# Predicting Regulatory Elements in *P. falciparum*


**Chengyong Yang**

**Erliang Zeng**

**Giri Narasimhan**

**BioInformatics Research Group (BioRG)**
**School of Computer Science**
**Florida International University, Miami, FL.**

**Kalai Mathee**

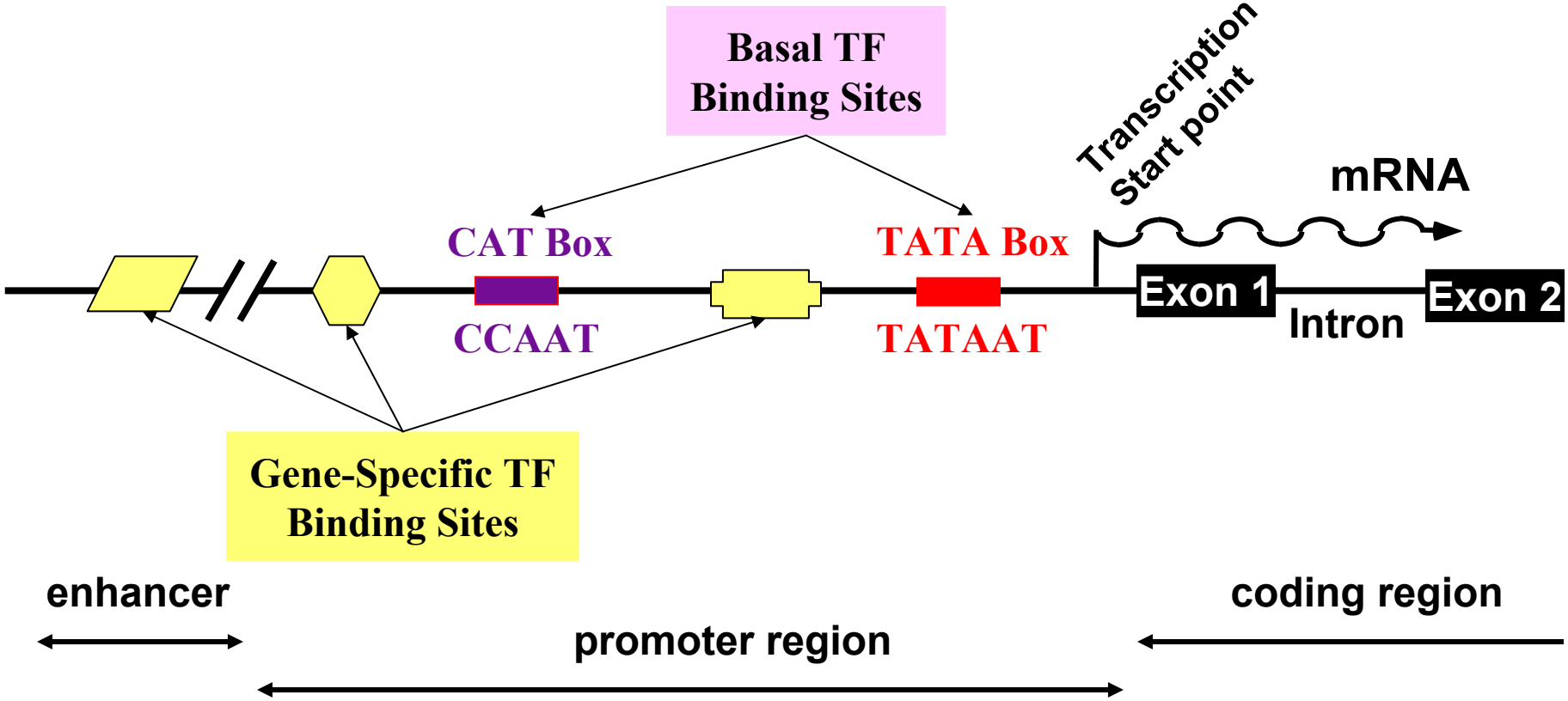
**Department of Biological Sciences**
**Florida International University , Miami, FL.**


BIORG
BioInformatics Research Group

# Outline

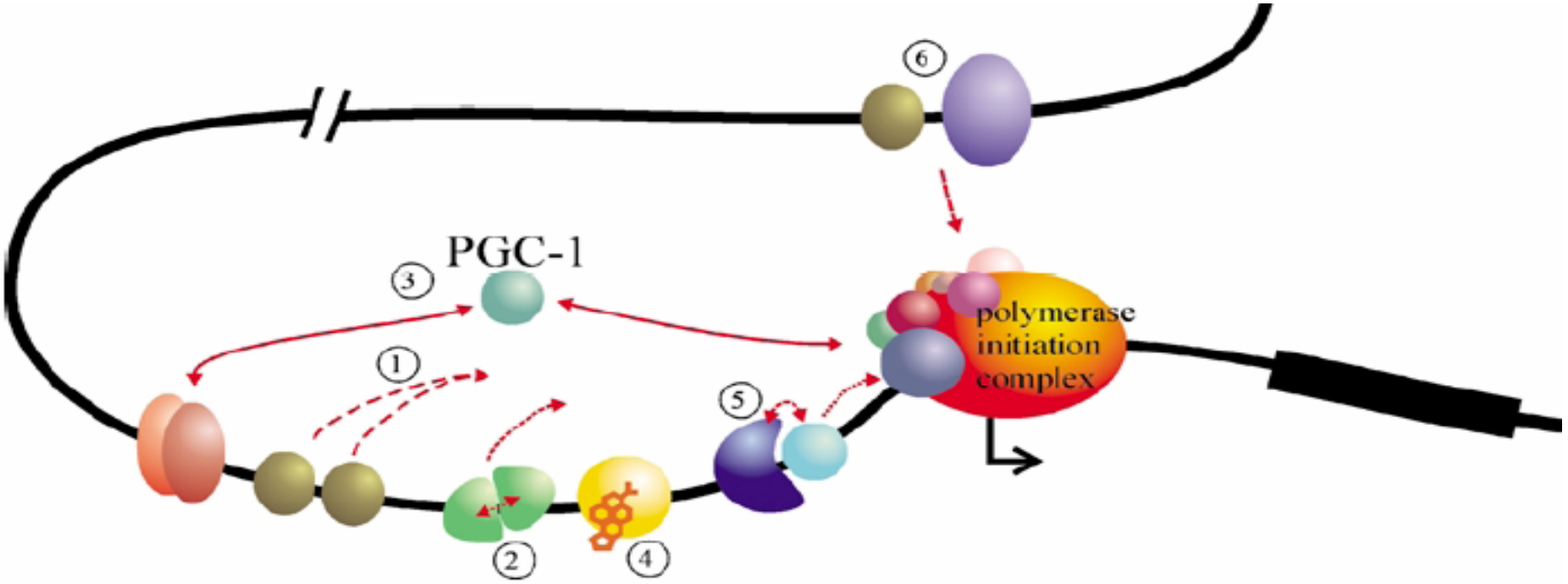
- Biology of Transcription Regulation
- Mining Regulatory Elements (or Transcription Factor Binding Motifs or TFBMs)
- Experimental Results
- Conclusion
- Future Work

# Transcription Regulation



# Transcription Regulation

[ Goffart *et al. Exp. Physiology* (2003) ]



# Outline

- Biology of Transcription Regulation
- Mining Regulatory Elements (or Transcription Factor Binding Motifs or TFBMs)
- Experimental Results
- PlasmoTFBM database & Web Query Interface
- Conclusion
- Future Work

# Transcription Factor Binding Motifs (TFBM)

- **Why look for TFBMs?**
  - **Which TFs regulate a specific gene?**
  - **Which genes are co-regulated by same TF?**
  - **Understand strength of gene expression.**
  - **Understand gene regulatory pathways.**

# How to Find TF Binding Motifs?

## Direct Experimental Assays

- Electrophoretic mobility shift assay
- Nuclease protection assay

Need to know the TFs.

## Computational Methods

- Search for known motifs
- Predict sites based on pattern discovery in upstream sequences

Need to know the TFBMs

Only need to know the upstream sequences

# CAMDA Data Set

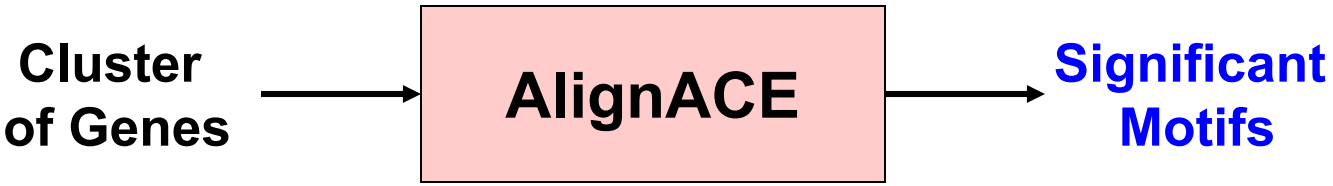
- Microarray data from DeRisi lab
- 46 data sets for a 48 hour time period for *P. falciparum* during the **intraerythrocytic** development life cycle.
- During the 48 hour period, *P. falciparum* goes through 4 stages:
  - **Ring** (1-15 hpi)
  - **Trophozoite** (16-28 hpi)
  - **Schizont** (29-42 hpi)
  - **Merozoite** (43-48 hpi)

# Broad Questions Raised

- Are there transcriptional events that distinguish the 4 stages of the organism?
- Are there functional similarities in the genes that share motifs?

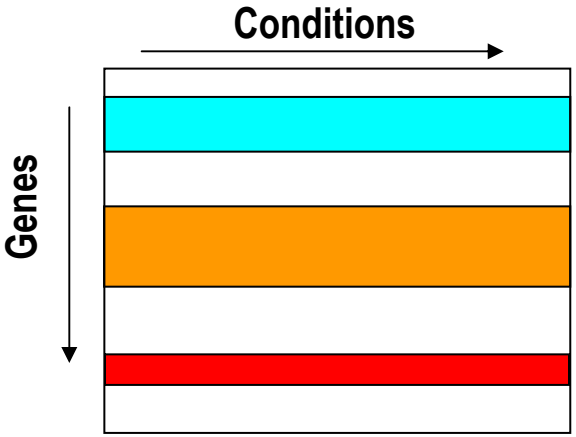
# AlignACE

[ Roth *et al.* Nature Biotechnology (1998) ]



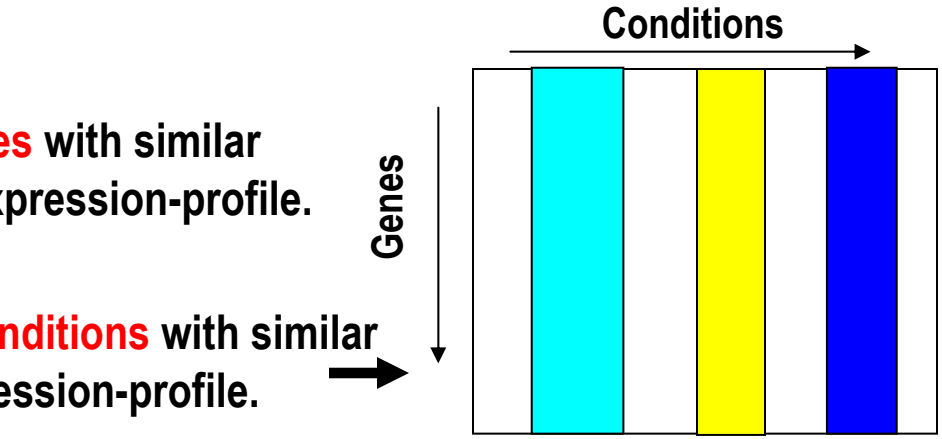

- Uses Gibbs Sampling to find good alignments of upstream sequences.
- Maximizes relative entropy to find significant motifs.
- Significant motifs: must over-represent in the input set and must have small probability of occurring by chance.

# Clustering of Samples or/and Genes

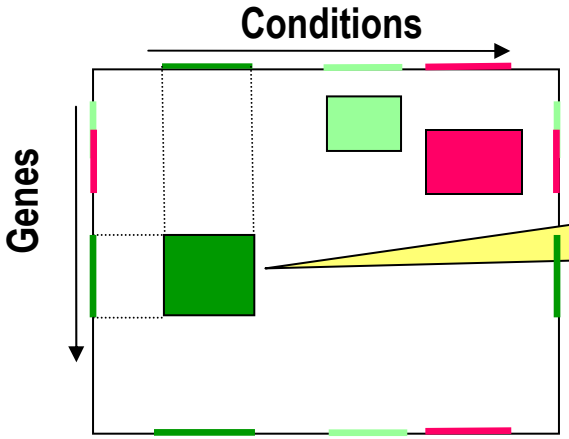



Cluster genes with similar condition expression-profile.

Cluster conditions with similar gene expression-profile.




Cluster genes and conditions simultaneously




Transcription Module



BIORG
BioInformatics Research Group

# Transcription Module

Transcription Module: a set of genes G and a set of conditions C such that the genes in G are co-regulated under conditions C.

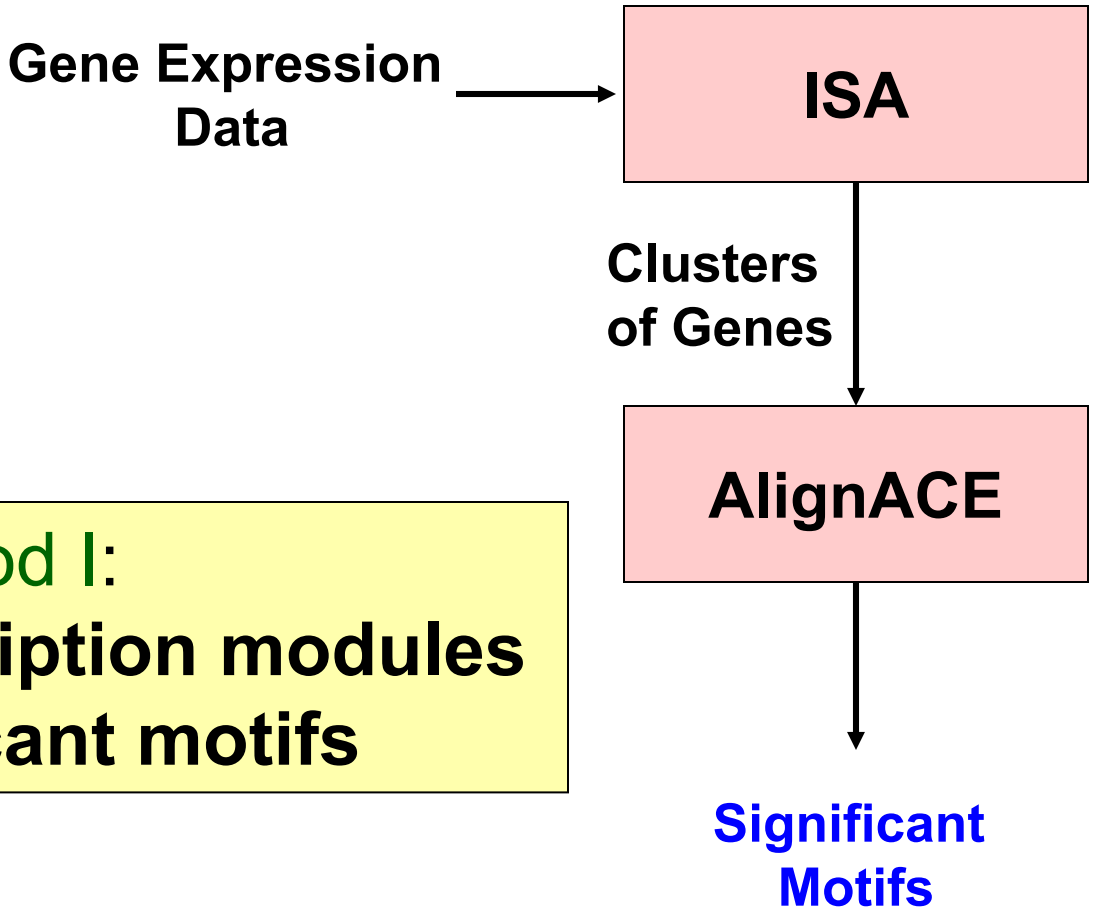
[ Ihmel *et al.* Nature Genetics (2002) ]

## Iterative Signature Algorithm (ISA)

- Defines the score of a set of genes and conditions.
- Iteratively refines the set of genes and conditions until a "stable" transcription module is obtained.

[ Ihmel *et al.* Bioinformatics (2004) ]

# Predicting TFBMs: Method I




Using Method I:

**106 transcription modules**

**840 significant motifs**

# Strength of TFBMs

- TFs bind to DNA in sequence-specific manner.
- If the motif is “strong”, then the binding is strong and the regulation is strong.
- Correlation between gene expression and the strength of its upstream TFBMs.
- MotifRegressor [Conlon *et al.*, PNAS 2003] exploits this correlation.

# Motif Regressor

[Conlon *et al.*, PNAS, 2003]

1. Rank all genes by expression and obtain upstream sequences of highly ranked genes.
2. Use MDscan to find motifs from most induced and most repressed genes.
3. Score each upstream sequence for matches to each MDscan reported motif.
4. Perform linear regression between motif matching score and gene expression and identify significant motifs.

**46 separate runs of MotifRegressor resulted in 637 significant motifs.**

# PlasmoTFBM Database

- All results were put into a searchable MySQL database containing:
  - Modules
  - Motifs
  - Gene Annotation information
  - Gene Expression data
  - Upstream sequence data
  - Miscellaneous data

# Outline

- Biology of Transcription Regulation
- Mining Regulatory Elements (or Transcription Factor Binding Motifs or TFBMs)
- Experimental Results
- Conclusion
- Future Work

# Results

A. Validation of known motifs
   1. G-Box motif
   2. *var* gene family
B. Motif clusters & motif-stage correlations
C. All Motifs in single gene of interest
D. Gene Family Analysis (*SERA* genes)

# A: G-Box Motifs

- *P. falciparum* genome is AT-rich (15% GC)
- G-box: a unique regulatory element
- Identified in upstreams of heat shock proteins (*hsp*).

**Published Motif**

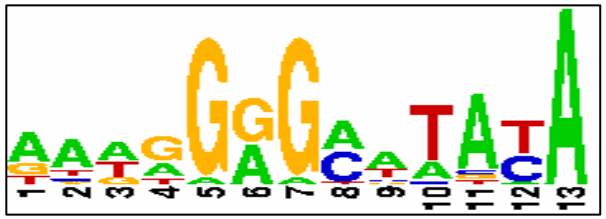
[Militello *et al.*, MBP, 2004]

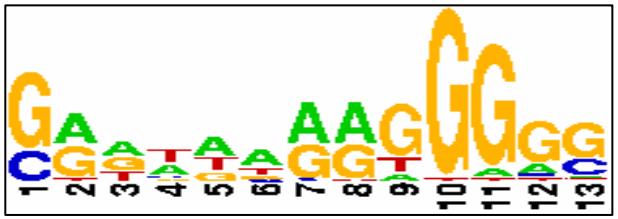
(A/G)N**GGGG**(C/A)

# A: G-Box Motifs

(A/G)NGGGG(C/A) [Militello *et al.*, MBP, 2004]

**G-Box from PlasmoTFBM**





**TG-box**

New

# A: *var* Gene Family

[Voss *et al.*, Mol Microbiol, 2003]

- 50 diverse *var* genes
- Coding for variants of *P. falciparum* erythrocyte membrane protein 1 (PfEMP1)
- Ability to switch the expression of PfEMP1
- Allows the parasite to escape specific immune responses

# A: Significant Motifs in *var* Genes

[Voss *et al.*, Mol Microbiol, 2003]

## SPE2

TGTGCATAGTG

## Repressed 38 hpi

PF08_010 PFL0935c
PF10_040 PFB0010w
PFI1830c PFA0765c

## CPE

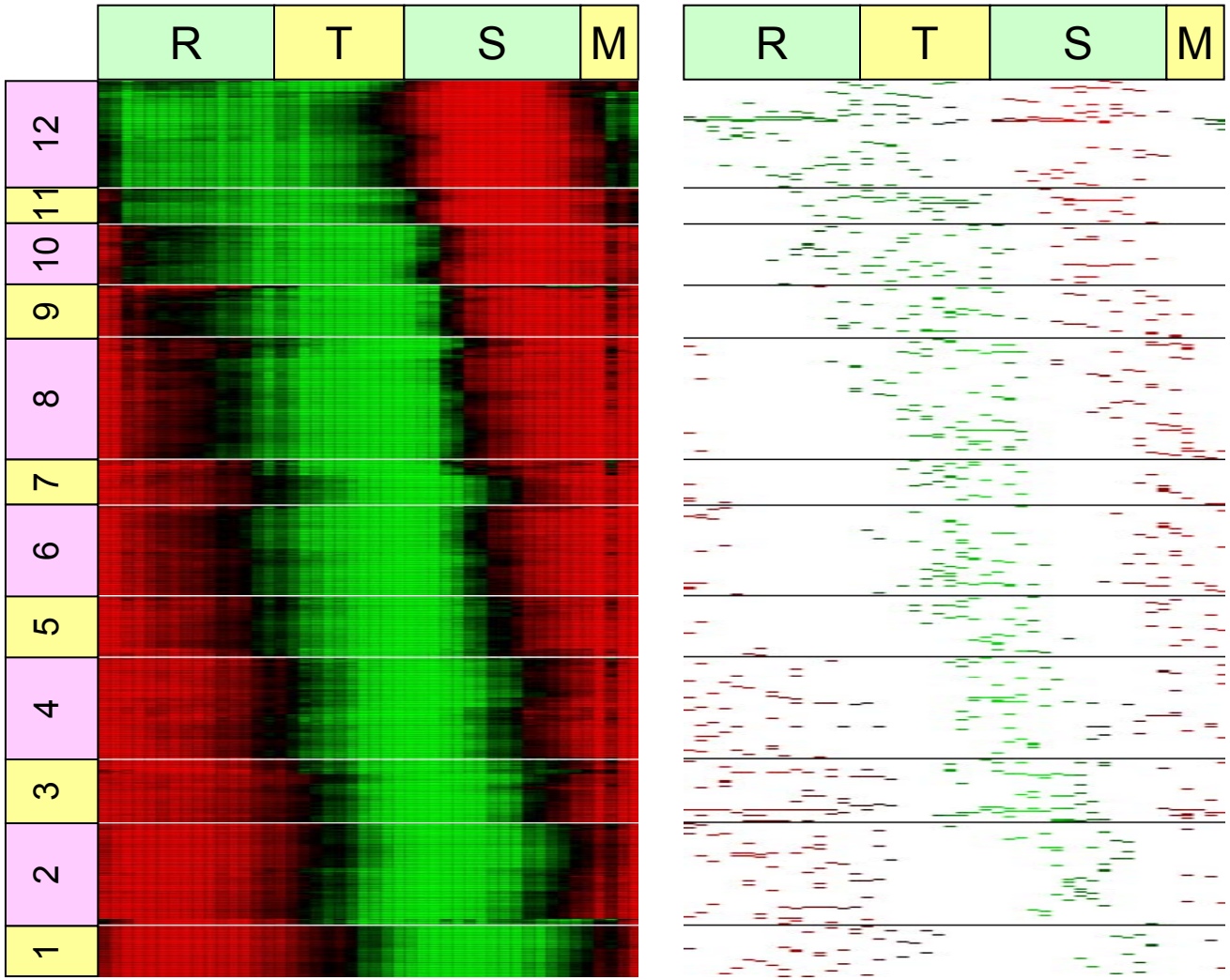
ATGTTGTACAT

## Induced 11 hpi

PFL0935c PF14_048
PFB0010w PF08_0103
PFI1830c PF10_0406
PFL1955w PFA0765c
PFD0615c

# B: Motif Clusters

[Genesis, http://genome.tugraz.at/Software]



# C: *EBA140*

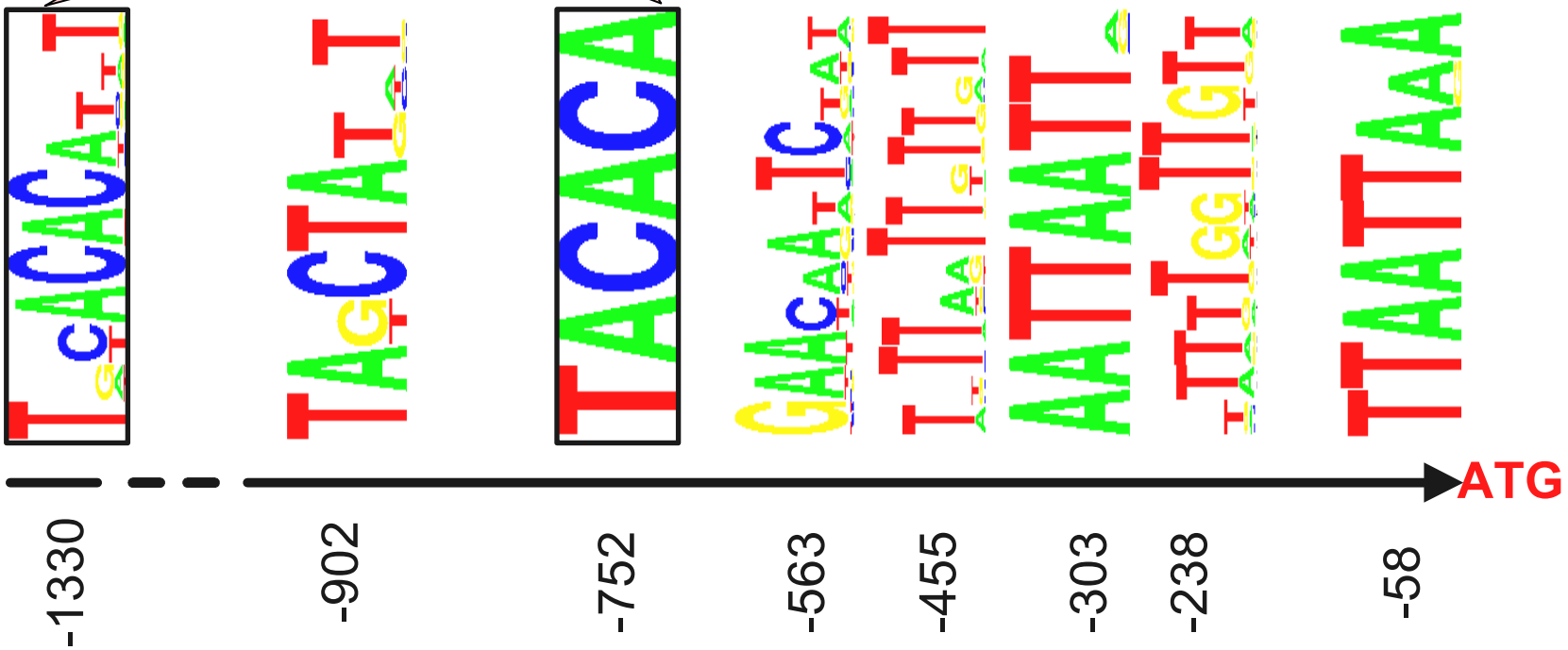
[Thompson *et al.*, Mol Microbiol, 2001]

- **EBA140 is implicated in merozoite invasion on erythrocytes**
- **Putative vaccine target**
- **Share sequence homology and structural features with EBA175**

# C: Motifs Found in *EBA140*

Shared by 77 genes including MAL7P1.86 (TF)

**Correction:**
**Figure 4, Pg 19 of Abstracts**






BIORG
BioInformatics Research Group

# D: *SERA* Gene Family

[Miller *et al.*, JBC, 2002]

- Serine repeat antigen (*SERA*)
- Adjacent, co-regulated genes from Chr 2
- Highly expressed in late blood cycle
- Target of protective immune response
- Possesses a protease function domain
- Serves both as a vaccine and a drug target to control *P. falciparum*

# D: Motif Discovered in *SERA*

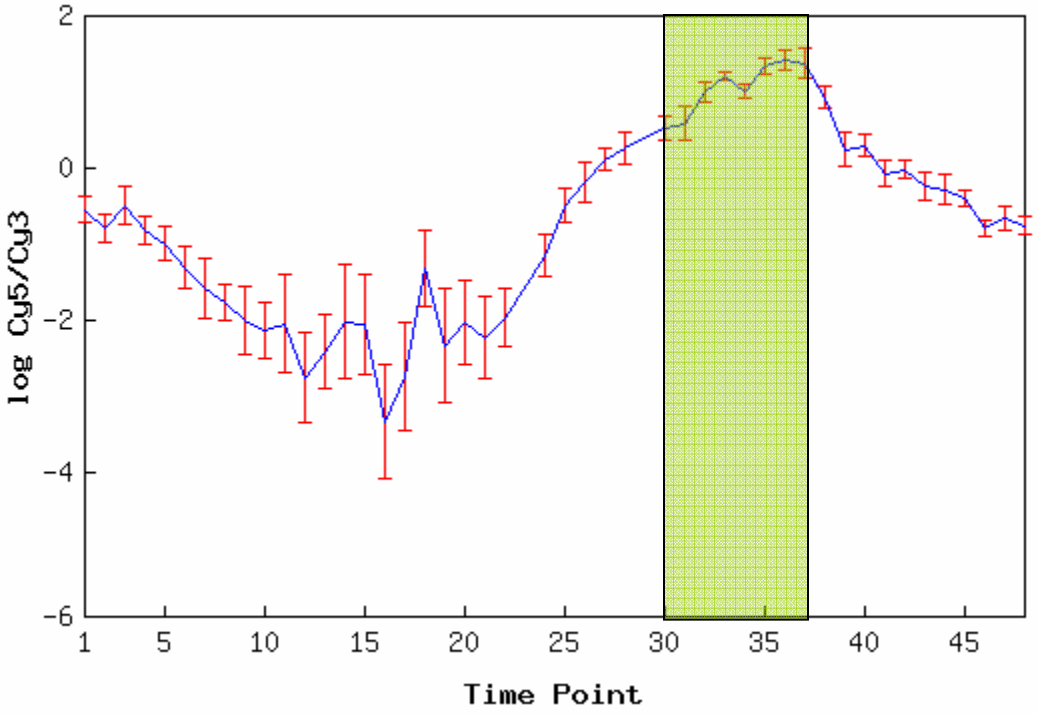
| *SERA* | Modules | Motif |
|---|---|---|
| PFB0325c<br>PFB0330c<br>PFB0335c<br>PFB0340c<br>PFB0345c<br>PFB0350c<br>PFB0355c<br>PFB0360c | rand5-80_m22_1<br>PFE0415w_g1.3_c8 |  |



http://biorg.cs.fiu.edu/TFBM/tfbm.php

# D: Module Average Expression Profile

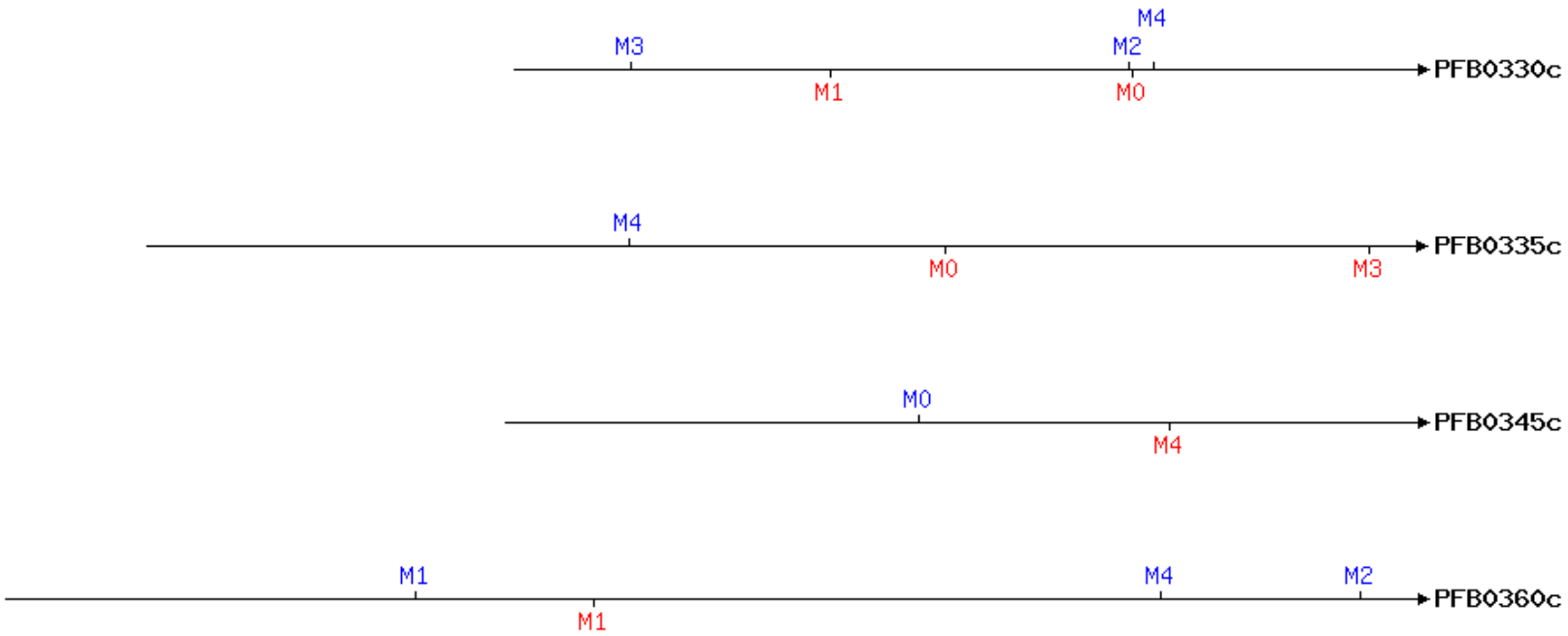
**PFE0415w_g1.3_c8**






BIORG
BioInformatics Research Group

# D: *SERA* TFBMs Visualization




BIORG
BioInformatics Research Group

# Conclusions

- **PlasmoTFBM: first comprehensive database of *P. falciparum* TFBMs**
- **Validated many known *P. falciparum* motifs**
- **Discovered new interesting motifs**
- **Web query interface built for biologists**

# Acknowledgements


- BioRG members (Tao Li, Gaolin Zheng, Tom Milledge)
- Prof. Shirley Liu, Harvard (MotifRegressor)
- Haifeng Wang, Jing Zhai & Wei Shi


http://biorg.cs.fiu.edu/CAMDA2004
http://biorg.cs.fiu.edu/TFBM/tfbm.php



BIORG
BioInformatics Research Group