proteins

Functional Group: plastid genome			
PlasmoDB ID	Oligo ID	IDC Expression Profile	PlasmoDB Annotation
Clp	pclp		plastid encoded Clp protease
LSUrRNA	plsu		plastid large subunit rRNA
ORF129	porf129		hypothetical, plastid encoded
ORF91	porf91		hypothetical, plastid encoded
rpl14	prpl14		plastid ribosomal protein 14, large subunit
rpl16	prpl16		plastid ribosomal protein 16, large subunit
rpl2	prpl2		plastid ribosomal protein 2, large subunit
rpl23	prpl23		plastid ribosomal protein 23, large subunit
rpl36	prpl36		plastid ribosomal protein 36, large subunit
rpl4	prpl4		plastid ribosomal protein 4, large subunit
rpl6	prpl6		plastid ribosomal protein 6, large subunit
rps11	prps11		plastid ribosomal protein 11, small subunit
rps12	prps12		plastid ribosomal protein 12, small subunit
rps17	prps17		plastid ribosomal protein 17, small subunit
rps19	prps19		plastid ribosomal protein 19, small subunit
rps3	prps3		plastid ribosomal protein 3, small subunit
rps5	prps5		plastid ribosomal protein 5, small subunit
rps7	prps7		plastid ribosomal protein 7, small subunit
rps8	prps8		plastid ribosomal protein 8, small subunit
PtRNA-Gln	ptrnagln		plastid tRNA-Gln
PtRNA-Gly	ptrnagly		plastid tRNA-Gly
PtRNA-Gly2	ptrnagly2		plastid tRNA-Gly2
PtRNA-Phe	ptrnaphe		plastid tRNA-Phe
PtRNA-Pro	ptrnapro		plastid tRNA-Pro

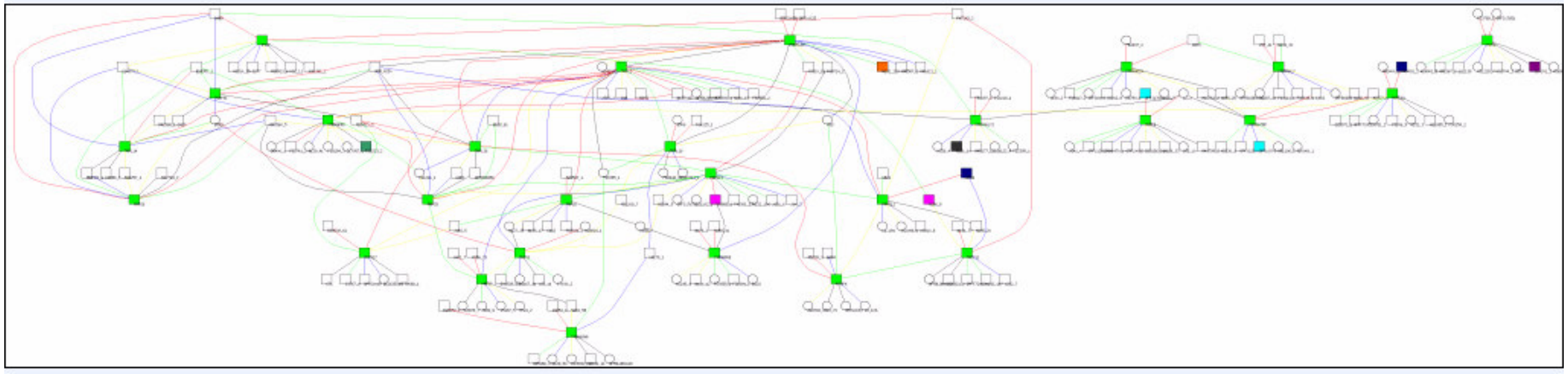
Apicoplast PGN network (singlets)



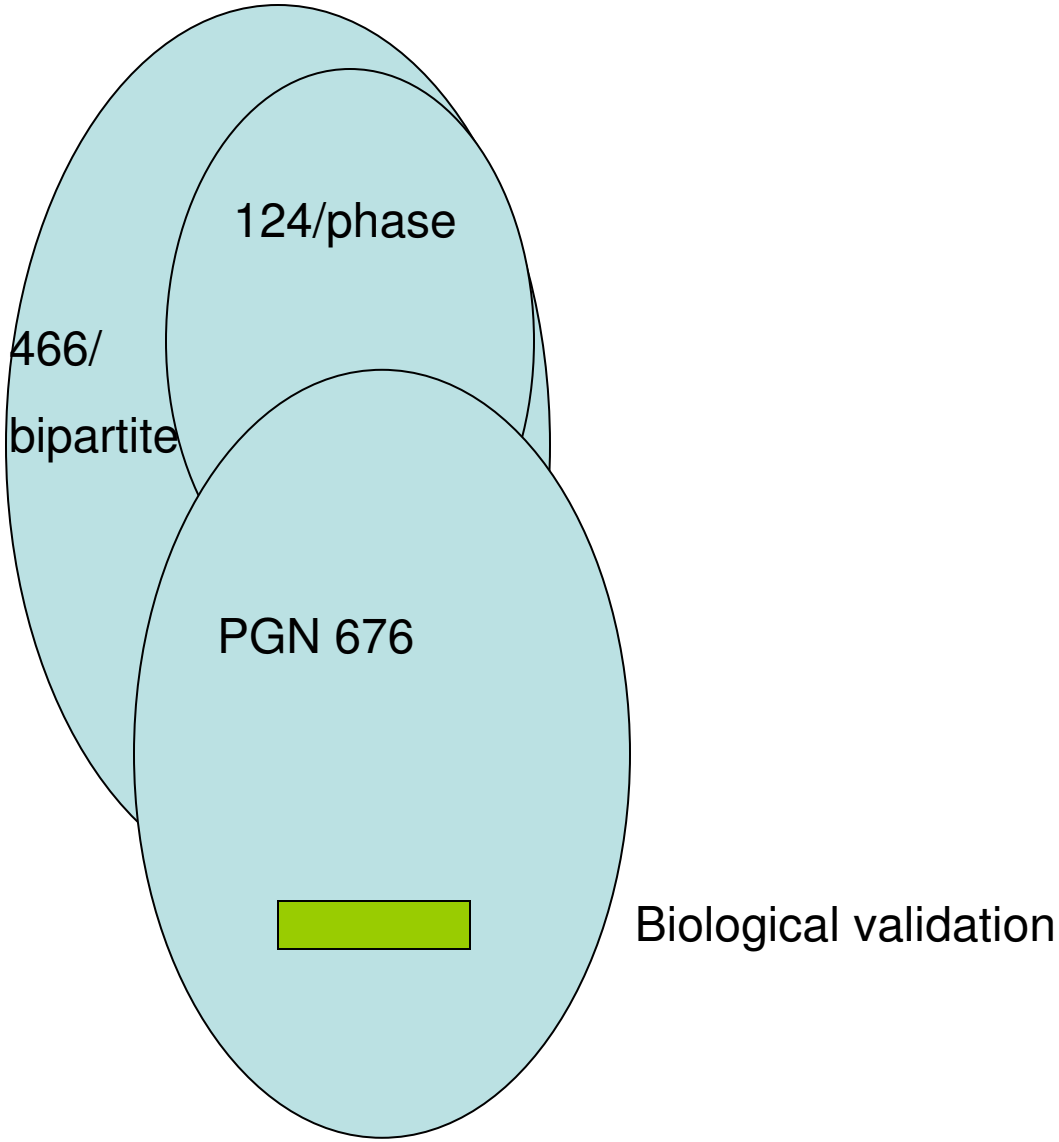
- glycolysis
- transcription machinery
- cytoplasmic translation
- ribonucleotide synthesis
- deoxynucleotide synthesis
- DNA replication
- proteoasome
- plastid genome
- kinases
- actin myosoin motors
- mitochondrial

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

Apicoplast PGN network (doublets)

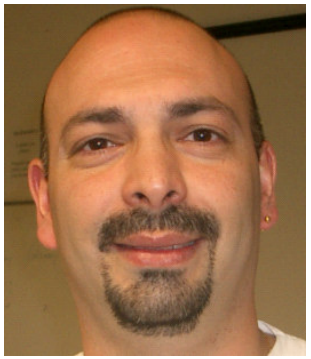


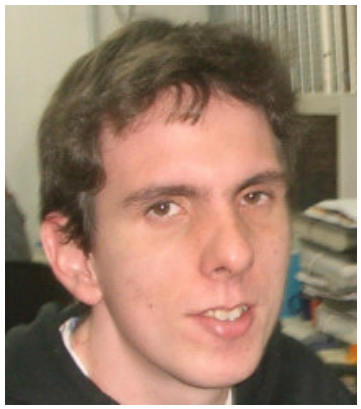
- glycolysis
- transcription machinery
- cytoplasmic translation
- ribonucleotide synthesis
- deoxynucleotide synthesis
- DNA replication
- proteoasome
- plastid genome
- kinases
- actin myosoin motors
- mitochondrial


124/phase
466/
bipartite
PGN 676
Biological validation







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

FAPESP