Predicting functions in Plasmodium and other Apicomplexa

Chris Stoeckert, Ph.D.
Dept. of Genetics, Center for Bioinformatics
U. Penn
Graduate Seminars in Parasitology
Dec. 10, 2007
Structure and Function Assignments of Most Eukaryotic Parasite Genes Are Largely Based on Bioinformatic Analysis

- Not a parasitologist
  - Databases for parasite genomics
  - Data integration to understand (parasite) biology
- Bioinformatic approaches to predict protein function
  - Initial genome annotation is primarily bioinformatic with manual curation (i.e. not directly based on experiment)
- Share our observations on what is common among Apicomplexa and what is different in terms of protein families.
- Discuss what are the implications of these findings.
NIAID Bioinformatics Resource Centers for Biodefense and Emerging or Re-emerging Infectious Diseases

- focus on data related to multiple organisms selected from the NIAID lists of Category A-C priority pathogens and other pathogens causing emerging and re-emerging diseases
  - e.g., Category B includes *Cryptosporidium parvum*, *Giardia lamblia*, *Toxoplasma*
  - e.g., Category C includes *Trichomonas vaginalis*
Welcome to BRC Central

BRC Central is a repository linking to eight Bioinformatics Resource Centers (BRCs) sponsored by the National Institute of Allergy and Infectious Diseases (NIAID). The BRCs are providing web-based resources to scientific community conducting basic and applied research on organisms considered potential agents of biowarfare or bioterrorism or causing emerging or re-emerging diseases.

These centers support existing and newly developed techniques for bioinformatic analysis aimed at obtaining a deeper understanding of the fundamental biology of a specific set of pathogenic organisms, and efforts to counter the threats posed by these pathogens.

For Researchers

Here you will find the tools researchers can use across all BRC organisms. Click here to see a complete list of all tools available for pathogen researchers.

- BLAST
- Keyword Search
- Organism Search
- Taxonomy Browser
- Genome Summary
- FTP Downloads

More Tools for Researchers >>

For BRC Contributors

Here you will find the tools BRC contributors can use to submit data to BRC Central. Click here to see a complete list of all tools available for BRC contributors.

- Annotation Clearing House
- Dbxref Reference
- Conferences Sign Up

More Tools for BRC Contributors >>

For BRC Administrators

Here you will find the tools BRC administrators can use to track data on BRC Central. Click here to see a complete list of all tools available for BRC administrators.

- Genome Summary
- Sequence Source
- Dbxref Reference
- Internal Documentation

More Tools for BRC Administrators >>

Latest News

- October 26, 2007: Release of the new BRC Central website providing web-based resources to scientific community
  more >>
- September 26, 2007: Release of the new BRC Central data: 145,489 new genes, more >>
ApiDB.org

- Portal to Apicomplexan web sites
  - http://apidb.org/apidb/
  - PlasmoDB, ToxoDB, CryptoDB
- Also supporting GiardiaDB and TrichDB
- Not Apicomplexa but we can easily apply our database infrastructure to other pathogens
## Genomes in ApiDB
(Mouse over organism for more information)

<table>
<thead>
<tr>
<th>Organism/Strain</th>
<th>Last Updated</th>
<th>Genome Size(Mb)</th>
<th>Gene Count</th>
<th>Multiple Strains</th>
<th>SNPs ESTs</th>
<th>M</th>
<th>Pr</th>
<th>Pa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cryptosporidium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. hominis 7U022</td>
<td>11/2009</td>
<td>9.74</td>
<td>3590</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. muris</td>
<td>08/2007</td>
<td>8.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. parvum JW17A</td>
<td>08/2007</td>
<td>9.09</td>
<td>3888</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. berghei ANKA</td>
<td>06/2007</td>
<td>18.60</td>
<td>12345</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. chabaudi AS</td>
<td>06/2007</td>
<td>10.69</td>
<td>15095</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. falciparum 307</td>
<td>06/2007</td>
<td>23.27</td>
<td>5635</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. gallinaceum</td>
<td>06/2007</td>
<td>18.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. knowlesi</td>
<td>06/2007</td>
<td>25.44</td>
<td>5157</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. reichenowi</td>
<td>06/2007</td>
<td>7.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. vivax Salvator 1</td>
<td>06/2007</td>
<td>20.09</td>
<td>6416</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. yoelii 17XNL</td>
<td>06/2007</td>
<td>20.17</td>
<td>7971</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theileria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. annulata Anancare</td>
<td>03/2009</td>
<td>8.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. parva Muguga</td>
<td>03/2009</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toxoplasma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. gondii ME49</td>
<td>06/2007</td>
<td>63.50</td>
<td>8032</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Identify Genes by:

#### Genomic Position
- Chromosomal Location
- Proximity to Centromeres
- Proximity to Telomeres
- Non-nuclear Genomes

#### Gene Attributes
- Type (e.g. rRNA, tRNA)
- Exon/Intron Structure

#### Other Attributes
- Keyword
- List of IDs
- Species
- Available Reagents

### Transcript Expression
- EST Evidence
- SAGE Tag Evidence
- Microarray Evidence

### Protein Expression
- Mass Spec. Evidence

### Putative Function
- GO Term
- EC Number
- Metabolic Pathway
- VHDH Interaction
- Predicted Interaction

### Molecular Weight
- Isoelectric Point
- Protein Structure
- Epitopes

### Evolution
- Orthologs/Paralogs
- Orthology Profile
- Homology Profile
- Phylogenetic Tree

### Population Biology
- SNPs
- Microsatellites

### Identify SNPs by:
- SNP ID
- Gene ID
- AlleleFrequency
- Chromosomal Location

### Identify ORFs by:
- ORF ID
- Mass Spec. Evidence
- ORF Motif

### Identify ESTs by:
- EST ID
- BLAST Similarity
- EST Motif

### Identify Genomic Sequences by:
- Sequence ID
- Species
- BLAST Similarity
- DNA Motif

---

**ApiDB News**
- NEW! Tutorials offered also for Windows
- GiardiaDB and TrichDB have been released
- CryptDB 3.5 is released
- Status of the C. muris Genome Sequencing Project
- PlasmodDB 5.4 is released
- T. gondii gene prediction IDs now searchable
- ToxoDB 4.2 is released

**My ApiDB Account**
- Email: [Enter]
- Password: [Enter]
- Login

**ApicDB Outreach**
- Events with ApiDB Presence
- ApiDB Workshops
- ApiDB Publications

**Community Resources**
- BRC Central
- OrthoDB
- GeneDB
- ModBase at UCSF
- Tetrahymena Genome
- NCB: Entrez - PubCrawler
- More Resources...

**Information and Help**
- Click on the logos below to access taxon-specific sites.
ApiDB Personnel

• UGA
  – Jessica Kissinger
  – John Miller
  – Eileen Kraemer
  – Mark Heiges
  – Cristina Aurrecoechea
  – Cary Pennington
  – Doug Brewer
  – Haiming Wang
  – Alan Gingle
  – Kelly Storm
  – Zhiming Wang
  – Yunzhou Wu
  – Yolanda Kowalewski

• Upenn
  – David Roos
  – Chris Stoeckert
  – Brian Brunk
  – Steve Fischer
  – Bindu Gajria
  – Debbie Pinney
  – John Iodice
  – Jennifer Dommer
  – Jerric Gao
  – John Brestelli
  – Frank Innamorato
  – Jonathan Schug
  – Greg Grant
  – Charles Treatman
Apicomplexa are obligate intracellular parasites.
Some facts about *Plasmodium*

- *Plasmodium falciparum* is the causal organism of the most lethal form of Malaria
- Malaria is one of the three big killers (with TB and HIV/AIDS)
- >40% of the world’s population is exposed to the disease
- **300-500 Million** cases every year, and up to **2.7 Million** deaths (mostly children under the age of 5, ~ one death / 30 seconds).
- Drug-resistant malaria strains found in Asia, Africa and South America
## Plasmodium falciparum Genome

<table>
<thead>
<tr>
<th></th>
<th>P. falciparum</th>
<th>S. cerevisiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>22 Mb</td>
<td>12 Mb</td>
</tr>
<tr>
<td>No. of genes</td>
<td>5,444</td>
<td>5,770</td>
</tr>
<tr>
<td>Avg. gene length</td>
<td>2,283 kb</td>
<td>1,424 kb</td>
</tr>
<tr>
<td>G+C content</td>
<td>19.4 %</td>
<td>38.3 %</td>
</tr>
<tr>
<td>Hypothetical proteins</td>
<td>3537 (~65%)</td>
<td>~30%</td>
</tr>
<tr>
<td>Hypothetical proteins w/o pfam domain</td>
<td>2684 (~50%)</td>
<td></td>
</tr>
</tbody>
</table>

Characterizing these hypothetical proteins will increase our options for drug targets and vaccines.
Conventional approach to genome sequence annotation

...ACTGCGTATGCGTGCCTAGCTAGCATCGATCGATGCATCGATGCATCGATGCATCGATGCATCG;

Predict gene models

...ACTGCGTATGCGTGCCTAGCTAGCATCGATCGATGCATCGATGCATCGATGCATCGATGCATCG...

Similarity to characterized protein? (BLAST, homology)

Yes! Name it after that protein. (maybe add “-like”)  
No. Call it a “hypothetical protein.”
PlasmoDB.org hosts genomic and proteomic data (and more) for different species of the parasitic eukaryote Plasmodium, the cause of Malaria. It brings together data provided by numerous laboratories worldwide (see the Data Sources page), and adds its own data analysis. Publications relying on PlasmoDB should please acknowledge the database developers and the scientists who have made their data available. PlasmoDB is part of an NIH/NIAID funded Bioinformatics Resource Center to provide Apicomplexan Database Resources.

Features not yet available in PlasmoDB 5.4 may still be accessed via PlasmoDB 4.4, and the results of PlasmoDB 4.4 queries may be exported to PlasmoDB 5.4 (see PlasmoDB 4.4 Query History).

Related Sites:
- ApID
- ToxID
- CryptoDB
- OrthoMCL-DB
- Recent Plasmodium publications

News:
- 31 October 2007 PlasmoDB 5.4 is released
- 7 October 2007 Tutorials offered now for more platforms
- 19 June 2007 PlasmoDB 5.3 is released
- 14 February 2007 PFGRG offers course on Microarray Tech., April 2007
- All PlasmoDB News

PlasmoDB Events | Release Notes

Quick Tools [ Genome browser | BLAST | Sequence Retrieval | PlasmoCyc]

Search Genes

ID: PF11_0344
Keyword (search product name, notes, etc.): membrane
Gene Type: protein coding

Search Genomic Sequences

Genomic Sequence ID: MAL4

All available queries:
- Search Genes
- Search Genomic Sequences
- Search ESTs
- Search ORFs
- Search SNPs

PlasmoDB 4.4 queries/tools not yet in 5.4 >>
Find Vaccine Targets: Cell surface proteins expressed in schizonts not found in mammals and under diversifying selection.
# Gene Results

## Query

**SNP Characteristics**

**Details**

- **Reference**: Pf-3D7
- **Comparator(s)**: Pf-GHANA1
- **SNP Class**: Non-Synonymous
  - Number of SNPs of above class $\geq 5$
  - Number of SNPs of above class $\leq 1000$
  - Non-synonymous / synonymous SNP ratio $\leq 0$
  - Non-synonymous / synonymous SNP ratio $\leq 100$
  - SNPs per KB (CDS) $\geq 0$

## Results

- **392** (showing 1 to 20)

## Download

- **Combine with other results**
- **Orthologs**
- **Revise query**

## Table

<table>
<thead>
<tr>
<th>Gene</th>
<th>Product Description</th>
<th>Total SNPs</th>
<th>Non-synonymous SNPs</th>
<th>Synonymous SNPs</th>
<th>Nonsense SNPs</th>
<th>Non-coding SNPs</th>
<th>Non-syn/syn SNP ratio</th>
<th>SNPs per Kb (CDS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFL0030c</td>
<td>erythrocyte membrane protein 1 (PfEMP1)</td>
<td>616</td>
<td>469</td>
<td>147</td>
<td>5</td>
<td>0</td>
<td>3.19</td>
<td>67.17</td>
</tr>
<tr>
<td>PFE1640w</td>
<td>erythrocyte membrane protein 1 (PfEMP1), truncated</td>
<td>210</td>
<td>158</td>
<td>52</td>
<td>2</td>
<td>0</td>
<td>3.04</td>
<td>22.12</td>
</tr>
<tr>
<td>PFD1160w</td>
<td>surface-associated interspersed gene 4.2, (SURFIN4.2)</td>
<td>170</td>
<td>146</td>
<td>24</td>
<td>1</td>
<td>0</td>
<td>6.08</td>
<td>23.8</td>
</tr>
<tr>
<td>PF10_0374</td>
<td>Pf 11-1 protein</td>
<td>163</td>
<td>82</td>
<td>81</td>
<td>1</td>
<td>0</td>
<td>1.01</td>
<td>5.68</td>
</tr>
<tr>
<td>PF07_0004</td>
<td>hypothetical protein</td>
<td>159</td>
<td>124</td>
<td>32</td>
<td>2</td>
<td>3</td>
<td>3.88</td>
<td>53.89</td>
</tr>
<tr>
<td>PF10_0355</td>
<td>Erythrocyte membrane protein, putative</td>
<td>156</td>
<td>107</td>
<td>49</td>
<td>3</td>
<td>0</td>
<td>2.18</td>
<td>68.15</td>
</tr>
<tr>
<td>PFD0100c</td>
<td>surface-associated interspersed gene</td>
<td>118</td>
<td>101</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>5.04</td>
<td>18.24</td>
</tr>
</tbody>
</table>
### Gene query history

<table>
<thead>
<tr>
<th>ID</th>
<th>Query</th>
<th>Size</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Rename SNP Characteristics</td>
<td>392</td>
<td>orthologs view download revise delete</td>
</tr>
<tr>
<td>4</td>
<td>Rename Orthology Phylogenetic Profile</td>
<td>2815</td>
<td>orthologs view download revise delete</td>
</tr>
<tr>
<td>3</td>
<td>Rename Expression Timing (P.f.)</td>
<td>1116</td>
<td>orthologs view download revise delete</td>
</tr>
<tr>
<td>2</td>
<td>Rename Transmembrane Domain Count</td>
<td>1756</td>
<td>orthologs view download revise delete</td>
</tr>
<tr>
<td>1</td>
<td>Rename Predicted Signal Peptide</td>
<td>866</td>
<td>orthologs view download revise delete</td>
</tr>
</tbody>
</table>

**Combine results:** 1 and 2 and 3 and 4 and 5  
*Get Combined Result*  
*[eg: 1 or ((4 and 3) not 2)]*

**Delete all queries**

*Be careful: This will delete all your queries on PlasmoDB.*

**Understanding AND, OR and NOT:**

- **1 and 2**  
  Genes that 1 and 2 have in common. You can also use "1 intersect 2".

- **1 or 2**  
  Genes present in 1 or 2, or both. You can also use "1 union 2".

- **1 not 2**  
  Genes in 1 but not in 2. You can also use "1 minus 2".

* If you want to delete a query, you must first delete all other boolean queries that uses this one as a component.*
Query: 1 and 2 and 3 and 4 and 5

Details: Hide

INTERSECT

Query: Predicted Signal Peptide
Parameter: minimum SignalP conclusion score: 3
Organism: Plasmodium falciparum

INTERSECT

Query: Transmembrane Domain Count
Parameter: Minimum Number of Transmembrane Domains: 1
Maximum Number of Transmembrane Domains: 100
Organism: Plasmodium falciparum

INTERSECT

Query: Expression Timing (P.f.)
Parameter: Microarray time course(s) to query: Any of the time courses
Timing of maximal expression: 40 hours
Maximal expression time plus or minus: 8 hours
Timing of minimum expression: 1 hour
Minimum expression time plus or minus: 1 hour
Induction ratio cut-off: >= 2 fold induction
Maximum percentile cut-off: >= 80th percentile

INTERSECT

Query: Orthology Phylogenetic Profile
Parameter: Included Species: pfa, pvi
Excluded Species: Mammals
Organism: Plasmodium falciparum

Query: SNP Characteristics
Parameter: Reference: Pf-3D7
Comparator(s): Pf-GHANA1
SNP Class: Non-Synonymous
Number of SNPs of above class: >= 5
Number of SNPs of above class: <= 1000
Non-synonymous / synonymous SNP ratio: >= 0
Non-synonymous / synonymous SNP ratio: <= 100
SNPs per KB (CDS): >= 0

Results: 8 (showing 1 to 8)
<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
<th>Genomic Location</th>
<th>Product Description</th>
<th># TM Domains</th>
<th>Pf-iRBC expr profile graph (GS array)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAL7P1.176</td>
<td>P. falciparum 3D7</td>
<td>MAL7: 1,413,431 - 1,417,944 (+)</td>
<td>erythrocyte binding antigen</td>
<td>1</td>
<td><img src="image1" alt="Graph" /></td>
</tr>
<tr>
<td>PF10_0082</td>
<td>P. falciparum 3D7</td>
<td>MAL10: 351,703 - 355,593 (+)</td>
<td>hypothetical protein</td>
<td>10</td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td>PF10_0177</td>
<td>P. falciparum 3D7</td>
<td>MAL10: 721,733 - 732,855 (-)</td>
<td>erythrocyte membrane-associated antigen</td>
<td>1</td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>PF10_0189</td>
<td>P. falciparum 3D7</td>
<td>MAL10: 796,008 - 799,922 (+)</td>
<td>hypothetical protein</td>
<td>2</td>
<td><img src="image4" alt="Graph" /></td>
</tr>
</tbody>
</table>
By combining computational analysis (signal peptide and transmembrane motif sequence search, ortholog groups) with experimental data (expression analysis and SNP analysis), known and novel vaccine targets can be identified. The result includes hypothetical proteins for which we now have clues about function based on associations and sequence motifs.
Computational models: going beyond combining query results

Model the ‘functional interactome’ of the parasite

– Use functional linkages from computational and experimental functional genomics datasets to generate a protein-protein interaction network.

– Instead of using datasets independently, combine them within a Bayesian framework, which simultaneously allows standardization and incorporation of expert knowledge (against predefined ‘gold standards’).

**Definition: Functional linkages**
Proteins that are members of the same pathway, or are a part of the same protein complex or even a cellular system are functionally linked.
Interactome Modeling: Previous Studies


Interactome modeling: Datasets

• Computational functional genomics datasets
  – Phylogenetic profile linkages (Date & Stoeckert, Penn)
  – Rosetta stone linkages (Date & Stoeckert, Penn)

• Experimental functional genomics datasets
  – Microarray expression time-series (DeRisi Lab, UCSF)
  – Microarray expression data for all stages (Winzeler Lab, Scripps)
  – Mass spectrometry (Carruci Lab, Navy; Mann Lab, Denmark)

• Annotation datasets (used as gold standards)
  – Gene Ontology (GO) annotations (from Sanger & TIGR)
  – KEGG Pathway annotations
Phylogenetic profiles are a description of the presence or absence of a given protein in a set of reference genomes (Pellegrini et al, *PNAS* 96, 1999).

<table>
<thead>
<tr>
<th>Genomes</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
<th>G5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein1</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Protein2</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Protein3</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Simple Binary Profile (1 = presence, 0 = absence)**

> Protein1 1 0 0 1 1
Phylogenetic profile data

A phylogenetic profile constructed using BLAST E-values

\[ p_{ij} = -1 / \log (\text{BLAST E-value}) \]

>gi|9654538|conserved hypothetical [Caulobacter crescentus]  

1.00 1.00 1.00 1.00 1.00 1.00 1.00 0.02 1.00 1.00 0.01 1.00 0.14 1.00 1.00 1.00 0.06 1.00 1.00 1.00 1.00 1.00 1.00 0.14 1.00 1.00 0.18 1.00 1.00 1.00 0.00* 1.00 1.00 1.00 1.00 1.00 1.00 1.00

Absence (1.00)  Strong Presence (0.00)

Archaea  Bacteria  Eukaryotes

Final dataset:
Profiles of 2813 proteins that were found in at least one other organism, other than \( P. \ falciparum \). Similarity between phylogenetic profiles was defined by their mutual information.
Proteins that appear as a single fused protein in one organism, but as two or more separate proteins in either the same or a different organism.

**Yeast Topo II**
- *E. coli* gyrA
- *E. coli* gyrB

**Known RS**
- Unknown protA
- Unknown protB

---

**Final dataset:**
Fusion protein links between 993 proteins that were found in at least one other organism, other than *P. falciparum*. Linkage confidence was measured using the hypergeometric distribution.
Generating the interactome

Approach: Combine functional genomics data (instead of using each data set separately) within a Bayesian framework

- G - gold standards
- P – phylogenetic profiles
- R – Rosetta stone links
- E₁ – expression set 1
- E₂ – expression set 2
- M₁ – mass spec. set 1
- M₂ – mass spec. set 2

Workflow for Interactome Modeling:

1. Assemble data sets
2. Calculate likelihood scores for individual data sets, by comparing with gold standards
3. Use a reference prior to calculate a likelihood score threshold (predict links with odds of 1 or greater)
4. Measure the quality of the isolated (links above the score threshold)
5. Is the quality acceptable?
   - NO
   - YES
6. Use links to create a network
7. Validate trends and data quality by varying benchmarks
8. Is the quality acceptable?
   - NO
   - YES
9. Use network to draw biological inferences
Gold standards (KEGG + GO)

Positives: Pair proteins in the same KEGG pathway (remove promiscuous components) \( (G_P) \)

Negatives: Pair all proteins that are not in the same KEGG pathway \( (G_N) \), filter out proteins with shared GO terms (up to 7 levels) \( (G'_N) \)

Final dataset:

10,267 positive pairs and 44,812 negative pairs
P. falciparum interaction network: Examples

- Assemble data sets
  - Calculate likelihood scores for individual data sets, by comparing with gold standards
  - Use a reference prior to calculate a likelihood score threshold (predict links with odds of 1 or greater)
  - Measure the quality of the isolated links above the score threshold
  - Is the quality acceptable?
    - NO
    - YES
    - Use links to create a network
  - Validate trends and data quality by varying benchmarks
  - Is the quality acceptable?
    - YES
    - Use network to draw biological inferences

Diagram: Interacting proteins and pathways in P. falciparum, showing examples of RNA-binding proteins, hypothetical proteins, and splicing factors.
**P. falciparum** interaction network

A mix of hypothetical proteins, known splicing factors, and proteins which bind RNA.

**Resource**

Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale

Shailesh V. Date¹ and Christian J. Stoeckert Jr.

Center for Bioinformatics, Department of Genetics, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Shailesh Date
Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale

Shailesh V. Date and Christian J. Steckert, Jr
Center for Bioinformatics, Department of Genetics,
School of Medicine, University of Pennsylvania, Philadelphia, PA
Please address all correspondence to: SVD

Companion website

Abstract

Many thousands of proteins encoded by the genome of *Plasmodium falciparum*, the causative organism of the deadliest form of malaria, are of utmost importance that these proteins be characterized. If we are to develop combinatorial strategies against malaria based on protein function on a genome-wide scale, we computationally modeled the *P. falciparum* interactome, elucidating local products. The resulting interaction network, constructed by integrating in silico and experimental functional genomics data and protein interaction data, provides functional inferences for more than 2000 uncharacterized proteins, based on their observed novel strategy, whereby we incorporated continuously updated, uniform reference priors in our Bayesian model. This method is suitable for application to other genomes, where protein interactions are known. Additionally, we discovered potential pathways in the *P. falciparum* interactome, leading to the prediction of new interacting partners and new potential targets to discover new treatments for malaria.

All results of this study are available for download at http://cbi.upenn.edu/plasmoMAP/.

Dynamic queries

- Generate a data set of any desired cutoff
  - Strain (threshold min/max):

Done

Shailesh Date, Ghislain Bidaut

Data access: PlasmoMAP and PlasmoDB
Hypothetical protein from the vaccine target query

P. falciparum 3D7 protein coding on MAL10 from 351,703 to 355,593 (3891 bp)

Genomic Context

SNPs Summary

Annotation

Add a comment on PF10_0082

User Comments

External Links

Database Link

GeneDB

Literature Database

Ontology Based pattern Identification

OrthoMCL

PASA ESTs

PlasmoMAP

UCSC Plasmodium falciparum genome browser
PF10_0082 :: hypothetical protein

This page lists functional linkage information available from interactome modeling analysis of *P. falciparum* (Date & Stoeckert, Genome Res., 2006). These data were used to reconstruct a genome-scale interaction map of all proteins, and can be used to annotate individual proteins based on their connections with known proteins, using the 'guilt-by-association' principle.

Jump to: GO Annotations | KEGG Annotations | Fusion information | plasmoMAP links

**GO ANNOTATIONS**

Top 3 over-represented GO categories of proteins linked to PF10_0082 (out of 50 total):

- 5 instances of GO:0030163 protein catabolism
- 5 instances of GO:0016043 cell organization and biogenesis
- 5 instances of GO:0009056 catabolism
- 5 instances of GO:0006508 proteolysis and peptidolysis
- 5 instances of GO:0009057 catabolism macromolecule catabolism
- 4 instances of GO:0006810 cell growth and/or maintenance transport
- 4 instances of GO:0006996 organelle organization and biogenesis
- 4 instances of GO:0007028 cytoplasm organization and biogenesis
- 4 instances of GO:0007010 cytoskeleton organization and biogenesis
- 3 instances of GO:0016310 phosphorylation

Note: Categories represent GO process information. Ubiquitous GO categories- GO:0008151, GO:0008152, GO:0000004, GO:0009058, GO:0019538 and GO:0006139 have been filtered out from the top 3 links [see all GO links for PF10_0082].

**KEGG ANNOTATIONS**

Top 3 over-represented KEGG categories of proteins linked to PF10_0082 (out of 3 total):

- 1 instances of 00720 Reductive carboxylate cycle (CO2 fixation)
- 1 instances of 00710 Carbon fixation
- 1 instances of 00620 Pyruvate metabolism

Note: Ubiquitous KEGG categories- 00230, 00240 and 03010 have been filtered out from the top 3 links [see all KEGG links for PF10_0082].
Going beyond the
Conventional approach to genome sequence annotation

...ACTGCGTATGCGTAGCTAGCATCGATCGATGCATCGATGCATCGATGCATCGATGCATCG...;

Predict gene models

Similarity to characterized protein? (BLAST, homology)

Yes! Name it after that protein. (maybe add “-like”)

No. Call it a “hypothetical protein.”
Comparisons of protein sequence similarity within genomes to identify common features

**Investigate types of proteins found**

- Analyze organization of the proteome with regard to
  - Distribution of proteins in families (number of members in families: singletons, family of size 2, family of size 3, etc.)
  - Percentage of the proteome occupied by family types
  - Distribution characteristics as compared with other genomes

- Check ability to generate functional information for proteins based on their membership in particular groups

**Definition: Protein families or clusters**
Proteins that share the most domains over the entire length of the protein. No assertion as to evolutionary past (i.e., not necessarily orthologs). May contain multiple known (e.g., PFAM) or unknown protein domains.
Proteome organization

- All vs. all pair-wise comparisons using BLASTClust
- Create clusters based on sequence similarity
- Compare with clusters from other organisms
- Explore and characterize individual clusters

Filter: 30/30 rule

P. falciparum proteome

Shailesh Date, Debbie Pinney
Cluster characteristics include proteins found in all Apicomplexa and those found only in Plasmodium.

Cluster 1 includes hypothetical protein PF10_0082 from vaccine target query.
Other hypothetical protein from vaccine target query, PF10_0189 does not match any other proteins (singleton).
A Cluster shared by all Apicomplexan: 78% are hypotheticals. Known proteins include hydrolases.
Proteome organization: domain analysis

- Family composed on hypothetical protein members
Domain analysis: isolation of a novel domain?

PsiBLAST analysis (5 iterations): No hits to proteins other than P. falciparum
UniProt clusters No new proteins
OrthoMCL One ortholog in P. vivax
NCBI CDD No hits
Pfam No hits in Pfam A, one in Pfam B
Novel domain is putatively involved in pathogenesis

Information from plasmoMAP

>PFA0035c,
>PFB0056c,
>PFL2650w,
>PF10_0007,
>PF14_0765,
>PF10060c

GO Annotations
8 GO:0020033 antigenic variation
8 GO:0020012 evasion of host immune response
Proteome organization

- **P. falciparum** proteome
- All vs. all pair-wise comparisons using *BLASTClust*
- Create clusters based on sequence similarity
- Filter: 30/30 rule
- Compare with clusters from other organisms
- Explore and characterize individual clusters

Shailesh Date, Debbie Pinney
Proteome organization differs in *P. falciparum*

1. Is the bias Apicomplexan specific?
2. Is the bias ‘AT-rich genome’ specific?

**Genome coverage**
- 1: 71.3
- 2: 4.55
- 3: 1.22
- 4: 0.74

**Cluster types**
- Clusters of size 1: 3858
- Clusters of size 2: 246
- Clusters of size 3: 22
- Clusters of size 4: 10

**Plasmodium falciparum**
- Total number of clusters: 4042
- Cluster types: 17
- Largest cluster size: 693 prots
- Smallest cluster size: 1 prot

1. Is the bias Apicomplexan specific?
2. Is the bias ‘AT-rich genome’ specific?
Proteome organization is shared among Apicomplexa
Proteome organization in Apicomplexa differs from Non-apicomplexa
Other parasites, *E. cuniculi* and *G. lamblia* have a similar proteome organization.

Not all parasites (e.g., *E. histolytica*) share the Apicomplexan signature.
Proteome organization: Inferences

★ Observations:

– Distribution of proteins (in families) in apicomplexan genomes shows marked differences, compared to prokaryotes and most other eukaryotes
– More singletons, and fewer small families were noted

★ Possible causes:

1. Apicomplexan proteomes are old and singletons reflect lack of genome duplication
2. Suppression of lineage-specific expansion of protein families (gene families by proxy), except for families related to invasion and pathogenesis.
3. Paralogs have diverged or have been lost as parasites adapt to niches.
4. Singletons reflect genes acquired from hosts.
5. Annotation of genomes is a work in progress
Summary

- What makes pathogens unique also makes them hard to understand
- BRCs provide a foundation for study bringing together available genomic information
- Genomic and functional genomic datasets may be integrated to predict functional interactions
- Study of proteome characteristics reveals clues about Apicomplexan proteins
Welcome to the Computational Biology and Informatics Laboratory in the Center for Bioinformatics, part of the Genomics Institute at the University of Pennsylvania.

### Tools
- OrthoMCL
- GO Function Predictions
- Mouse Chr. 5
- tplWY
- PaGE
- GSIM
- TESS
- STAC
- COGRIM

### Databases
- PlasmoDB
- EPConDB
- plasmoMAP
- ApiDots
- AllGenes
- ToxoDB
- ApiDB
- GUS
- SCGAP
- RAD
- MTIR

### News
**December 6 2007**
CryptoDB released version 3.6 and ToxoDB released version 4.3 joining PlasmoDB version 5.4 (Oct. 31) with providing new features and data this Fall. ApiDB version 3.1 will be released shortly.

**October 7 2007**
*GiardiaDB* and *TrichDB* have been released to support the efforts of researchers studying the pathogens, *Giardia lamblia* and *Trichomonas vaginalis*, respectively. These resources reflect the expansion of the ApiDB Bioinformatics Resource Center to additional organisms that are eukaryotic pathogens though not Apicomplexa. GUS and WDK were used to build the databases and web sites.

**August 7 2007**
Version 4 of EPConDB has been released using the new WDK (the technology used by the ApiDB sites). As part of this release, additional lists have been added to the improved query for differentially expressed genes.